Problem Set 1

Due: September 28, 2023 (11:59 PM EST)

Please turn in a PDF with solutions typed in LaTeX. You may collaborate on this problem set, but please attempt the problems yourself first. It goes without saying, please do not try to look up solutions online. Feel free to use Wikipedia, etc. for material pertaining to reasonable course prerequisites (e.g. basic probability, linear algebra, etc.) If you find a bug in a problem, please let me know.

Exercise 1 (Dobrushin Uniqueness and Ising Models)**.** *Let μ be a probability measure on $[q]^n$, where recall that for a positive integer $q \in \mathbb{N}$, $[q] \stackrel{\mathsf{def}}{=} \{1, \ldots, q\}$. For a subset $S \subseteq [n]$ and a (partial) configuration $\tau : S \to [q]$, we let μ^τ denote the induced conditional measure, i.e.*

$$\mu^\tau(\sigma) \propto \begin{cases} \mu(\sigma), & \text{if } \sigma(i) = \tau(i), \forall i \in S \\ 0, & \text{otherwise.} \end{cases}$$

For $i \in [n]$ and $\mathfrak{c} \in [q]$, we also use the notation $i \leftarrow \mathfrak{c}$ for the case $S = \{i\}$ and $\tau(i) = \mathfrak{c}$. For all $j \in [n]$, we write μ_j for the marginal distribution induced on j, i.e.

$$\mu_j(\mathfrak{c}) = \sum_{\sigma \in [q]^n : \sigma(j) = \mathfrak{c}} \mu(\sigma), \qquad \forall \mathfrak{c} \in [q].$$

(a) Define the $n \times n$ Dobrushin influence matrix *$\mathscr{R} \in \mathbb{R}^{n \times n}$ as follows:*

$$\mathscr{R}(i \to j) = \max_{\tau : [n] \setminus \{i,j\} \to [q]} \max_{\mathfrak{b}, \mathfrak{c} \in [q]} d_{\mathsf{TV}}\left(\mu_j^{\tau, i \leftarrow \mathfrak{b}}, \mu_j^{\tau, i \leftarrow \mathfrak{c}}\right), \qquad \forall i \neq j.$$

(We set $\mathscr{R}(i \to i) = 0$ for the diagonal entries.) Prove that if

$$\|\mathscr{R}\|_{\ell_\infty \to \ell_\infty} \stackrel{\mathsf{def}}{=} \max_{i \in [n]} \sum_{j=1}^{n} |\mathscr{R}(i \to j)| = 1 - \delta,$$

for some constant $0 < \delta < 1$, then Glauber dynamics with stationary measure μ has mixing time $\mathrm{T}_{\mathsf{mix}}(\epsilon) \leq O((n/\delta) \log(n/\epsilon))$.

(b) For a symmetric $n \times n$ matrix $J \in \mathbb{R}^{n \times n}$ and $\beta \geq 0$, define a probability measure μ on $\{\pm 1\}^n$ by

$$\mu(\sigma) \propto \exp\left(\frac{\beta}{2} \sigma^\top J \sigma\right).$$

This is an Ising model *with interaction matrix J; here $\beta \geq 0$ is the* inverse temperature*, and is treated as a constant independent of n. Prove that for all $i \neq j$, $\mathscr{R}(i \to j) \leq |\tanh(\beta J_{ij})|$.*

(c) Consider the case where $J = \frac{1}{n}\mathbf{1}\mathbf{1}^\top$; this is the Curie–Weiss model.

- *Prove that when $\beta < 1$, Glauber dynamics mixes in $O(n \log n)$ steps.*
- *Prove that when $\beta > 1$, Glauber dynamics needs $\exp(\Omega(n))$ steps to mix.*
  (If it makes things convenient, you may use the approximation $\binom{n}{pn} \approx \exp(n \cdot H_e(p))$ for all $p \in [0, 1]$ as if it is an equality, where $H_e(p) \stackrel{\mathsf{def}}{=} -p \ln(p) - (1-p) \ln(1-p)$ is a slight modification of the standard binary entropy function.)

Exercise 2 (Correlation Inequalities via Markov Chains)**.** *Let μ, ν be two probability measures over $2^{[n]}$, endowed with the containment order. We say ν* stochastically dominates *μ if there exists a coupling ξ of μ, ν such that $S \subseteq T$ holds with probability 1 for $(S, T) \sim \xi$. Such a coupling is called* monotone.

(a) Now let P_μ, P_ν be Markov chains with stationary measures μ, ν, respectively. We say P_ν stochastically dominates *P_μ if for every $S \subseteq T$, the transition measure $P_\nu(T \to \cdot)$ stochastically dominates $P_\mu(S \to \cdot)$. In other words, there is a Markovian coupling $(X_t, Y_t)_{t=0}^\infty$ of P_μ, P_ν such that for all initial states $X_0, Y_0 \in 2^{[n]}$, $X_0 \subseteq Y_0$ implies $X_t \subseteq Y_t$ for all t with probability 1.*

Prove that if P_μ, P_ν have μ, ν as their unique *stationary measures, and P_ν stochastically dominates P_μ, then ν stochastically dominates μ.*

(b) Assume μ, ν are both strictly positive. Suppose for all $A, B \in 2^{[n]}$, we have the inequality

$$\nu(A \cup B) \cdot \mu(A \cap B) \geq \mu(A) \cdot \nu(B).$$

Prove that ν stochastically dominates μ.

(c) Prove that if ν stochastically dominates μ, then for every increasing function $f : 2^{[n]} \to \mathbb{R}$, we have the inequality $\mathbb{E}_\mu[f] \leq \mathbb{E}_\nu[f]$.[1]

(d) Suppose $\mu(S) \propto \exp(F(S))$ where $F : 2^{[n]} \to \mathbb{R}$ is a supermodular set function*, i.e.*

$$F(S \cup T) + F(S \cap T) \geq F(S) + F(T), \qquad \forall S, T \in 2^{[n]}.$$

In other words, μ is a log-supermodular point process*. Prove that μ is positively correlated in the following sense:*

$$\Pr_{S \sim \mu}[i, j \in S] \geq \Pr_{S \sim \mu}[i \in S] \cdot \Pr_{S \sim \mu}[j \in S], \qquad \forall i, j \in [n].$$

Exercise 3 (Concentration of Trajectories)**.** *The goal of this exercise is to prove the following theorem: Let P be an ergodic Markov chain on a finite state space Ω which is reversible w.r.t. a probability measure μ. Let $(X_t)_{t\geq 0}^\infty$ be a trajectory of P with the initial state X_0 drawn from the stationary distribution μ. Then for every nonnegative integer $T \in \mathbb{N}$, every $\epsilon > 0$, and every 1-bounded function $f : \Omega \to [0, 1]$, we have the concentration inequality*

$$\Pr\left[\left|\frac{1}{T}\sum_{t=0}^{T-1} f(X_t) - \mathbb{E}_\mu[f]\right| \geq \boldsymbol{\lambda}_2^*(\mathsf{P}) + \epsilon\right] \leq 2\exp\left(-C\epsilon^2 T\right),$$

for some universal constant $C > 0$, where $\boldsymbol{\lambda}_2^*(\mathsf{P}) = \max_{i>1}\{|\boldsymbol{\lambda}_i(\mathsf{P})|\}$ gives the second largest eigenvalue of P in absolute value.

(a) Like most proofs of Chernoff-like concentration inequalities, the crucial step is obtaining a good bound on the moment generating function*. Let $s \geq 0$ be a parameter to be determined later. Establish the following bound:*

$$\mathbb{E}\left[\exp\left(s\sum_{t=0}^{T-1} f(X_t)\right)\right] \leq \|M_f\|_{\mathsf{op}}^{T-1} \cdot \mathbb{E}_\mu\left[e^{sf}\right],$$

where $M_f \in \mathbb{R}^{\Omega\times\Omega}$ is the symmetric matrix given by

$$M_f = \operatorname{diag}\left(e^{sf/2}\right) \cdot \operatorname{diag}(\mu)^{1/2}\,\mathsf{P}\,\operatorname{diag}(\mu)^{-1/2} \cdot \operatorname{diag}\left(e^{sf/2}\right),$$

and $\|\cdot\|_{\mathsf{op}}$ denotes the usual operator norm w.r.t. the standard Euclidean inner product.

(b) Complete the proof of the theorem.

(Hint: Aim for the upper bound $\|M_f\|_{\mathsf{op}} \leq \boldsymbol{\lambda}_2^*(\mathsf{P}) \cdot e^s + (1 - \boldsymbol{\lambda}_2^*(\mathsf{P})) \cdot \mathbb{E}_\mu\left[e^{sf}\right]$. It may also be helpful to use the inequality $1 - e^{-x} \geq x - \frac{1}{2}x^2$, which holds for all $x \geq 0$.)

[1] This is the original definition of stochastic domination. However, Strassen's Theorem says that these are equivalent.

Remark 1. Better results are known. In particular, it is known there is a universal constant $C > 0$ such that for all $\epsilon > 0$,

$$\Pr\left[\left|\frac{1}{T}\sum_{t=0}^{T-1} f(X_t) - \mathbb{E}_\mu[f]\right| \geq \epsilon\right] \leq 2\exp\left(-C(1-\lambda_2^*(\mathsf{P}))\cdot\epsilon^2 T\right).$$

This stronger bound greatly generalizes the classical Chernoff bound, where P is the trivial Markov chain $\mathbf{1}\mu^\top$. These types of results are really useful in practice due to the following reason: The usual Chernoff bound assumes independence, so naïvely, you would run $O\left(\frac{1}{\epsilon^2\cdot\mathbb{E}_\mu[f]}\log(1/\delta)\right)$ independent copies of the chain until mixing to be able to do Monte Carlo estimation. The running time cost becomes this factor multiplied by the mixing time (unless you run the chains in parallel). This result allows you instead to run a single copy of the chain until mixing, and take the next $O\left(\frac{1}{(1-\lambda_2^*(\mathsf{P}))\cdot\epsilon^2\cdot\mathbb{E}_\mu[f]}\log(1/\delta)\right)$ states in the trajectory as your samples. You pay an extra multiplicative factor of $\frac{1}{1-\lambda_2^*(\mathsf{P})}$ in the number of samples, but the combined running time is additive with the mixing time, not multiplicative.

There are loads of additional applications to *derandomization* as well, where P is typically the simple random walk on a d-regular expander graph with $d = O(1)$.