CS 298 Lecture 4 (9/23/16)
Professor: Umesh Vazirani
Scribes: Jeffrey Epstein, Roy Kun Tu

Quantum Techniques for Classical Functions

Suppose that we would like to know something about a "classical" function $f : X \to Y$. We've seen that we can construct two natural unitary operators from f:

$$\begin{aligned} U_f &: |x\rangle \otimes |y\rangle \mapsto |x\rangle \otimes |y \oplus f(x)\rangle \\ (-1)^f &: |x\rangle \mapsto (-1)^{f(x)} |x\rangle . \end{aligned} \tag{1}$$

We can actually build the second operator from the first and some elementary gates by first applying U_f, then applying a phase gate to the second register (Z if $Y = \{0,1\}$), and then applying U_f again. An application of U_f is called a query of an oracle, and in this context we're often interested in the minimum number of queries needed to answer some question about f.

A naive hope is that we can learn a lot about f with a single call to the oracle, say by feeding it the uniform superposition over inputs:

$$|X|^{-1/2} \sum_{x \in X} |x\rangle \otimes |0\rangle \overset{U_f}{\longmapsto} |X|^{-1/2} \sum_{x \in X} |x\rangle \otimes |f(x)\rangle . \tag{2}$$

Now we've evaluated f on all possible inputs, or taken all possible paths through the computation, or computed in parallel universes, or something! But to get any information out, we have to make a measurement, and the amount of information we can extract is limited[1]. The simplest thing we could do is to measure both registers in the computational basis. Then we get the state $|x\rangle |f(x)\rangle$ with probability $1/|X|$. This is even worse than the classical strategy of feeding each value of x to the oracle one at a time to learn the whole function, since we might end up with redundant outcomes. Any hope for a quantum speedup can't depend on this technique.

Instead of this generally fruitless approach, we can identify two useful techniques:

1. Project onto a level set of f by measuring the output register:

$$\sum_{x \in X} \alpha_x |x\rangle \otimes |0\rangle \overset{U_f}{\longmapsto} \sum_{x \in X} \alpha_x |x\rangle \otimes |f(x)\rangle \overset{\Pi_a}{\longmapsto} \left\{ \sum_{x \in X : f(x) = a} |\alpha_x|^2 , \sum_{x \in X : f(x) = a} \alpha_x |x\rangle \otimes |a\rangle \right\} \tag{3}$$

   where the notation $\{p_i, |\psi_i\rangle\}$ refers to the ensemble of preparing state $|\psi_i\rangle$ with probability p_i. The states are left unnormalized for compactness.

2. Imprint f on an input state via a phase factor:

$$\sum_{x \in X} \alpha_x |x\rangle \overset{(-1)^f}{\longmapsto} \sum_{x \in X} \alpha_x (-1)^{f(x)} |x\rangle \tag{4}$$

   where here for simplicity $\text{Im} f = \{0,1\}$.

Lower Bound on Quantum Search Query Complexity

Can we use quantum mechanics to speed up the process of searching for a marked element in a set? One way to formalize this question is to consider a discrete set X and the maps $f_a : X \mapsto \{0,1\}$ where $f_a(x) = \delta_a(x)$ for $a \in X$, i.e. the indicator functions on the one-element subsets of X (we'll assume exactly one element is marked). If we're given a black box that can evaluate f_a on inputs $x \in X$, the search problem becomes the

[1] Holevo's theorem makes this precise.

problem of determining a, and we'd like a lower bound on the number of times we have to use the black box (query the oracle) to get a correct answer with high probability.

Classically, each call to the oracle takes an input x and gives us $f(x)$, so all we can do is repeatedly ask whether the statement $a = a'$ is true for different values of the guess $a' \in X$. We might get lucky and get $f(a') = 1$, i.e. $a' = a$ the first time. But we might get unlucky and run through all $|X|$ possible inputs before hitting on the right answer. Therefore $\mathcal{O}(|X|)$ is a lower bound for the number of calls to the oracle we have to make (in the worst case). Another way to see this is to note that the random variable encoding the value of a has Shannon entropy $\log|X|$ (assuming a is drawn uniformly at random) and the random variable $f_a(x)$ has entropy

$$H = \frac{1}{|X|}\log|X| + \frac{|X|-1}{|X|}\log\frac{|X|}{|X|-1}. \tag{5}$$

For $|X| \gg 1$, this goes as $\log|X| \,/\, |X|$, indicating that we need at least $\mathcal{O}(|X|)$ calls to the oracle to get all of the information about a.

In the quantum setting, the oracle should become a unitary operator. Let $\{\psi_x\}_{x\in X}$ be an orthonormal set of states spanning some Hilbert space $\mathcal{H}_1$, and let $\mathcal{H}_2$ be some other finite-dimensional Hilbert space. For some function $f : X \mapsto \{0,1\}$, define

$$U_f : \psi_x \otimes \phi \mapsto \psi_x \otimes W_{f(x)}\phi \tag{6}$$

where W_0 and W_1 are unitary operators on $\mathcal{H}_2$. This definition extends by linearity to a unique unitary operator on $\mathcal{H}_1 \otimes \mathcal{H}_2$, i.e. on the input register and any ancilla system we might want to include in the computation:

$$U_f = \sum_{x\in X} \Pi_x \otimes W_{f(x)} \tag{7}$$

where Π_x is the projector onto the subspace of $\mathcal{H}_1$ spanned by ψ_x. If we let $\mathcal{H}_2 = \mathbb{C}^2$, $W_0 = \mathbb{1}$, and $W_1 = X$, then we have

$$U_f : \psi_f \otimes |0\rangle \mapsto \psi_f \otimes |f(x)\rangle\,. \tag{8}$$

Preparing the state $\psi_f \otimes |0\rangle$, applying U_f, and measuring the ancilla qubit is then equivalent to querying a classical oracle for f, so this is an appropriate generalization of the classical problem. Equivalently, we could see the classical problem as adding to the quantum problem the restriction that we're only allowed to apply U_f to basis states ψ_f as opposed to arbitrary superpositions.

In this framework we can ask how many calls to a quantum oracle (applications of U_f) we need to determine a, given that we're promised $f = f_a$ for some $a \in X$ and can apply any unitary operation we like. The strategy of the proof of the lower bound will be to note that we want to be able to distinguish the output of the algorithm in the cases $f = f_a$ and $f = f_b$, given that $a \neq b$ and to try to upper bound the norm distance between these two states by some quantity depending on the number of calls to the oracle.

Theorem 1. *Any quantum algorithm that can reliably distinguish f_a from f_b for any $a, b \in X$, $a \neq b$ must involve $\mathcal{O}(|X|^{1/2})$ queries.*

Proof. The output of an algorithm that uses t calls to the oracle may be represented as

$$\psi_{\text{out}} = V_t U_f V_{t-1} U_f \dots V_1 U_f \psi \tag{9}$$

where we're free to choose the input state ψ and the unitary gates V_i. These choices specify the algorithm. We can bound the difference Δ_{out} between the output states of this algorithm in the cases $f = f_a$ and $f = f_b$ by imagining comparing the outputs of the algorithm using U_{f_a} for all queries except the last, which uses U_{f_b}, then the outputs of the algorithm that calls U_{f_a} for all but the last two and the algorithm that calls U_{f_a} for all but the last, and so on. This is the "hybrid argument" mentioned in lecture. Formally, it amounts to repeated application of the triangle inequality for norms. For compactness, I'll write $U_a = U_{f_a}$:

$$\begin{aligned}
\|\Delta_{\text{out}}\| &= \|V_t U_a V_{t-1} U_a \ldots V_1 U_a \psi - V_t U_b V_{t-1} U_b \ldots V_1 U_b \psi\| \\
&\le \|V_t U_a V_{t-1} U_a \ldots V_2 U_a V_1 U_a \psi - V_t U_a V_{t-1} U_a \ldots V_2 U_a V_1 U_b \psi\| \\
&\quad + \|V_t U_a V_{t-1} U_a \ldots V_2 U_a V_1 U_b \psi - V_t U_a V_{t-1} U_a \ldots V_2 U_b V_1 U_b \psi\| \\
&\qquad + \ldots \\
&\qquad\quad + \|V_t U_a V_{t-1} U_b \ldots V_2 U_b V_1 U_b \psi - V_t U_b V_{t-1} U_b \ldots V_2 U_b V_1 U_b \psi\| \\
&= \|U_a \psi - U_b \psi\| \\
&\quad + \|U_a V_1 U_b \psi - U_b V_1 U_b \psi\| \\
&\qquad + \ldots \\
&\qquad\quad + \|U_a V_{t-1} U_b \ldots V_2 U_b V_1 U_b \psi - U_b V_{t-1} U_b \ldots V_2 U_b V_1 U_b \psi\| \qquad (10) \\
&= \|U_a \psi^{(0)} - U_b \psi^{(0)}\| + \|U_a \psi^{(1)} - U_b \psi^{(1)}\| + \ldots + \|U_a \psi^{(t-1)} - U_b \psi^{(t-1)}\| \\
&= \sum_{i=0}^{t-1} \|U_a \psi^{(i)} - U_b \psi^{(i)}\|.
\end{aligned}$$

The first inequality is the triangle inequality, the following equality is due to the unitary invariance of the norm, and the next is just the introduction of some notation. Now we can apply the Cauchy-Schwarz inequality to find:

$$\|\Delta_{\text{out}}\|^2 = \left(\sum_{i=0}^{t-1} \|U_a \psi^{(i)} - U_b \psi^{(i)}\| \right)^2 \le t \sum_{i=0}^{t-1} \|U_a \psi^{(i)} - U_b \psi^{(i)}\|^2. \qquad (11)$$

Now suppose that we draw $a, b \in X$ uniformly and independently at random. The expectation of the norm difference squared is then:

$$\begin{aligned}
\mathbb{E}\left[\|\Delta_{\text{out}}\|^2\right] &\le \frac{t}{|X|^2} \sum_{a,b \in X} \sum_{i=0}^{t-1} \|U_a \psi^{(i)} - U_b \psi^{(i)}\|^2 \\
&= \frac{t}{|X|^2} \sum_{i=0}^{t-1} \sum_{a,b \in X} \| \sum_{y \in X} \left(\Pi_y \otimes W_{\delta_a(y)} - \Pi_y \otimes W_{\delta_b(y)} \right) \psi^{(i)} \|^2 \\
&= \frac{t}{|X|^2} \sum_{i=0}^{t-1} \sum_{a,b \in X} \|\Pi_a \otimes (W_1 - W_0)\, \psi^{(i)} - \Pi_b \otimes (W_1 - W_0)\, \psi^{(i)}\|^2 \\
&\le \frac{t}{|X|^2} \sum_{i=0}^{t-1} \sum_{a,b \in X} \Big(\|\Pi_a \otimes (W_1 - W_0)\, \psi^{(i)}\|^2 + \|\Pi_b \otimes (W_1 - W_0)\, \psi^{(i)}\|^2 \Big) \qquad (12) \\
&= \frac{t}{|X|^2} \sum_{i=0}^{t-1} \left(\sum_{b \in X} \| \sum_{a \in X} \Pi_a \otimes (W_1 - W_0)\, \psi^{(i)}\|^2 + \sum_{a \in X} \| \sum_{b \in X} \Pi_b \otimes (W_1 - W_0)\, \psi^{(i)}\|^2 \right) \\
&= \frac{t}{|X|^2} \sum_{i=0}^{t-1} \left(\sum_{b \in X} \|\mathbb{1} \otimes (W_1 - W_0)\, \psi^{(i)}\|^2 + \sum_{a \in X} \|\mathbb{1} \otimes (W_1 - W_0)\, \psi^{(i)}\|^2 \right) \\
&\le \frac{t}{|X|^2} \sum_{i=0}^{t-1} (4\,|X| + 4\,|X|) \\
&= \frac{8t^2}{|X|}.
\end{aligned}$$

Here we've used the Pythagorean theorem to pull sums inside norms. The first inequality is due to the fact that the terms in the double sum with $a = b$ actually vanish. The triangle inequality and the unitary invariance of the norm give the second inequality. From this result we can conclude that there is some pair of elements $a, b \in X$ for which

$$\|\Delta_{\text{out}}\| \leq \frac{2\sqrt{2}t}{|X|^{1/2}}. \tag{13}$$

If ψ_a and ψ_b are the outputs of the algorithm given that $f = f_a$ and $f = f_b$, respectively, then

$$\|\Delta_{\text{out}}\| = \|\psi_a - \psi_b\| = \langle \psi_a - \psi_b, \psi_a - \psi_b \rangle \geq 2 - 2\left|\langle \psi_a, \psi_b \rangle\right|. \tag{14}$$

Then there are some pair of choices $a, b \in X$ such that the fidelity between the two output states is lower bounded:

$$|\langle \psi_a, \psi_b \rangle| \geq 1 - \frac{\sqrt{2}t}{|X|^{1/2}}. \tag{15}$$

Therefore we need $t = \mathcal{O}(|X|^{1/2})$ for the fidelity to be close to zero, i.e., for the two output states to be distinguishable with high probability. In more operational terms, if you give me a family of algorithms with that uses fewer than $\mathcal{O}(|X|^{1/2})$ queries, then for large $|X|$, I can choose to mark an item $x \in X$ that you will, at least almost half of the time, fail to identify correctly. □

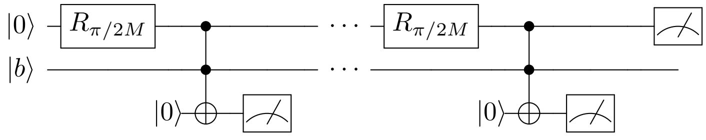
This quantum lower bound of $\mathcal{O}(|X|^{1/2})$ is better than the classical lower bound of $\mathcal{O}(|X|^{1/2})$. We'll see in a bit that there is in fact an algorithm that achieves the quantum lower bound, so quantum mechanics does offer some speedup. Usually, we think of $X = \{0,1\}^n$, so that $|X| = 2^n$. Then in terms of n, we have $\mathcal{O}(2^{n/2})$ instead of $\mathcal{O}(2^n)$. In terms of n, both lower bounds are exponential, so the quantum speedup isn't quite the miracle we hoped.

Elitzur-Vaidman Bomb Detector

The Elitzur-Vaidman bomb detection experiment imagines a scenario in which we need to determine whether or not a bomb is a dud. The catch - the working bombs have such sensitive triggers that any interaction with the bombs (i.e., any attempt to measure them) will set them off. Can we *safely* conclude one way or the other?

To answer this question, we'll have to make some kind of model of the situation. Let's say the bomb is a qubit (it could also be taken here just to be a classical bit) which is either in the state $|0\rangle$ (a dud) or the state $|1\rangle$ (a live bomb). To capture the notion of a measurement setting off a live bomb, we'll do the following: suppose that we take an ancilla qubit initialized in the state $|0\rangle$, perform a CNOT from the bomb to the ancilla, and then measure the ancilla. If we measure 0, then the bomb is a dud, and we're safe. If we measure 1, then the bomb was functional, and we consider that our lab has been blown up.

Clearly this approach doesn't satisfy our desire for a *safe* procedure, since any time we find that the bomb is live, we also set it off. But consider the following circuit:



There are M applications of the basic unit of rotation, CCNOT, and measurement, and $R_{\pi/2M}$ is a rotation by $\frac{\pi}{2M}$ in the $|0\rangle - |1\rangle$ plane, i.e. the unitary operator $\exp(i\pi X/2M)$, with $X = |0\rangle\langle 1| + |1\rangle\langle 0|$. Let's examine what happens in the two cases:

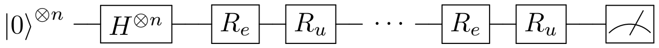
Case 1: The bomb is a dud, and $b = 0$. In this case, none of the CCNOT gates are performed, so the only action of the circuit is to perform a rotation by $\pi/2$ on the first qubit. Then we will always measure the first qubit to be in the state $|1\rangle$. Since the bit flip is never applied to the ancillae qubits, we measure all of them to be in the state $|0\rangle$, and no explosion occurs.

Case 2: The bomb is live, and $b = 1$. If any of the ancilla measurements yields $|1\rangle$, the bomb detonates, so we're only interested in the case in which all of these measurements yield $|0\rangle$. Since the CNOT (the second C in CCNOT is always on) decoheres the first qubit, we can think of it as being probabilistically flipped back and forth from $|0\rangle$ to $|1\rangle$ as each unit of the circuit is applied (from the point of view of the first qubit, the CNOT is a measurement). The probability of a bit flip is $\sin^2(\pi/2M)$. So the probability that no bit flip occurs, which is the same as the probability that no CNOT is ever applied, and thus that no ancilla measurement causes an explosion, is $\cos^{2M}(\pi/2M)$. This probability goes to 1 as $M \to \infty$ (it takes $M = 123$ to achieve a probability of greater than 98%). For large M, $\cos^{2M}(\pi/2M) \approx 1 - \pi^2/4M$. Note that in this case, we also measure the first qubit to be in the state $|0\rangle$.[2]

Since this protocol, with high probability ($1 - \mathcal{O}(1/M)$), gives us a way to determine if the bomb is live or not (by checking the outcome of the measurement of the first qubit) without setting off an explosion, it is a safe bomb detection test of the kind we wanted. Whether this model captures the fundamental nature of the question is a question for another time. For a start, is conditioning on the state of the bomb something that can be done "without interacting" with it? This thought experiment is often discussed in the context of quantum optics, and that may be a more compelling analogue from this point of view.

Grover's Algorithm

Suppose we're given a black box/oracle of the type described above that evaluates a function $f : \{0,1\}^n \to \{0,1\}$, and promised that there is a unique $a \in \{0,1\}^n$ such that $f(a) = 1$. Grover's algorithm for finding this value of a is implemented by the following circuit:



where the gates R_e and R_u are the unitary operators defined by

$$\begin{aligned} R_e &: |x\rangle \mapsto (-1)^{\delta_a(x)} |x\rangle \\ R_u &: |k\rangle_f \mapsto -(-1)^{\delta_0(k)} |k\rangle_f \end{aligned} \tag{16}$$

where $|x\rangle$ are the computational basis elements (this is the basis we're measuring in) and $|k\rangle_f$ are the elements of the Fourier basis, i.e. $|k\rangle_f = H^{\otimes n} |k\rangle$. The block consisting of the two R gates is repeated $\mathcal{O}\left(2^{n/2}\right)$ times. We've seen that R_e can be implemented with two calls to the oracle.

So why does this circuit do what is claimed? Consider the two-dimensional subspace of the 2^n-dimensional Hilbert space spanned by $|a\rangle$ and $|u\rangle = H^{\otimes n} |0\rangle^{\otimes n} = |+\rangle^{\otimes n} = |0\rangle_k$. In this subspace, define

$$|e\rangle = \frac{(\mathbb{1} - |a\rangle \langle a|) |u\rangle}{\sqrt{\langle u| (\mathbb{1} - |a\rangle \langle a|) |u\rangle}} \tag{17}$$

so that $|a\rangle$ and $|e\rangle$ form an orthonormal basis. In this subspace, R_e performs a reflection across the line spanned by $|e\rangle$, and R_u performs a rotation across the line spanned by $|u\rangle$. Recall that the composition of two reflections is a rotation by twice the angle between the two axes of reflection. In this case, the angle is

$$\begin{aligned} \cos\theta = \langle u|e\rangle = \frac{\langle u| (\mathbb{1} - |a\rangle \langle a|) |u\rangle}{\sqrt{\langle u| (\mathbb{1} - |a\rangle \langle a|) |u\rangle}} &= \sqrt{\langle u| (\mathbb{1} - |a\rangle \langle a|) |u\rangle} = \sqrt{2^{-n} \sum_{x,y\in\{0,1\}^n} \langle x| (\mathbb{1} - |a\rangle \langle a|) |y\rangle} \\ = \sqrt{2^{-n} \sum_{x,y,z\in\{0,1\}^n; z\neq a} \langle x|z\rangle \langle z|y\rangle} &= \sqrt{2^{-n} \sum_{x\in\{0,1\}^n; x\neq a} 1} = \sqrt{\frac{2^n - 1}{2^n}} = \sqrt{1 - 2^{-n}}. \end{aligned} \tag{18}$$

[2] Turning this thought experiment on its head, we can think about trying to "freeze" the first qubit in the state $|0\rangle$ even while successive rotations are applied. This analysis shows that a way to do this is by making measurements between each rotation, with the probability of successful freezing going to one as the measurements become more and more frequent. This is the quantum Zeno effect.

Rearranging, we find $\sin\theta = 2^{-n/2}$. For $n \gg 1$, then $\theta \approx 2^{-n/2}$. Since for large n, we also have $|u\rangle \approx |e\rangle$, after $\mathcal{O}\left(2^{n/2}\right)$ applications of the two rotation gates, we will have rotated the input state $|u\rangle$ by an angle of $\mathcal{O}(1)$ towards $|a\rangle$. Then measuring in the computational basis, with high probability we get the outcome a. If we demand that we rotate by exactly $\pi/2$, so that we always measure outcome a, we'll have to be slightly more careful about keeping constant multiplicative factors around, and to be really careful we should remember that the unitary R_e required *two* calls to the oracle. But this doesn't change the scaling behavior, so is irrelevant from a complexity point of view.

Shor's Algorithm

Given a number that can be decomposed into two primes $N = N_1 N_2$, how do we factor N? We know that factoring is not NP-complete (factoring $\in$ NP $\cap$ CoNP) but it is still very difficult classically.
Naively, we can divide by all primes from $1 \ldots \sqrt{N}$, but this is polynomial in N (which is often very large, $\approx 2^n$). The most efficient classical algorithm is sub-exponential, $O(e^{1.9(\log N)^{1/3}(\log\log N)^{2/3}})$. Shor's algorithm is $O((\log N)^{1/3}\log\log N)$, which is significantly faster.

Non-trivial Square Roots over $\mathbb{F}_p$

We consider a different problem that, when solved, lets us solve the factoring problem quickly:

Given a natural number N which is odd, not a prime, and not a prime factor, find $x \neq \pm 1$ such that: $x^2 \equiv 1 \pmod{N}$. That is, find a non-trivial square root over $\mathbb{F}_p$.

We show how this can be used to solve the factoring problem. Assume we have solved x for a particular N. Then:

$$\begin{aligned} & x^2 \equiv 1 \pmod{N} \\ \implies & x^2 - 1 \equiv 0 \pmod{N} \\ \implies & (x+1)(x-1) \equiv 0 \pmod{N} \\ \implies & (x+1)(x-1) = cN \end{aligned}$$

and $\gcd(x+1, N)$ is a factor of N. Hence, factoring N reduces to the problem of finding non-trivial square roots over $\mathbb{F}_{\mathbb{N}}$.

Period-finding

We consider yet another problem, which allows us to solve the non-trivial square root problem with high probability:

Given $N \in \mathbb{N}, x$ find r such that $x^r \equiv 1 \pmod{N}$.

If we can solve this problem quickly, then we can solve for non-trivial square roots quickly. Here's how:

1. Pick a random x and solve for r such that $x^r \equiv 1 \pmod{N}$.
2. If r is even, then $x^{r/2}$ is a non-trivial square root. Proof:

$$x^r \equiv 1 \pmod{N} \implies (x^{r/2})^2 \equiv 1 \pmod{N}$$

If r is odd, or if $x \equiv 1 \pmod{N}$, we choose a different x until r is even and x is non-trivial. We assert – and it can be proved that – if we keep choosing random x's, the probability that r is even converges to 1. So now our concern becomes solving for r (called the **order**) given an N, x. Here's a quantum algorithm: Consider the function $f : a \to x^a \pmod{N}$. Note that this function is periodic, with period r:

$$\begin{aligned} & x^0 \equiv x^r \equiv x^{2r} \equiv \ldots \equiv 1 \pmod{N} \\ & x^1 \equiv x^{r+1} \equiv x^{2r+1} \equiv \ldots \equiv x \pmod{N} \\ & x^2 \equiv x^{r+2} \equiv x^{2r+2} \equiv \ldots \equiv x^2 \pmod{N} \end{aligned}$$

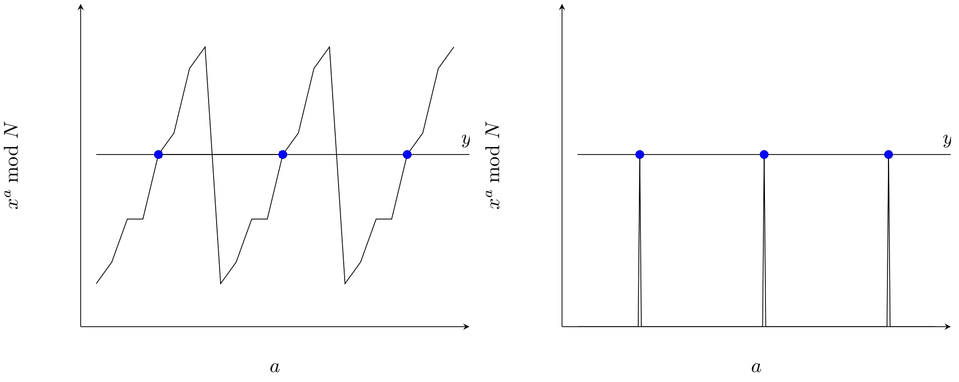


Figure 1: *Left*: Before measurement. Measurement selects some subspace of inputs a where $f(a) = y$, collapsing the function into a Dirac comb with nonzero values at the marked a values. *Right*: After measurement.

The algorithm goes:

1. Start with a superposition of all possible orders:

$$\sum_a |a\rangle |0\rangle \xrightarrow{f} \sum_a |a\rangle |x^a \bmod N\rangle$$

2. Measure the second subspace. This collapses the first subspace into all possible preimages:

$$\sum_a |a\rangle |x^a \bmod N\rangle \xrightarrow{\text{measure}} \sum_{a:f(a)=y} |a\rangle |y\rangle$$

Notice that the elements of this space have period r! That is, the first subspace contains kets like $|c\rangle, |c+r\rangle, |c+2r\rangle \ldots$ with some offset c. If we measure in this subspace right now, we actually get no information, since we have no control over this offset. But fortunately we have a way of removing the offset, with the following lemma:

Lemma: If $\text{DFT}_M(\alpha_0, \alpha_1, \ldots, \alpha_{m-1}) = (\beta_0, \beta_1, \ldots, \beta_{m-1})$ then a cyclic permutation of the α's corresponds to a phase factor multiplied to the β's. That is, if ω is the root of unity:

$$\text{DFT}_M \begin{pmatrix} \alpha_{m-1} \\ \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{m-2} \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \omega\beta_1 \\ \omega^2\beta_2 \\ \vdots \\ \omega^{m-1}\beta_{m-1} \end{pmatrix}$$

The crucial point is, if we take a Fourier transformation, the cyclic offset manifests itself only in phase factors on the amplitudes, which don't affect measurement!

This completes the algorithm from a high level, but there is one detail we ignored. If we take DFT_M of a Dirac comb with some offset c and some period r, then the output is another Dirac comb with period M/r only if M divides r! Since in general this is not the case, we have to consider the case where $r \nmid M$.
First, we consider the case where $r \nmid M$. We assume that the offset is 0 for simplicity (the proof can be easily generalized). Then:

$$\frac{1}{\sqrt{M/r}} \sum_{j=0}^{M/r-1} |jr\rangle \xrightarrow{DFT_M} \frac{1}{\sqrt{M/r}} \sum_{j=0}^{M/r-1} \left[\frac{1}{\sqrt{M}} \sum_{y=0}^{M-1} \omega^{jry} |y\rangle \right] \equiv \sum_{y=0}^{M-1} \alpha_y |y\rangle$$

with $\alpha_y \equiv \sum_{j=0}^{M/r-1} \frac{\sqrt{r}}{M}\omega^{jry}$. If $(M/r) \nmid y$, $\omega^{jry} = \omega^{jMc} = 1$ for some $c \in \mathbb{N}$. Otherwise, $\alpha_y = 0$. Hence we get constructive interference at any multiple of M/r, and destructive interference everywhere else – a perfect Dirac comb.

What if $r \nmid M$? Assume M is large, $M \gg 2N^2$ (this is sensible – we need many cycles to be sampled in order to find our period!) Then, we get constructive interference iff $|ry \bmod M| \leq r/2$. Intuitively, this constraint means the vectors don't deviate more than 180 degrees from each other, and hence do not interfere destructively. If this constraint is satisfied, how do we find r? The problem is as follows:

Given y, M, determine r such that $|ry - cM| \leq r/2$ for some $c \in \mathbb{N}$.

This constraint implies:

$$|ry - cM| \leq r/2 \implies |\frac{y}{M} - \frac{c}{r}| \leq \frac{1}{2M}$$

This bounds the difference between two vectors with different $c, c' \in \mathbb{N}$ as follows:

$$|\left(\frac{y}{M} - \frac{c}{r}\right) - \left(\frac{y}{M} - \frac{c'}{r}\right)| = |\frac{c}{r} - \frac{c'}{r}| \leq \frac{1}{M} \leq \frac{1}{2N^2}$$

where we used the triangle inequality. Since $\frac{1}{2N^2}$ is small, the interference pattern closely approximates that of a Dirac comb with period M/r and we can determine r with a technique like continued fraction expansion.