PSET 3

Due on **Thursday, October 23, 2025 at 11:59 pm ET**

COMPSCI 2233 / PHYSICS 272: Quantum Learning Theory, Fall 2025 (Sitan Chen, Jordan Cotler) Version: **1.0**

Submitted by **<name>**

This problem set will cover concepts from shadow tomography and Gibbs state learning.
The questions have been labeled with the date of the lecture in which the relevant material is covered, to help you budget your time. The questions are meant to be challenging, so do not feel discouraged if you get stuck and are unable to solve some of them.
If you find that you are running low on time to finish all the problems, our recommendation is to try to aim for breadth rather than depth – e.g., it is better to complete a few parts of each of the three questions, than to completely solve one of the three questions and skip the others.
Below we provide hints for the various problems in this assignment. While these may help you solve the problems more easily, you are not required to follow the hints as long as the proofs you provide are correct.

1 (25 pts.) Learning Parameterized Quantum States (10/8 and 10/15)

Motivation: Full quantum state tomography requires a number of samples exponential in the number of qubits, which is generally intractable. However, many quantum states encountered in physics and quantum algorithms are not arbitrary but belong to families described by a small number of parameters. In this problem, we will explore how to leverage this structure to learn an unknown state from such a family using only $\tilde{\mathcal{O}}(n)$ copies. This problem will combine several powerful ideas taught in class: covering nets, efficient Gibbs state preparation, online learning, and shadow tomography.

Setup: We are given classical knowledge of a family of n-qubit pure states parameterized by a vector $\vec{x} = (x_1, \ldots, x_k) \in [-1, 1]^k$, where the number of parameters is $k = \mathcal{O}(\log n)$. The parameterized state is given by a polynomial-depth quantum circuit $U(\vec{x})$ such that $|\psi(\vec{x})\rangle = U(\vec{x})\,|0\rangle^{\otimes n}$. We are also told this family is Lipschitz continuous, that is,

$$\| \, |\psi(\vec{x})\rangle - |\psi(\vec{x}')\rangle \, \|_2 \leqslant \|\vec{x} - \vec{x}'\|_2 \, .$$

We are given $\tilde{\mathcal{O}}(n)$ copies of an unknown state ρ with the promise that $\rho = |\psi(\vec{x}_{\text{true}})\rangle \langle\psi(\vec{x}_{\text{true}})|$ for some unknown $\vec{x}_{\text{true}} \in [-1, 1]^k$. Our goal is to find a parameter vector $\vec{x}^*$ such that

$$\|\rho - |\psi(\vec{x}^*)\rangle\langle\psi(\vec{x}^*)| \, \|_1 < 0.01 \, ,$$

where $\| \cdot \|_1$ is the trace norm.

1.A. (5 pts.) **(Constructing a Covering Net)** First, we must discretize the continuous parameter space. Show that it is possible to construct a finite set of parameter vectors $\mathcal{N} \subset [-1, 1]^k$, called a net, such that for any potential true parameter vector $\vec{x}_{\text{true}}$, there exists a net vector $\vec{x}_i \in \mathcal{N}$ satisfying $\| \, |\psi(\vec{x}_{\text{true}})\rangle\langle\psi(\vec{x}_{\text{true}})| - |\psi(\vec{x}_i)\rangle\langle\psi(\vec{x}_i)| \, \|_1 \leqslant 0.005$. Prove that the size of this net, $M = |\mathcal{N}|$, is polynomial in n. Let the set of states corresponding to this net be $\mathcal{S}_{\text{net}} = \{|\psi_i\rangle\}_{i=1}^M$.

1.B. (5 pts.) **(Computationally Efficient Online Learning for Rank-1 Observables)** Assume access to a $\mathrm{poly}(n)$-time quantum algorithm that can prepare a single copy of the state $\exp(-H)/\operatorname{tr}\exp(-H)$ given any Hermitian operator H with polynomial rank (such an algorithm was developed by Brandão et al. in "Quantum SDP Solvers"). Using this oracle, describe a complete, computationally efficient quantum algorithm for online learning of quantum states when the observables are restricted to be rank 1. Your proof of correctness can cite as a black box any guarantees that were stated in class about matrix multiplicative weights.

1.C. (7 pts.) **(Computationally Efficient Shadow Tomography for Rank-1 Observables)** The complete shadow tomography protocol taught in class is only guaranteed to be sample-efficient; its computational complexity can be exponential. However, by replacing its internal online learning subroutine with your efficient implementation from the previous question, we can make the entire protocol computationally efficient for our specific problem.

Describe how to create a quantum algorithm for shadow tomography for observables $O_1 = |\psi_1\rangle\langle\psi_1| \, , \ldots, O_M = |\psi_M\rangle\langle\psi_M|$ that runs in $\mathrm{poly}(n)$ time, uses $\tilde{\mathcal{O}}(n)$ copies of an unknown state ρ, and predicts the expectation values $\operatorname{tr}(|\psi_i\rangle\langle\psi_i| \, \rho)$ up to any small constant error for all $i \in \{1, \ldots, M\}$. You may use Theorem 115 from the lecture notes as a black box (while we did not stipulate that the blended measurements algorithm is computationally efficient, you may assume that without proof for this problem).

1.D. (8 pts.) **(The Full Learning Algorithm)** You now have all the necessary components. Combine them to develop a final quantum learning algorithm that solves the original problem of learning parametrized quantum states with high probability in $\mathrm{poly}(n)$ time. Your description should first outline the complete sequence of steps. Then, provide a proof of correctness, showing that the output $\vec{x}^*$ satisfies the desired accuracy bound. Your proof should also justify the sample complexity, explaining why the number of copies of ρ needed is only $\tilde{\mathcal{O}}(n)$.

Solution:

1.A.

1.B.

1.C.

1.D.

2 (40 pts.) Learning Classical Hamiltonians from Gibbs States (10/15 and 10/20)

Motivation: In class we were introduced to the problem of Hamiltonian learning from Gibbs states. It turns out that this question has also been the subject of considerable interest in the *classical* learning theory community. Classical Gibbs states enjoy a number of nice properties that quantum Gibbs states do not, which enable the use of algorithmic techniques which do not yet have suitable quantum analogues. In this exercise, you will explore one such technique, the *pseudolikelihood method*.

In this question, a *classical Gibbs state with pairwise interactions*, sometimes known as an *Ising model*, is a probability distribution μ over the Boolean hypercube $\{-1,1\}^n$ which is specified by a symmetric *interaction matrix* $A \in \mathbb{R}^{n\times n}$ and which has probability mass function

$$\mu(\boldsymbol{x}) = \frac{1}{Z}\exp(-\boldsymbol{x}^\top A\boldsymbol{x}) \qquad \text{for} \qquad Z \triangleq \sum_{\boldsymbol{x}\in\{-1,1\}^n} \exp(-\boldsymbol{x}^\top A\boldsymbol{x})\,.$$

Given symmetric interaction matrix A' corresponding to classical Gibbs state μ', define the *pseudolikelihood loss* $\mathrm{PL}(A') \geqslant 0$ with respect to μ to be the quantity

$$\mathrm{PL}(A') \triangleq \sum_{i=1}^n \mathrm{PL}^{(i)}(A') \qquad \text{for} \qquad \mathrm{PL}^{(i)}(A') \triangleq \mathop{\mathbf{E}}_{\boldsymbol{z}\sim\mu}\Big[-\log(\mu'(\boldsymbol{x}_i = \boldsymbol{z}_i \mid \boldsymbol{x}_{\setminus i}))\Big]\,.$$

In this exercise, you will show that if $\mathrm{PL}(\mu')$ is minimized over all Ising models μ', then the minimizer is exactly μ.

2.A. (5 pts.) Let $i \in [n]$ and let $\boldsymbol{x}_{\setminus i} \in \{-1,1\}^{n-1}$ be any assignment to the coordinates of $\boldsymbol{x}$ outside of the i-th coordinate. Compute the conditional probability

$$\mu(\boldsymbol{x}_i = 1 \mid \boldsymbol{x}_{\setminus i})\,.$$

(Your derivation should be entirely self-contained, e.g., do not invoke any facts that were not proved in class.)

Conclude that $\mathrm{PL}^{(i)}_\mu(A')$ only depends on the off-diagonal entries of A' in the i-th row (recall that A' is symmetric).

Thus, henceforth, given a symmetric matrix $M \in \mathbb{R}^{n\times n}$ we will write $M[i] \in \mathbb{R}^{n-1}$ to denote the vector consisting of all off-diagonal entries in the i-th row. We will use $\mathrm{PL}^{(i)}_\mu(A'[i])$ in place of $\mathrm{PL}^{(i)}_\mu(A'[i])$ so that in all subsequent parts, $\mathrm{PL}^{(i)}_\mu$ is a function $\mathbb{R}^{n-1} \to \mathbb{R}$.

Next, given a differentiable function $F : \mathbb{R}^d \to \mathbb{R}$, $x \in \mathbb{R}^d$, define the *curvature* of f around x by

$$\delta F_x(\Delta) = F(x+\Delta) - F(x) - \langle \Delta, \nabla F(x)\rangle$$

for all $\Delta \in \mathbb{R}^d$. Our goal in the next two parts will be to lower bound the curvature of the pseudolikelihood loss.

Below, you may use without proof the following elementary calculus fact:

> **Fact:** For $\psi(x) \triangleq \log(1+\exp(-2x))$, $\delta\psi_x(\Delta) \geqslant \exp(-2\gamma)\Delta^2/2$ for any $\Delta \in \mathbb{R}$ if $\max(|x|, |x+\Delta|) \leqslant \gamma$.

2.B. (7 pts.) Show that for any A, A' for which the off-diagonal entries in the i-th column of either have magnitudes summing to at most γ, then for $\Delta \triangleq A'[i] - A[i]$,

$$\delta\mathrm{PL}^{(i)}_A(\Delta) \geqslant \frac{\exp(-2\gamma)}{2}\mathrm{Var}_\mu\Big[\sum_{j:j\neq i}\Delta_{ij}\boldsymbol{x}_j\Big]\,.$$

2.C. (15 pts.) Conclude that for A, A' satisfying the assumptions of the previous part, $\delta\mathrm{PL}^{(i)}_A(\Delta) \geqslant \frac{\exp(-4\gamma)}{2}\|\Delta[i]\|_\infty^2$.
Hint: Computing expectations of quadratic functions over μ can be unwieldy. You may find it helpful to use the law of total variance, as well as your characterization of the conditional marginals of μ from Part **2.A.**.

In the last part of this exercise, we will show how to go from a bound on the curvature to a statement about the optimization landscape of the pseudolikelihood loss. Let Ω denote the set of matrices A' for which the sum of the off-diagonal entries in any row is at most γ, and suppose $A \in \Omega$. You may use without proof that $\|\nabla\mathrm{PL}^{(i)}_\mu(A')\|_1 \leqslant C$ for some finite quantity C for all $A' \in \Omega$.

2.D. (8 pts.) Suppose A' satisfies $\mathrm{PL}(A') \leqslant \min_{A^*}\mathrm{PL}(A^*) + \eta$. Give an upper bound on $\|A' - A\|_\infty$ that depends only on C, γ, n, η and which tends to zero as $\eta \to 0$.

2.E. (5 pts.) Describe informally in a few sentences how one might use the above ingredients to design an algorithm that takes samples from μ and learns an approximation of A. You do not need to prove this algorithm works, but the intuition needs to be correct.

Solution:

2.A.

2.B.

2.C.

2.D.

2.E.

3 (40 pts.) Lower bound for learning Hamiltonians (10/15 and 10/20)

Motivation: In this problem you will derive an information-theoretic lower bound for learning a Hamiltonian from copies of its Gibbs state in the ℓ_∞ metric. The proof reduces learning to a multi-hypothesis test between closely related Gibbs distributions and applies Fano's lemma. As we will see, the proof ingredients are entirely classical; later in this course, we will revisit these tools and develop "inherently quantum" analogues.

Setup: Let $\rho_\beta(H) = e^{-\beta H}/\mathrm{tr}(e^{-\beta H})$ be the Gibbs state at inverse temperature $\beta > 0$. Consider the Pauli matrices $Z \otimes I$, $I \otimes Z$, and $Z \otimes Z$, which are diagonal with

$$Z \otimes I = \mathrm{diag}(1, 1, -1, -1), \quad I \otimes Z = \mathrm{diag}(1, -1, 1, -1), \quad Z \otimes Z = \mathrm{diag}(1, -1, -1, 1).$$

For any $\varepsilon \in (0, \frac{1}{2}]$, define two diagonal 2-qubit Hamiltonians

$$H_0 = (-1)(Z \otimes I) - \frac{1}{2}(I \otimes Z) - \frac{1}{2}(Z \otimes Z) = \mathrm{diag}(-2, 0, 1, 1),$$

$$H_1 = (-1)(Z \otimes I) + \big(-\frac{1}{2} + \varepsilon\big)(I \otimes Z) + \big(-\frac{1}{2} - \varepsilon\big)(Z \otimes Z) = \mathrm{diag}(-2, 0, 1 + 2\varepsilon, 1 - 2\varepsilon).$$

Let q_0, q_1 denote the classical probability vectors obtained from the diagonal entries of $\rho_\beta(H_0), \rho_\beta(H_1)$, respectively, and write $D_{\mathrm{KL}}(p\|q) = \sum_j p_j \log(p_j/q_j)$.

3.A. (8 pts.) **Warm-up: explicit Gibbs distributions.** Compute q_0 and q_1, and then derive the exact expression for $D_{\mathrm{KL}}(q_1\|q_0)$.

3.B. (14 pts.) **Bounding the KL divergence.** Show that

$$D_{\mathrm{KL}}(q_1\|q_0) \leqslant 8\,\beta^2\varepsilon^2\,e^{-3\beta+2\beta\varepsilon} \leqslant 8\,\beta^2\varepsilon^2\,e^{-2\beta} \quad \text{for } \varepsilon \in (0, \frac{1}{2}].$$

Hints: Use $\log x \geqslant 1 - \frac{1}{x}$ for $x > 0$ and $1 - e^{-x} \leqslant x$ for $x \geqslant 0$, and note $e^{-3\beta+2\beta\varepsilon} \leqslant e^{-2\beta}$ when $\varepsilon \leqslant \frac{1}{2}$.

3.C. (6 pts.) **Product construction and the chain rule.** Consider a system of $2N$ qubits organized into N disjoint pairs of 2 qubits each. We construct $N + 1$ hypotheses about the system's total Hamiltonian:

- **Null hypothesis:** All N pairs have Hamiltonian H_0. This induces the product distribution $p_0 = q_0^{\otimes N}$.
- **Alternative hypothesis i (for $i \in [N]$):** The i-th pair has Hamiltonian H_1; all other pairs have H_0. This induces the product distribution $p_i = q_0 \otimes \cdots q_0 \otimes \underbrace{q_1}_{i\text{th slot}} \otimes q_0 \otimes \cdots q_0$.

With S i.i.d. samples from these distributions, prove that:

$$D_{\mathrm{KL}}\big(p_i^{\otimes S} \,\big\|\, p_0^{\otimes S}\big) = S\, D_{\mathrm{KL}}(q_1\|q_0).$$

This shows the KL divergence scales linearly with the number of samples and depends only on the single-pair divergence.

Hint: Use the chain rule for KL divergence (you may use this without proof) and the fact that the hypotheses differ on exactly one pair.

3.D. (12 pts.) **From Fano's lemma to a lower bound.** You may use the following form of Fano's lemma.

Fact 1: Fano's lemma. Let $P_0, P_1, \ldots, P_N$ be distributions such that $\frac{1}{N+1}\sum_{j=1}^{N} D_{\mathrm{KL}}(P_j\|P_0) \leqslant \alpha$ with $0 < \alpha < \log N$. Then the minimax error of testing among $\{P_0, \ldots, P_N\}$ is at least

$$1 - \frac{\log 2 + \alpha}{\log N}.$$

Here, the minimax error of testing is $\inf_\tau \max_{j\in\{0,\ldots,N\}} \mathrm{Prob}_{X\sim P_j}[\tau(X) \neq j]$, where $\tau(X)$ is any test that attempts to identify which distribution generated the observation X. This represents the best achievable worst-case error probability across all possible testing procedures.

Fano's lemma thus provides a fundamental lower bound: when the distributions are close on average (small α), no test can reliably distinguish between them. Apply Fano's lemma to $\{p_0^{\otimes S}, p_1^{\otimes S}, \ldots, p_N^{\otimes S}\}$ together with your bound from part (b) to show that any learner achieving ℓ_∞ error at most $\varepsilon/2$ with failure probability at most, say, $1/3$ must use

$$S = \Omega\left(\frac{e^{2\beta}}{\beta^2\varepsilon^2}\,\log N\right) \quad \text{copies of the Gibbs state.}$$

Hint: Learning to ℓ_∞ error $< \varepsilon/2$ distinguishes which pair corresponds to H_1.

Solution:

3.A.

3.B.

3.C.

3.D.