Ph 220: Quantum Learning Theory – Lecture Note 3: Learning and Predicting Properties of a Quantum State

Hsin-Yuan Huang (Robert)

Caltech

In this lecture, we explore the fundamental problems of learning and predicting properties of an unknown quantum state. We begin by looking at an extremely simple approach for learning the full description of a quantum state. We will then study random unitaries and randomized measurements to develop a better approach for learning and predicting properties of quantum states.

1 A Simple Approach for Learning a Quantum State

Given access to multiple copies of an n-qubit quantum state ρ, our goal is to produce a classical description, $\hat{\rho}$, that is close to the true state ρ.

Q: How do we obtain a full classical description of an n-qubit quantum state ρ?

To represent the $2^n \times 2^n$ density matrix ρ, we can express it in a convenient basis. The set of n-qubit Pauli operators, $\{I, X, Y, Z\}^{\otimes n}$, forms an orthonormal basis for the space of Hermitian matrices under the Hilbert-Schmidt inner product, $\langle A, B\rangle = \mathrm{tr}(A^\dagger B)$. Any state ρ can be uniquely decomposed as:

$$\rho = \sum_{P \in \{I,X,Y,Z\}^{\otimes n}} \alpha_P \frac{P}{2^n}$$

where the Pauli coefficients $\alpha_P = \mathrm{tr}(P\rho)$ are the real expectation values of the corresponding Pauli operators. Since $\mathrm{tr}(\rho) = 1$, the coefficient for the identity Pauli string $P = I^{\otimes n}$ is always $\alpha_I = 1$. The problem of learning the state ρ is therefore equivalent to learning the remaining $4^n - 1$ unknown Pauli coefficients, which are all between -1 and $+1$.

1.1 Algorithm

A straightforward way to learn these coefficients is to measure the expectation value of each Pauli operator directly. For each Pauli operator $P \in \{I, X, Y, Z\}^{\otimes n}$ (excluding the identity):

1. Prepare a copy of the state ρ.
2. Measure the observable P. This is done by measuring each qubit i in the basis corresponding to the single-qubit Pauli operator P_i. If $P_i = X$, we measure in the $\{|+\rangle, |-\rangle\}$ basis; if $P_i = Y$, we measure in the $\{|y+\rangle, |y-\rangle\}$ basis; if $P_i = Z$ or I, we measure in the $\{|0\rangle, |1\rangle\}$ basis.
3. The measurement yields an outcome $b_i \in \{+1, -1\}$ for each qubit. The overall outcome is the product $b_1 b_2 \cdots b_n$, which is an eigenvalue of P.

4. Repeat the above steps k times and average the results to obtain an estimate $\hat{\alpha}_P$ for the true expectation value $\alpha_P = \text{tr}(P\rho)$.

After estimating all the coefficients, we construct an estimate of the state: $\hat{\rho} = \sum_P \hat{\alpha}_P \frac{P}{2^n}$, where we set $\hat{\alpha}_I = 1$. Our goal is to determine the total number of copies of ρ (the sample complexity) required to ensure that our estimate $\hat{\rho}$ is close to ρ with high probability.

1.2 Sample Complexity Analysis

Let's analyze the number of samples needed for this simple tomography scheme. Each measurement of a Pauli operator P yields an outcome in $\{+1, -1\}$. The average of k such outcomes, $\hat{\alpha}_P$, can be related to its expectation α_P using Hoeffding's inequality:

$$\Pr\left[|\hat{\alpha}_P - \alpha_P| \geq \varepsilon\right] \leq 2e^{-k\varepsilon^2/2}$$

To reconstruct the full state, we need to estimate all $4^n - 1$ coefficients accurately at the same time. Using the union bound, the probability that *any* of our estimates is off by more than ε is at most $4^n \cdot 2e^{-k\varepsilon^2/2}$. To make this failure probability smaller than some δ, the number of measurements *per Pauli basis* must be:

$$k = O\left(\frac{\log(4^n/\delta)}{\varepsilon^2}\right) = O\left(\frac{n + \log(1/\delta)}{\varepsilon^2}\right)$$

Next, we relate the coefficient error ε to the error in the final state estimate, measured by the squared Frobenius norm:

$$\|\hat{\rho} - \rho\|_F^2 = \text{tr}\left(\left(\sum_P (\hat{\alpha}_P - \alpha_P)\frac{P}{2^n}\right)^2\right) = \frac{1}{2^n}\sum_P (\hat{\alpha}_P - \alpha_P)^2$$

If we guarantee that $|\hat{\alpha}_P - \alpha_P| \leq \varepsilon$ for all P, the squared Frobenius norm is bounded by $\|\hat{\rho} - \rho\|_F^2 \leq \frac{1}{2^n}\sum_P \varepsilon^2 = 2^n\varepsilon^2$. To achieve a final state error of $\|\hat{\rho} - \rho\|_F \leq \epsilon$, we must set the precision for each coefficient to $\varepsilon \leq \epsilon/\sqrt{2^n}$. Substituting this into our expression for k, the total sample complexity $N = 4^n \cdot k$ becomes:

$$N = 4^n \cdot O\left(\frac{(n + \log(1/\delta))}{(\epsilon/\sqrt{2^n})^2}\right) = O\left(\frac{(n + \log(1/\delta))8^n}{\epsilon^2}\right)$$

This exponential scaling in 8^n makes the protocol prohibitively expensive for even a small number of qubits, motivating the search for more efficient methods.

1.3 From State Tomography to Predicting Observables

A key operational meaning of "learning a state" is the ability to accurately predict the expectation values of arbitrary observables. Let O be an observable with eigenvalues bounded between -1 and $+1$ (i.e., its operator norm $\|O\|_\infty \leq 1$). The error in predicting $\text{tr}(O\rho)$ using our learned state $\hat{\rho}$ is:

$$|\text{tr}(O\hat{\rho}) - \text{tr}(O\rho)| = |\text{tr}(O(\hat{\rho} - \rho))|$$

By Hölder's inequality for matrices, this error is bounded by the trace norm (Schatten 1-norm) of the state error:

$$|\text{tr}(O(\hat{\rho} - \rho))| \leq \|O\|_\infty \|\hat{\rho} - \rho\|_1 \leq \|\hat{\rho} - \rho\|_1$$

This shows that to bound the prediction error for *any* such observable, we must bound the trace distance between $\hat{\rho}$ and ρ. We can relate the trace norm to the Frobenius norm using the inequality $\|\cdot\|_1 \leq \sqrt{d}\|\cdot\|_F$ for matrices of dimension $d = 2^n$. This gives:

$$\|\hat{\rho} - \rho\|_1 \leq \sqrt{2^n}\|\hat{\rho} - \rho\|_F$$

To guarantee a prediction error of at most ϵ for any bounded observable, we need $\|\hat{\rho} - \rho\|_1 \leq \epsilon$. This requires us to achieve a Frobenius norm error of $\|\hat{\rho} - \rho\|_F \leq \epsilon/\sqrt{2^n}$. Plugging this new, more stringent error target into our sample complexity formula yields:

$$N = O\left(\frac{(n + \log(1/\delta))8^n}{(\epsilon/\sqrt{2^n})^2}\right) = O\left(\frac{(n + \log(1/\delta))8^n \cdot 2^n}{\epsilon^2}\right) = O\left(\frac{(n + \log(1/\delta))16^n}{\epsilon^2}\right)$$

The sample complexity to learn the state well enough for universal prediction is even worse, scaling as 16^n. This underscores the extreme inefficiency of this simple approach for practical tasks.

2 Randomized Measurements

The poor scaling of direct Pauli measurement motivates us to consider alternative strategies. One powerful alternative is to use randomized measurements. The core idea is to apply a random unitary transformation to the state and then perform a simple, fixed measurement, such as a measurement in the computational basis (Z basis). The procedure for a single copy is:

$$\rho \to U \to \text{Z-basis Measurement} \to b \in \{0, 1\}^n$$

To analyze this, we first need to understand how to integrate over the unitary group.

2.1 The Haar Measure and Haar Integration

The uniform distribution over the group of $d \times d$ unitary matrices $U(d)$ (where $d = 2^n$) is called the **Haar measure**, denoted $d\mu_H(U)$. It is a unique probability measure that is invariant under left and right multiplication. For any integrable function f and any fixed unitary $V \in U(d)$:

$$\int_{U(d)} f(U)d\mu_H(U) = \int_{U(d)} f(VU)d\mu_H(U) = \int_{U(d)} f(UV)d\mu_H(U).$$

This invariance allows us to compute averages of functions over the unitary group, a process often called "twirling" in the quantum information science literature.

2.1.1 First Moment Integrals

Let's compute the integral of a unitary matrix itself. Using the left-invariance property with $V = -I \in U(d)$, we have

$$\int U d\mu_H(U) = \int (-I)U d\mu_H(U) = -\int U d\mu_H(U).$$

This implies that $\int U d\mu_H(U) = 0$.

Next, let's compute the average of $UXU^\dagger$ for a fixed operator X. Let $M(X) = \int UXU^\dagger d\mu_H(U)$. By the invariance of the Haar measure, for any unitary V:

$$VM(X)V^\dagger = V\left(\int UXU^\dagger d\mu_H(U)\right)V^\dagger = \int (VU)X(VU)^\dagger d\mu_H(U) = M(X).$$

This implies $[V, M(X) = 0$ for all unitaries V. By considering $V = e^{-itH}$ and taking derivative in t, we have $[H, M(X)] = 0$ for all Hermitian matrices H. Because any matrix A can be written as $A = H_1 + iH_2$ for Hermitian matrices H_1, H_2, this implies that $[A, M(X)] = 0$ for all matrices A. This means $M(X)$ commutes with any matrix, hence $M(X)$ is proportional to the identity matrix. $M(X) = c(X) \cdot I$. To find the function $c(X)$, we take the trace:

$$\mathrm{Tr}(M(X)) = \int \mathrm{Tr}(UXU^\dagger) d\mu_H(U) = \int \mathrm{Tr}(X) d\mu_H(U) = \mathrm{Tr}(X).$$

Since we also have $\mathrm{Tr}(M(X)) = \mathrm{Tr}(c(X) \cdot I) = c(X) \cdot d$, we find $c(X) = \mathrm{Tr}(X)/d$. This gives the well-known twirling formula:

$$\int_{U(d)} UXU^\dagger d\mu_H(U) = \frac{\mathrm{Tr}(X)}{d} I.$$

2.1.2 Higher Moments and Schur-Weyl Duality

To analyze more complex protocols, we need higher moments of the Haar distribution. Specifically, we want to compute the integral of $U^{\otimes t} X U^{\dagger \otimes t}$ for a fixed operator X acting on the t-fold tensor product space $(\mathbb{C}^d)^{\otimes t}$. The primary tool for this is **Schur-Weyl duality**.

Let $M_t(X) = \int_{U(d)} U^{\otimes t} X U^{\dagger \otimes t} d\mu_H(U)$. This operator must commute with $V^{\otimes t}$ for any $V \in U(d)$, a fact that follows directly from the invariance of the Haar measure:

$$V^{\otimes t} M_t(X) (V^\dagger)^{\otimes t} = \int (VU)^{\otimes t} X (VU)^{\dagger, \otimes t} d\mu_H(U) = M_t(X)$$

Schur-Weyl duality states that the algebra of operators that commute with all tensor powers $\{V^{\otimes t} | V \in U(d)\}$ is precisely the algebra spanned by the permutation operators $\{\sigma \in S_t\}$ acting on the t Hilbert space copies. Therefore, the integral must be a linear combination of these permutation operators:

$$M_t(X) = \sum_{\sigma \in S_t} C_\sigma(X) \cdot \sigma$$

where the coefficients $C_\sigma(X)$ depend on the operator X.

Linear Independence of Permutation Operators. For $t \leq d$, we can utilize a useful fact that the permutation operators $\{\sigma \in S_t\}$ are linearly independent. Consider an orthonormal basis $\{|0\rangle, |1\rangle, \ldots, |d-1\rangle\}$ for the single-copy Hilbert space $\mathbb{C}^d$. We can form a specific state in the t-fold tensor product space $(\mathbb{C}^d)^{\otimes t}$ by taking the first t distinct basis vectors:

$$|\psi\rangle = |0\rangle \otimes |1\rangle \otimes \cdots \otimes |t-1\rangle$$

A permutation operator $\sigma \in S_t$ acts on this state by permuting the tensor factors:

$$\sigma |\psi\rangle = |\sigma(0)\rangle \otimes |\sigma(1)\rangle \otimes \cdots \otimes |\sigma(t-1)\rangle$$

Now, consider two distinct permutations, $\sigma \neq \tau$. The states $\sigma|\psi\rangle$ and $\tau|\psi\rangle$ are orthogonal:

$$\langle \psi | \tau^\dagger \sigma | \psi \rangle = \langle \tau\psi | \sigma\psi \rangle = (\langle \tau(0)| \cdots \langle \tau(t-1)|) (|\sigma(0)\rangle \cdots |\sigma(t-1)\rangle)$$

$$= \langle \tau(0)|\sigma(0)\rangle \langle \tau(1)|\sigma(1)\rangle \cdots \langle \tau(t-1)|\sigma(t-1)\rangle$$

Since σ and τ are different permutations, there must be at least one index j for which $\sigma(j) \neq \tau(j)$. Because we chose an orthonormal basis, $\langle \tau(j)|\sigma(j)\rangle = \delta_{\tau(j),\sigma(j)} = 0$. Therefore, the entire product

is zero. This orthogonality allows us to prove linear independence. Assume for contradiction that there is a non-trivial linear combination that equals zero:

$$\sum_{\sigma \in S_t} c_\sigma \sigma = 0$$

Applying this operator equation to our state $|\psi\rangle$ gives $\sum_{\sigma \in S_t} c_\sigma \sigma |\psi\rangle = 0$. Now, we take the inner product with the state $\langle \tau\psi|$ for an arbitrary $\tau \in S_t$:

$$\langle \tau\psi| \left(\sum_{\sigma \in S_t} c_\sigma \sigma |\psi\rangle \right) = \sum_{\sigma \in S_t} c_\sigma \langle \tau\psi|\sigma\psi\rangle = 0$$

As we just showed, $\langle \tau\psi|\sigma\psi\rangle = \delta_{\tau,\sigma}$. The sum thus collapses to a single term:

$$\sum_{\sigma \in S_t} c_\sigma \delta_{\tau,\sigma} = c_\tau = 0$$

Since this holds for any $\tau \in S_t$, all coefficients must be zero. The set of permutation operators is therefore linearly independent.

Deriving the General Form of the Integral. Since the integral $M_t(X)$ is a linear map of X, each coefficient $C_\sigma(X)$ must be a linear functional of X. This can be shown using

$$|\psi\rangle = |0\rangle \otimes |1\rangle \otimes \cdots \otimes |t-1\rangle .$$

More precisely, we have

$$\langle \sigma\psi| M_t(X) |\psi\rangle = C_\sigma(X).$$

Hence, $C_\sigma(X)$ is a linear function of X. For any linear functional on the space of matrices, there exists a fixed matrix C'_σ such that the functional can be represented as an inner product: $C_\sigma(X) = \mathrm{Tr}(C'_\sigma X)$. This gives us:

$$M_t(X) = \sum_{\sigma \in S_t} \mathrm{Tr}(C'_\sigma X)\sigma$$

Now we must determine the structure of the operators C'_σ. The map $M_t(\cdot)$ has an additional symmetry. For any fixed unitary V:

$$M_t(V^{\otimes t} X V^{\dagger,\otimes t}) = \int U^{\otimes t}(V^{\otimes t} X V^{\dagger,\otimes t})U^{\dagger,\otimes t} d\mu_H(U)$$

From the invariance of Haar measure, we have:

$$M_t(V^{\otimes t} X V^{\dagger,\otimes t}) = M_t(X)$$

for all $V \in U(d)$. Let's apply our formula to both sides of this identity:

$$\mathrm{RHS} = M_t(X) = \sum_{\sigma \in S_t} \mathrm{Tr}(C'_\sigma X)\sigma.$$

$$\mathrm{LHS} = M_t(V^{\otimes t} X V^{\dagger,\otimes t}) = \sum_{\sigma \in S_t} \mathrm{Tr}(C'_\sigma (V^{\otimes t} X V^{\dagger,\otimes t}))\sigma = \sum_{\sigma \in S_t} \mathrm{Tr}(V^{\dagger,\otimes t} C'_\sigma V^{\otimes t} X)\sigma.$$

By the linear independence of $\{\sigma\}$, we can equate the coefficients for each σ:

$$\mathrm{Tr}(C'_\sigma X) = \mathrm{Tr}((V^{\dagger,\otimes t} C'_\sigma V^{\otimes t})X).$$

This equality must hold for all X. This implies that the operators themselves must be equal:

$$C'_\sigma = V^{\dagger,\otimes t} C'_\sigma V^{\otimes t} \quad \implies \quad [V^{\otimes t}, C'_\sigma] = 0.$$

This means that each operator C'_σ must commute with $V^{\otimes t}$ for any unitary V. But this is the same condition we started with! By Schur-Weyl duality, any operator C'_σ that commutes with $V^{\otimes t}$ for all $V \in U(d)$ must itself be in the algebra spanned by the permutation operators. Therefore, each C'_σ can be written as a linear combination of permutations:

$$C'_\sigma = \sum_{\tau \in S_t} W_{\sigma,\tau^{-1}} \tau$$

where $W_{\sigma,\tau^{-1}}$ are scalar coefficients. Substituting this final piece back into our expression for the integral gives the general formula for Haar integration of tensor powers:

$$\int_{U(d)} U^{\otimes t} X U^{\dagger \otimes t} d\mu_H(U) = \sum_{\sigma,\tau \in S_t} W_{\sigma,\tau^{-1}} \mathrm{Tr}(\tau X)\sigma = \sum_{\sigma,\tau \in S_t} W_{\sigma,\tau} \mathrm{Tr}(\tau^{-1} X)\sigma.$$

The matrix W with entries $W_{\sigma,\tau}$ is the **Weingarten matrix**.

2.1.3 Weingarten Calculus

To find the coefficients $W_{\sigma,\tau}$, we can set X to be a permutation operator σ'. Since σ' commutes with $U^{\otimes t}$, the integral is simply σ'.

$$\sigma' = \sum_{\sigma,\tau \in S_t} W_{\sigma,\tau} \mathrm{Tr}(\tau^{-1}\sigma')\sigma.$$

By linear independence of $\{\sigma\}$, we can equate coefficients:

$$\delta_{\sigma,\sigma'} = \sum_{\tau \in S_t} W_{\sigma,\tau} \mathrm{Tr}(\tau^{-1}\sigma').$$

Let W be the matrix with entries $W_{\sigma,\tau}$ and G be the **Gram matrix** with entries $G_{\tau,\sigma'} = \mathrm{Tr}(\tau^{-1}\sigma')$. The equation is $I = W \cdot G^T$. Since G is symmetric, we have $W = G^{-1}$. The matrix W is the Weingarten matrix. The entries of the Gram matrix are $G_{\tau,\sigma} = \mathrm{Tr}(\tau^{-1}\sigma) = d^{\#\mathrm{cycles}(\tau^{-1}\sigma)}$.

The Second Moment ($t = 2$). For the important case of $t = 2$, the symmetric group is $S_2 = \{I, \mathrm{Swap}\}$. The Gram matrix is:

$$G = \begin{pmatrix} \mathrm{Tr}(I^{-1}I) & \mathrm{Tr}(I^{-1}\mathrm{Swap}) \\ \mathrm{Tr}(\mathrm{Swap}^{-1}I) & \mathrm{Tr}(\mathrm{Swap}^{-1}\mathrm{Swap}) \end{pmatrix} = \begin{pmatrix} \mathrm{Tr}(I) & \mathrm{Tr}(\mathrm{Swap}) \\ \mathrm{Tr}(\mathrm{Swap}) & \mathrm{Tr}(I) \end{pmatrix} = \begin{pmatrix} d^2 & d \\ d & d^2 \end{pmatrix}.$$

Its inverse is the Weingarten matrix $W = G^{-1}$:

$$W = \frac{1}{d^4 - d^2} \begin{pmatrix} d^2 & -d \\ -d & d^2 \end{pmatrix} = \frac{1}{d^2 - 1} \begin{pmatrix} 1 & -1/d \\ -1/d & 1 \end{pmatrix}.$$

We can now use this to write the explicit formula for the second-moment twirling channel. The general formula is:

$$\int U^{\otimes 2} X U^{\dagger \otimes 2} d\mu_H(U) = \sum_{\sigma,\tau \in S_2} W_{\sigma,\tau} \mathrm{Tr}(\tau^{-1} X)\sigma$$

Expanding the sum over $\sigma, \tau \in \{I, \mathrm{Swap}\}$ gives four terms:

$$\begin{aligned}\int U^{\otimes 2} X U^{\dagger \otimes 2} d\mu_H(U) = \quad & W_{I,I} \mathrm{Tr}(I^{-1} X) I \\ & + W_{I,\mathrm{Swap}} \mathrm{Tr}(\mathrm{Swap}^{-1} \cdot X) I \\ & + W_{\mathrm{Swap},I} \mathrm{Tr}(I^{-1} X) \mathrm{Swap} \\ & + W_{\mathrm{Swap},\mathrm{Swap}} \mathrm{Tr}(\mathrm{Swap}^{-1} \cdot X) \mathrm{Swap}\end{aligned}$$

Grouping terms by I and Swap and substituting the values from the Weingarten matrix:

$$\int U^{\otimes 2} X U^{\dagger \otimes 2} d\mu_H(U) = \left(\frac{1}{d^2-1} \mathrm{Tr}(X) - \frac{1}{d(d^2-1)} \mathrm{Tr}(\mathrm{Swap} \cdot X) \right) I \tag{1}$$

$$+ \left(-\frac{1}{d(d^2-1)} \mathrm{Tr}(X) + \frac{1}{d^2-1} \mathrm{Tr}(\mathrm{Swap} \cdot X) \right) \mathrm{Swap} \tag{2}$$

This is the complete formula for the second-moment Haar integral, often used in randomized benchmarking and the analysis of randomized measurements.

2.1.4 Unitary t-Designs

Implementing a truly Haar-random unitary is exponentially hard (since there is an exponential number of random bits that one needs to sample from). A **unitary t-design** is a distribution μ over unitaries that reproduces the first t moments of the Haar measure, i.e.,

$$\int U^{\otimes t} X U^{\otimes t,\dagger} d\mu(U) = \int U^{\otimes t} X U^{\otimes t,\dagger} d\mu_H(U).$$

The Clifford group on n qubits forms a unitary 3-design. Since a random Clifford circuit can be implemented in $\mathcal{O}(n)$ depth on 1D circuit and $\mathcal{O}(\log n)$ depth on all-to-all-connected circuits, it is often used in quantum information science as a surrogate for Haar-random unitaries.

2.2 Data Collection Phase

The protocol for generating "classical shadows" is surprisingly simple. To generate N shadows, we repeat the following procedure N times:

1. Take a fresh copy of the state ρ.
2. Sample a unitary U_i from a suitable distribution (e.g., the Haar measure or a unitary 3-design like the Clifford group). Apply it to the state: $\rho \to U_i \rho U_i^\dagger$.
3. Measure the resulting state in the Z basis, obtaining a bit string outcome $b_i \in \{0,1\}^n$.
4. The classical data recorded for this shot is the pair (U_i, b_i).

The key question is: how can we use this collection of classical data $\{(U_1, b_1), \ldots, (U_N, b_N)\}$ obtained from the quantum experiments to learn about the unknown state ρ?

2.3 Constructing an Unbiased Estimator

From each measurement shot (U_i, b_i), we can form a random matrix $\hat{\sigma}_i = U_i^\dagger |b_i\rangle\langle b_i| U_i$. Let's compute the expectation value of this random matrix, averaging over both the choice of unitary U and the probabilistic measurement outcome b.

$$\mathbb{E}[\hat{\sigma}] = \mathbb{E}_U \left[\mathbb{E}_{b|U}[U^\dagger |b\rangle\langle b| U] \right] = \int d\mu_H(U) \sum_{b \in \{0,1\}^n} \mathrm{Tr}(\rho U^\dagger |b\rangle\langle b| U) U^\dagger |b\rangle\langle b| U$$

This integral can be solved using the second moment of the Haar distribution (or a unitary 2-design). The result is a simple affine map on the state ρ:

$$\mathbb{E}[\hat{\sigma}] = \frac{\rho + \mathbb{I}}{d+1} \quad \text{where } d = 2^n.$$

This shows that $\hat{\sigma}$ is a biased estimator for ρ. However, we can easily invert this linear map to construct an *unbiased* estimator. Let's define the single-shot estimator, or "classical snapshot," as:

$$\hat{\rho}_i = (d+1)\hat{\sigma}_i - \mathbb{I} = (d+1) U_i^\dagger |b_i\rangle\langle b_i| U_i - \mathbb{I}$$

Taking the expectation value confirms it is unbiased:

$$\mathbb{E}[\hat{\rho}_i] = (d+1)\mathbb{E}[\hat{\sigma}_i] - \mathbb{I} = (d+1)\left(\frac{\rho + \mathbb{I}}{d+1}\right) - \mathbb{I} = \rho$$

Thus, from each randomized measurement on a single copy of the unknown state ρ, we can construct a matrix $\hat{\rho}_i$ that, on average, is equal to the true state ρ.

2.4 Prediction Phase

The real power of this formalism is not in reconstructing the full state ρ (which is still inefficient), but in efficiently predicting expectation values of observables. Given an observable O, we want to estimate $\mathrm{tr}(O\rho)$. We can define a set of scalar random variables X_i by taking the trace of the observable O with our single-shot snapshots:

$$X_i = \mathrm{Tr}(O\hat{\rho}_i)$$

The expectation value of each X_i is our desired quantity:

$$\mathbb{E}[X_i] = \mathbb{E}[\mathrm{Tr}(O\hat{\rho}_i)] = \mathrm{Tr}(O\mathbb{E}[\hat{\rho}_i]) = \mathrm{Tr}(O\rho)$$

This means we can estimate $\mathrm{tr}(O\rho)$ by simply aggregating the values of X_i (the aggregation can be simply taking the average of using the median-of-means estimator in the bonus lecture). The accuracy of this procedure depends on the variance of X_i.

2.5 Variance Analysis

Let's compute the variance of the random variable $X_i = \mathrm{Tr}(O\hat{\rho}_i)$. A crucial first step is to simplify the observable. It is convenient to work with the traceless part of O, defined as:

$$\tilde{O} = O - \frac{\mathrm{Tr}(O)}{d}\mathbb{I}$$

The single-shot estimator $\hat{\rho}_i$ is traceless by construction, as is the true state ρ. Therefore, their difference, $\hat{\rho}_i - \rho$, is also traceless. This means it is orthogonal to the identity operator, so $\mathrm{Tr}((\hat{\rho}_i - \rho)\mathbb{I}) = 0$. This allows us to substitute $\tilde{O}$ for O when analyzing the deviation from the mean:

$$X_i - \mathbb{E}[X_i] = \mathrm{Tr}(O(\hat{\rho}_i - \rho)) = \mathrm{Tr}\left(\left(\tilde{O} + \frac{\mathrm{Tr}(O)}{d}\mathbb{I}\right)(\hat{\rho}_i - \rho)\right) = \mathrm{Tr}(\tilde{O}(\hat{\rho}_i - \rho))$$

The variance is given by the expectation of the square of this deviation:

$$\mathrm{Var}[X_i] = \mathbb{E}[(X_i - \mathbb{E}[X_i])^2] = \mathbb{E}[(\mathrm{Tr}(\tilde{O}(\hat{\rho}_i - \rho)))^2]$$

We can bound the variance by the second moment, as the subtracted squared mean is non-negative:

$$\mathrm{Var}[X_i] = \mathbb{E}[(\mathrm{Tr}(\tilde{O}\hat{\rho}_i))^2] - (\mathrm{Tr}(\tilde{O}\rho))^2 \leq \mathbb{E}[(\mathrm{Tr}(\tilde{O}\hat{\rho}_i))^2]$$

Our main task is to compute this second moment. Let's expand the term inside the expectation. Recall that $\hat{\rho}_i = (d+1)U_i^\dagger|b_i\rangle\langle b_i|U_i - \mathbb{I}$ and $\mathrm{Tr}(\tilde{O}) = 0$.

$$\mathrm{Tr}(\tilde{O}\hat{\rho}_i) = (d+1)\mathrm{Tr}(\tilde{O}U_i^\dagger|b_i\rangle\langle b_i|U_i) = (d+1)\langle b_i|U_i\tilde{O}U_i^\dagger|b_i\rangle$$

The expectation $\mathbb{E}[\cdot]$ is taken over both the random unitary U_i and the random measurement outcome b_i. The probability of outcome b_i is $p(b_i|U_i) = \langle b_i|U_i\rho U_i^\dagger|b_i\rangle$. Combining these gives:

$$\mathbb{E}[(\mathrm{Tr}(\tilde{O}\hat{\rho}_i))^2] = (d+1)^2\mathbb{E}_{U_i}\left[\sum_{b_i} p(b_i|U_i)\left(\langle b_i|U_i\tilde{O}U_i^\dagger|b_i\rangle\right)^2\right] \tag{3}$$

$$= (d+1)^2\mathbb{E}_{U_i}\left[\sum_{b_i}\left(\langle b_i|U_i\rho U_i^\dagger|b_i\rangle\right)\cdot\left(\langle b_i|U_i\tilde{O}U_i^\dagger|b_i\rangle\right)\cdot\left(\langle b_i|U_i\tilde{O}U_i^\dagger|b_i\rangle\right)\right] \tag{4}$$

$$= (d+1)^2\cdot\sum_{b_i}\mathrm{Tr}\left((\rho\otimes\tilde{O}\otimes\tilde{O})\cdot\mathbb{E}_{U_i}\left[(U_i^\dagger|b_i\rangle\langle b_i|U_i)^{\otimes 3}\right]\right). \tag{5}$$

A simplified Weingarten calculus for for state t-design

The Weingarten calculus can be significantly simplified when we set the input matrix X to be $|\psi\rangle\langle\psi|^{\otimes t}$. This is given by the following theorem.

Theorem 1 (State t-design)**.** *Let $|\psi\rangle$ be any pure state in $\mathbb{C}^d$. The average of the t-th tensor power of its randomly rotated projector over the Haar measure is:*

$$\mathbb{E}_U\left[(U^\dagger|\psi\rangle\langle\psi|U)^{\otimes t}\right] = \frac{1}{(d+t-1)(d+t-2)\ldots(d+1)d}\cdot\sum_{\sigma\in S_t}\sigma.$$

Proof Hint. To prove this, start with the general form of the integral given by Schur-Weyl duality:

$$\mathbb{E}_U\left[(U^\dagger|\psi\rangle\langle\psi|U)^{\otimes t}\right] = \int U^{\otimes t}(|\psi\rangle\langle\psi|)^{\otimes t}U^{\dagger\otimes t}d\mu_H(U) = \sum_{\sigma\in S_t}C_\sigma\sigma$$

Now, use the key property of the operator $X = (|\psi\rangle\langle\psi|)^{\otimes t}$: it is symmetric under any permutation of its tensor factors. For any $\tau\in S_t$, we have $\tau X\tau^\dagger = X$. Use this symmetry, along with the invariance of the Haar measure, to show that the integral itself must be symmetric. This implies that the coefficients C_σ must be independent of σ, i.e., $C_\sigma = C$ for all σ. The constant of proportionality C can be found by taking the trace of both sides of the equation. □

Using this simplified Weingarten calculus for state t-design, we can evaluate the expression that upper bounds the variance: $(d+1)^2 \cdot \mathrm{Tr}\left((\rho \otimes \tilde{O} \otimes \tilde{O}) \cdot \mathbb{E}_{U_i}\left[(U_i^\dagger |b_i\rangle\langle b_i| U_i)^{\otimes 3}\right]\right)$ directly to obtain:

$$\mathbb{E}[(\mathrm{Tr}(\tilde{O}\hat{\rho}_i))^2] = \frac{d(d+1)^2}{d(d+1)(d+2)} \sum_{\sigma \in S_3} \mathrm{Tr}((\rho \otimes \tilde{O} \otimes \tilde{O})\sigma) = \frac{d+1}{d+2} \sum_{\sigma \in S_3} \mathrm{Tr}((\rho \otimes \tilde{O} \otimes \tilde{O})\sigma)$$

Evaluating the Trace Terms

We now evaluate the trace for each of the $3! = 6$ permutations σ in the symmetric group S_3. The key simplification comes from the fact that $\mathrm{Tr}(\tilde{O}) = 0$.

- $\sigma = \mathrm{id} = (1)(2)(3)$: $\mathrm{Tr}(\rho)\mathrm{Tr}(\tilde{O})\mathrm{Tr}(\tilde{O}) = 1 \cdot 0 \cdot 0 = 0$.
- $\sigma = (12)$: $\mathrm{Tr}(\rho\tilde{O})\mathrm{Tr}(\tilde{O}) = \mathrm{Tr}(\rho\tilde{O}) \cdot 0 = 0$.
- $\sigma = (13)$: $\mathrm{Tr}(\rho\tilde{O})\mathrm{Tr}(\tilde{O}) = \mathrm{Tr}(\rho\tilde{O}) \cdot 0 = 0$.
- $\sigma = (23)$: $\mathrm{Tr}(\rho)\mathrm{Tr}(\tilde{O}^2) = 1 \cdot \mathrm{Tr}(\tilde{O}^2) = \mathrm{Tr}(\tilde{O}^2)$. This term survives.
- $\sigma = (123)$: $\mathrm{Tr}(\rho\tilde{O}\tilde{O}) = \mathrm{Tr}(\rho\tilde{O}^2)$. This term survives.
- $\sigma = (132)$: $\mathrm{Tr}(\rho\tilde{O}\tilde{O}) = \mathrm{Tr}(\rho\tilde{O}^2)$. This term survives.

Summing the three non-zero terms, we get the exact expression for the second moment:

$$\mathbb{E}[(\mathrm{Tr}(\tilde{O}\hat{\rho}_i))^2] = \frac{d+1}{d+2}\left(\mathrm{Tr}(\tilde{O}^2) + 2\mathrm{Tr}(\rho\tilde{O}^2)\right).$$

To arrive at the final simple bound, we use two inequalities:

1. For a state ρ and positive semidefinite operator $\tilde{O}^2$, we have $\mathrm{Tr}(\rho\tilde{O}^2) \leq \mathrm{Tr}(\tilde{O}^2)$. This allows us to bound the sum:
$$\mathrm{Tr}(\tilde{O}^2) + 2\mathrm{Tr}(\rho\tilde{O}^2) \leq \mathrm{Tr}(\tilde{O}^2) + 2\mathrm{Tr}(\tilde{O}^2) = 3\mathrm{Tr}(\tilde{O}^2).$$
2. The squared Frobenius norm of the traceless part of an operator is never larger than that of the original operator:
$$\mathrm{Tr}(\tilde{O}^2) \leq \mathrm{Tr}(O^2).$$

Combining these results and noting that the prefactor $\frac{d+1}{d+2} \leq 1$, we get the final bound on the variance of a single-shot prediction:

$$\mathrm{Var}[X_i] \leq \mathbb{E}[(\mathrm{Tr}(\tilde{O}\hat{\rho}_i))^2] \leq 3\mathrm{Tr}(\tilde{O}^2) \leq 3\mathrm{Tr}(O^2)$$

This bound is the cornerstone of the classical shadow formalism. It is surprising because the variance of our estimator for $\mathrm{Tr}(O\rho)$ does not depend on the underlying state ρ or the system size n. It depends only on the Frobenius norm of the observable O we wish to predict.

3 Sample Complexity for Prediction

Now that we have established a robust, state-independent bound on the variance, $\mathrm{Var}[X_i] \leq 3\mathrm{Tr}(O^2)$, we can determine the number of samples N needed to predict $\mathrm{Tr}(O\rho)$ accurately and with low failure probability.

3.1 Standard Mean Estimator

The simplest approach is to use the sample mean, $\hat{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$. By applying Chebyshev's inequality with our variance bound (see the bonus lecture for a review on concentration inequality and median-of-means estimator), we get the following guarantee:

$$\Pr\left[|\hat{X} - \mathrm{Tr}(O\rho)| \geq \epsilon\right] \leq \frac{\mathrm{Var}[X_i]}{N\epsilon^2} \leq \frac{3\mathrm{Tr}(O^2)}{N\epsilon^2}$$

To make the failure probability at most δ, we must choose the number of samples N such that:

$$N \geq \frac{3\mathrm{Tr}(O^2)}{\epsilon^2\delta}$$

The dependence on $1/\delta$ is poor. To achieve a very low failure probability (e.g., $\delta = 10^{-6}$), this estimator requires an impractically large number of samples.

3.2 Median-of-Means Estimator for Robust Prediction

As detailed in the bonus lecture note, the median-of-means (MoM) estimator provides an exponentially better guarantee. By partitioning the N samples into blocks and taking the median of their means, the MoM estimator suppresses the influence of rare, large-deviation events.

The sample complexity for the MoM estimator to achieve an error of at most ϵ with failure probability at most δ is:

$$N = \mathcal{O}\left(\frac{\mathrm{Var}[X_i]}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$$

Substituting our state-independent variance bound, this becomes:

$$N = \mathcal{O}\left(\frac{\mathrm{Tr}(O^2)}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$$

The gentle logarithmic dependence on $1/\delta$ is a massive improvement, making it feasible to achieve extremely high confidence without an extremely large number of measurements.

3.3 The Classical Shadow Formalism

We can now combine our variance bound with the superior performance of the median-of-means estimator to state the main theorem of classical shadow tomography. The theorem addresses the practical task of predicting many observables from a single dataset.

Theorem 2 (Classical Shadow Tomography)**.** *Let $O_1, \ldots, O_M$ be a set of M observables, and let $B = \max_i \mathrm{Tr}(O_i^2)$. To predict the expectation value $\mathrm{Tr}(O_i\rho)$ for all M observables simultaneously from a single set of N classical shadows, up to an additive error ϵ and with a total success probability of at least $1-\delta$, it suffices to use N measurements, where:*

$$N = \mathcal{O}\left(\frac{B\log(M/\delta)}{\epsilon^2}\right)$$

Proof. The proof combines the median-of-means guarantee with a simple union bound.

1. **Single Observable Guarantee:** First, consider predicting a single observable O_i. The sample complexity required for the MoM estimator to fail with probability at most δ' is $N_1 = \mathcal{O}\left(\frac{\mathrm{Tr}(O_i^2)}{\epsilon^2}\log(\frac{1}{\delta'})\right)$. Using our variance bound, this becomes $N_1 = \mathcal{O}\left(\frac{B}{\epsilon^2}\log(\frac{1}{\delta'})\right)$.

2. **Union Bound:** To ensure that *all* M predictions are accurate, we require that none of them fail. The probability that at least one prediction fails is bounded by the sum of individual failure probabilities:

$$\Pr[\text{any failure}] = \Pr[\cup_{i=1}^{M} \text{failure}_i] \leq \sum_{i=1}^{M} \Pr[\text{failure}_i]$$

We set the desired total failure probability to δ. To achieve this, we can require the failure probability for each individual observable to be much smaller, namely $\delta' = \delta/M$. The total failure probability is then bounded by $\sum_{i=1}^{M} \delta' = M(\delta/M) = \delta$.

3. **Final Sample Complexity:** We substitute this individual failure probability $\delta' = \delta/M$ back into the sample complexity formula for a single observable. This gives the required number of samples N to guarantee success for all M observables:

$$N = \mathcal{O}\left(\frac{B}{\epsilon^2} \log\left(\frac{1}{\delta'}\right)\right) = \mathcal{O}\left(\frac{B}{\epsilon^2} \log\left(\frac{M}{\delta}\right)\right)$$

This completes the proof. □

Many of us found this result surprising when we first realized it. At the beginning of the lecture, we saw that learning a full description of the state ρ to accurately predict any observable required $N = \mathcal{O}(16^n)$ measurements. The classical shadow protocol achieves a similar goal with a sample complexity that scales only **logarithmically** with the number of observables M and completely independent from the number of qubits n. In contrast, if one directly measures all the observables $O_1, \ldots, O_M$, that would require $\mathcal{O}(M \log(M/\delta))/\epsilon^2$. So the classical shadow formalism exponentially improves the scaling with respect to the number of observables M we wish to predict. However, this comes at a cost: the observables must have bounded Frobenius norm $\text{Tr}(O^2)$ (e.g., even $Z^{\otimes n}$ has an exponentially large Frobenius norm). In the next lecture, we will look at how to handle the general case when $O_1, \ldots, O_M$ can be arbitrary.