1 Last lecture, and plan for today's lecture

Last lecture, we started discussing two related questions:

- given a secure PRF $F : \{0,1\}^{\ell_{\text{in}}} \times \{0,1\}^n \to \{0,1\}^{\ell_{\text{out}}}$, can we construct a secure PRP (with some domain and key space)? We saw a few attempts for this.
- given a secure PRP $P : \{0,1\}^{\ell_{\text{in}}} \times \{0,1\}^n \to \{0,1\}^{\ell_{\text{in}}}$, we know (by definition) that $P(\cdot, k)$, for a random k, 'looks like' a random permutation. Is P also a secure PRF (that is, is P also indistinguishable from a random function from $\{0,1\}^{\ell_{\text{in}}} \to \{0,1\}^{\ell_{\text{in}}}$)?

  Here, we proved the 'birthday bound', which says that if t numbers are sampled from $[N]$, u.a.r. and independently, then the probability of at least two samples being same is $\Theta(t^2/N)$. This will be useful for proving that a random function is indistinguishable from a random permutation.

2 PRFs vs PRPs

2.1 PRPs from PRFs

Given a secure PRF $F : \{0,1\}^{\ell_{\text{in}}} \times \{0,1\}^n \to \{0,1\}^{\ell_{\text{out}}}$, we can construct a permutation $P : \{0,1\}^{\ell_{\text{in}}+\ell_{\text{out}}} \times \{0,1\}^{3n} \to \{0,1\}^{\ell_{\text{in}}+\ell_{\text{out}}}$. This construction is very similar to Attempt 3 discussed in last class, except that we will use different keys each time we use P_1. More formally, consider the function $P(\cdot, (k_1, k_2, k_3)) \equiv P_1(\cdot, k_3) \circ \mathsf{swap} \circ P_1(\cdot, k_2) \circ \mathsf{swap} \circ P_1(\cdot, k_1)$. Pictorially, this is shown in Figure 1 below.

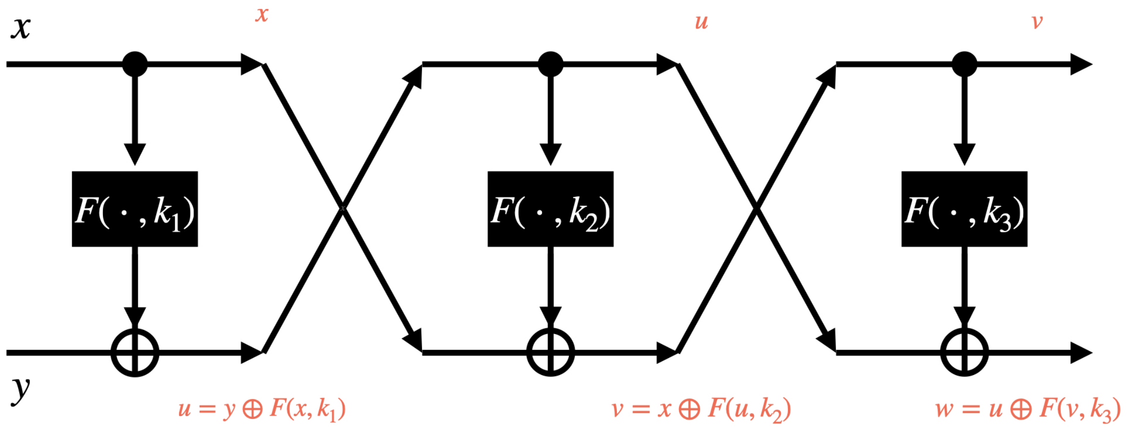

Figure 1: Implementation of keyed permutation P. The keyspace is $\{0,1\}^n$ here, input/output space is $\{0,1\}^{\ell_{\text{in}}+\ell_{\text{out}}}$ and it maps (x, y) to (v, w).

First, let us check that this is indeed a permutation. Check that the inverse of this function is $P_1(\cdot, k_1) \circ \mathsf{swap} \circ P_1(\cdot, k_2) \circ \mathsf{swap} \circ P_1(\cdot, k_3)$. As a sanity check, you should check that $P \circ P$ is not equal to identity.

Qn: Why do the previous attacks not work?

Ans: First, note that $P \circ P$ is not identity, this was the attack used to show that Attempt 3 is not a secure PRP. Next, recall the attack on Attempt 2 : we queried on inputs (x, y) and $(x, y \oplus 1^{\ell_{\text{out}}})$, and the reason this attack worked is because part of the output was of the form $y \oplus \text{some function of } x$. As a result, we could make changes to y and see the exact same changes in one of the output component. But now, if we apply P_1 thrice with different keys, check that the two output components are of the form $x \oplus \text{some function of } (x, y)$ and $y \oplus \text{some other function of } (x, y)$

This construction was proposed by Michael Luby and Charles Rackoff in 1988, and one can prove that if F is a secure PRF, then P is a secure PRP. The proof is not very complicated (and the theoretically inclined students are encouraged to attempt it). However, in the interest of time, we will not discuss the security proof in class.

2.1.1 Other attempts discussed in class

The following are some of the other candidates proposed in class:

- $P_1(\cdot, k) \circ \mathsf{swap} \circ P_1(\cdot, k) \circ P_1(\cdot, k)$: this will not be a secure PRP, because $P_1(\cdot, k) \circ P_1(\cdot, k)$ is identity, and therefore this function is simply $P_1(\cdot, k) \circ \mathsf{swap}$, which can be distinguished from a uniformly random permutation. Why?
- $P_1(\cdot, k) \circ \pi \circ \mathsf{swap} \circ P_1(\cdot, k) \circ \mathsf{swap} \circ P_1(\cdot, k)$ where π is some public permutation: this depends on the structure of π. For certain permutations (such as π being the identity function, or $\pi = \mathsf{swap}$), the resulting scheme will not be a secure PRP. However, if π is sufficiently complicated, then this might work.

2.2 PRFs from PRPs

Last lecture, we discussed the 'birthday bound'. Today, we will use it for showing that a random permutation with domain $\{0, 1\}^n$ looks like a random function if you're given only $\mathsf{poly}(n)$ queries.

Claim 12.01 (Birthday Bound).

$$\Pr\left[\text{There exist repetitions when sampling } t \text{ numbers, u.a.r. with replacement}\right] = \Theta\left(\frac{t^2}{N}\right)$$

Back to our PRP vs PRF problem: we would like to show that no (polynomial time) adversary can distinguish between a uniformly random permutation and a uniformly random function, given only polynomially many queries. To show this, we will introduce a new notion called the 'statistical distance' between two probability distributions, and use it to show that two distributions are indistinguishable.

2.2.1 Distinguishing two probability distributions

Consider the following toy problem: there are two coins. The first coin is a fair coin, and the second coin outputs H with probability $2/3$. I pick one of these two coins u.a.r., and tell you the outcome of the toss. Your job is to guess whether I picked the fair coin or the biased coin. Clearly, your strategy is simple: if the outcome was H, you will guess that I tossed the biased coin. Otherwise, you will guess that I tossed the fair coin. Using this strategy, your guess will be correct with probability $1/4 + 1/3$.

Let us now extend this problem: there are two dice. The first one is a fair dice (that is, all six numbers show up with equal probability). For the second dice, the probabilities are given by the following 'probability vector': $(1/2, 0, 1/6, 1/3, 0, 0)$. Again, I pick one of the two dice at random, roll it, and tell you the outcome. You have to guess whether I rolled the fair dice or the biased dice. As was suggested in class, to maximise your correctness probability, you would use the following strategy (or some close variant of it):

Sample	Your guess
1	biased dice
2	fair dice
3	guess randomly
4	biased dice
5	fair dice
6	fair dice

What is the probability of your guess being correct?
(Hint: It might be easier to compute the probability via the ’two-world’ formulation.)

The probability of your guess being correct can be described in terms of the difference between the two probability vectors. More formally, suppose there are two probability distributions $\mathcal{D}_0$ and $\mathcal{D}_1$ over some set Ω. You are given exactly one sample, either from $\mathcal{D}_0$ or $\mathcal{D}_1$, and you must guess whether it came from $\mathcal{D}_0$ or $\mathcal{D}_1$ (the source distribution is either $\mathcal{D}_0$ or $\mathcal{D}_1$ with probability 1/2). Let p_i (resp. q_i) be the probability that i is sampled if we use distribution $\mathcal{D}_0$ (resp. $\mathcal{D}_1$). Consider the following strategy: for every $i \in \Omega$, if i is the sample, and $p_i > q_i$, guess that distribution $\mathcal{D}_0$ was chosen, else guess that $\mathcal{D}_1$ was chosen.

Check that this strategy has winning probability equal to $1/2 + 1/4(\sum_{i\in\Omega} |p_i - q_i|)$.

The quantity $1/2(\sum_{i\in\Omega} |p_i - q_i|)$ is referred to as the statistical distance between $\mathcal{D}_0$ and $\mathcal{D}_1$, and is represented as $\mathsf{SD}(\mathcal{D}_0, \mathcal{D}_1)$.

Definition 13.01. Given two distributions $\mathcal{D}_0$ and $\mathcal{D}_1$ over the same sample space Ω, the statistical distance of $\mathcal{D}_0$ and $\mathcal{D}_1$ is defined as

$$\mathsf{SD}(\mathcal{D}_0, \mathcal{D}_1) = \frac{1}{2}\left(\sum_{i\in\Omega}\left|\Pr_{x\leftarrow\mathcal{D}_0}[x=i] - \Pr_{x\leftarrow\mathcal{D}_1}[x=i]\right|\right)$$

In the case of fair dice vs biased dice, the statistical distance between the two distributions is $1/2((1/2 - 1/6) + 1/6 + 0 + (1/3 - 1/6) + 1/6 + 1/6)$.

Properties of statistical distance: Here are a few properties of statistical distance that will be useful for our analysis.

Claim 13.01. For any distributions $\mathcal{D}_0, \mathcal{D}_1$ over Ω, let $p_i = \Pr_{x\leftarrow\mathcal{D}_0}[x=i]$ and $q_i = \Pr_{x\leftarrow\mathcal{D}_1}[x=i]$. Define $\Omega_1 = \{i \in \Omega : p_i > q_i\}$ and $\Omega_2 = \Omega \setminus \Omega_1$. Then

$$\mathsf{SD}(\mathcal{D}_0, \mathcal{D}_1) = \sum_{i\in\Omega_1}(p_i - q_i) = \sum_{i\in\Omega_2}(q_i - p_i)$$

Proof.

$$\begin{aligned}\mathsf{SD}(\mathcal{D}_0, \mathcal{D}_1) &= \frac{1}{2}\left(\sum_{i\in\Omega}|p_i - q_i|\right)\\ &= \frac{1}{2}\left(\sum_{i\in\Omega_1}|p_i - q_i|\right) + \frac{1}{2}\left(\sum_{i\in\Omega_2}|p_i - q_i|\right)\\ &= \frac{1}{2}\left(\sum_{i\in\Omega_1}(p_i - q_i)\right) + \frac{1}{2}\left(\sum_{i\in\Omega_2}(q_i - p_i)\right)\end{aligned}$$

In the last line, we are using the definitions of Ω_1 and Ω_2 to replace $|p_i - q_i|$ with either $(p_i - q_i)$ with either $(p_i - q_i)$ or $(q_i - p_i)$ depending on whether $i \in \Omega_1$ or $i \in \Omega_2$.

Next, we will show that $\sum_{i \in \Omega_1}(p_i - q_i)$ is equal to $\sum_{i \in \Omega_2}(q_i - p_i)$.

$$\begin{aligned}
\sum_{i \in \Omega_1}(p_i - q_i) &= \left(\sum_{i \in \Omega_1} p_i\right) - \left(\sum_{i \in \Omega_1} q_i\right) \\
&= \left[1 - \left(\sum_{i \in \Omega_1} q_i\right)\right] - \left[1 - \left(\sum_{i \in \Omega_1} p_i\right)\right] \\
&= \left(\sum_{i \in \Omega_2} q_i\right) - \left(\sum_{i \in \Omega_2} p_i\right) \\
&= \sum_{i \in \Omega_2}(q_i - p_i)
\end{aligned}$$

In the second last step, we use the fact that $\Omega_2 = \Omega \setminus \Omega_1$, and as a result, $\sum_{i \in \Omega_2} p_i = 1 - \sum_{i \in \Omega_1} p_i$ (and similarly for q_i). This concludes the proof of our claim. □

We saw a 'guessing strategy' where we decided our guess based on whether $p_i > q_i$ or $q_i \geq p_i$. This strategy has winning probability $1/2 + \mathsf{SD}(\mathcal{D}_0, \mathcal{D}_1)/2$ (it might be easier to use the two-world formulation to prove this).

Next, we will show that this is the 'best possible' strategy.

Claim 13.02. For any adversary $\mathcal{A}$, the winning probability in distinguishing $\mathcal{D}_0$ and $\mathcal{D}_1$ is at most $1/2 + \mathsf{SD}(\mathcal{D}_0, \mathcal{D}_1)/2$.

Proof. Since we are considering unbounded adversaries here, it suffices to consider deterministic adversaries only. The adversary, for any $i \in \Omega$, has a fixed 'guess' $g(i)$ corresponding to i. Consider the set $\Omega_0 = \{i : g(i) = 0\}$.

If we consider the 'two-world formulation', note that

$$\begin{aligned}
\beta_0 = \Pr\Big[\mathcal{A} \text{ outputs } 0 \text{ in world-}0\Big] &= \textstyle\sum_{i \in \Omega_0} p_i \\
\beta_1 = \Pr\Big[\mathcal{A} \text{ outputs } 0 \text{ in world-}1\Big] &= \textstyle\sum_{i \in \Omega_0} q_i \\
\beta_0 - \beta_1 &= \textstyle\sum_{i \in \Omega_0}(p_i - q_i)
\end{aligned}$$

Note that $|\beta_0 - \beta_1| \leq \mathsf{SD}(\mathcal{D}_0, \mathcal{D}_1)$, and therefore, in the guessing game, the probability of win is equal to $1/2 + (\beta_0 - \beta_1)/2$, which is at most $1/2 + \mathsf{SD}(\mathcal{D}_0, \mathcal{D}_1)/2$. □

Note: This guessing strategy requires $O(|\Omega|)$ time. The strategy requires the probabilities p_i, q_i of every $i \in \Omega$. However, the importance of statistical distance comes from the following observation: if we show an upper bound on the statistical distance between two distributions, then it implies a bound on the success probability of ANY adversary (not necessarily p.p.t).

2.2.2 Sampling with replacement vs sampling without replacement

Suppose you sample t numbers from the set $\{1, 2, \ldots, N\} \equiv [N]$, where each sample is drawn uniformly at random, and with replacement. Since all t samples are drawn uniformly at random, and with replacement, every vector $(r_1, \ldots, r_t)$ has equal probability ($= 1/N^t$) of being sampled.

On the other hand, suppose the t samples are drawn from $[N]$ without replacement. Then the probability of a vector $(r_1, \ldots, r_t)$ is $1/(N \cdot (N-1) \cdot (N-t+1))$ if all the r_i values are distinct; else the probability is 0.

Now consider the following distributions over $\Omega = [N]^t$ (the set of all t-tuples):

$$\mathcal{D}_0 = \{t \text{ numbers are sampled u.a.r. from } [N] \text{ with replacement}\}$$
$$\mathcal{D}_1 = \{t \text{ numbers are sampled u.a.r. from } [N] \textbf{ without } \text{replacement}\}$$

Can we show a bound on the statistical distance between $\mathcal{D}_0$ and $\mathcal{D}_1$? The statistical distance between $\mathcal{D}_0$ and $\mathcal{D}_1$ is closely related to the 'birthday bound'.

Claim 13.03.

$$\mathsf{SD}(\mathcal{D}_0, \mathcal{D}_1) = \Pr\left[\text{There exist repetitions when sampling } t \text{ numbers, u.a.r. with replacement}\right] = \Theta(t^2/N).$$

Proof. Let p_i (resp. q_i) denote the probability of i when sampled from $\mathcal{D}_0$ (resp. $\mathcal{D}_1$). First, as shown in Claim 13.01, $\mathsf{SD}(\mathcal{D}_0, \mathcal{D}_1) = \left(\sum_{i:p_i>q_i}(p_i - q_i)\right) = \left(\sum_{i:p_i<q_i}(q_i - p_i)\right)$. For all tuples in $[N]^t$ where at least two elements repeat, the distribution $\mathcal{D}_1$ assigns 0 probability to these tuples, while $\mathcal{D}_0$ will assign them $1/N^t$ probability.

Hence, if we just focus on the tuples where at least one element repeats, it follows that

$$\left(\sum_{i:p_i>q_i}(p_i - q_i)\right) = \Pr\left[\text{There exist repetitions when sampling } t \text{ numbers, u.a.r. with replacement}\right].$$

□

If $N = 2^n$ and $t = \mathsf{poly}(n)$, then it follows that even an exponential time adversary cannot distinguish between a sample from $\mathcal{D}_0$ and a sample from $\mathcal{D}_1$ (except with negligible advantage). We will use this observation for proving that a random permutation and a random function are indistinguishable.

2.2.3 Random permutation vs random function (with polynomial samples)

Suppose an adversary makes t queries, and uses the t queries to distinguish between a random permutation and a random function. We can assume, without loss of generality, that the adversary does not repeat any queries (since it will receive the same response on repeating a query). By definition of a random function, on any input, it receives a uniformly random element from the domain. Therefore, this is identical to t samples from $\{0,1\}^n$ drawn uniformly at random, with replacement.

Similarly, by definition of a random permutation, on the i^{th} query, it receives a uniformly random element subject to the restriction that it is not equal to any of the previous $(i-1)$ responses. Therefore, this is identical to sampling t elements uniformly at random from $\{0,1\}^n$ *without replacement*.

Therefore, if an adversary can distinguish between a uniformly random function and a uniformly random permutation with t queries, then there exists an adversary that can distinguish between t samples drawn u.a.r. with replacement, and t samples drawn u.a.r. without replacement. But we saw in Claim 13.03 that these two distributions have negligible statistical distance, and therefore no adversary can distinguish between them with non-negligible advantage.

3 Lecture summary, plan for next lecture, additional resources

Summary: Here are the main take-aways from this lecture:

- PRPs can be constructed from secure PRFs (the only modification needed was to use different keys for $P_1(\cdot, k)$.
- Any PRP is also a PRF. That is, a random permutation 'looks like' a random function if the domain is much larger than the number of queries to the permutation. The main take-away for this part is the analysis using statistical distance of distributions. This is a useful tool for showing that two distributions are indistinguishable (even for unbounded adversaries).

Next Lecture: We will define semantic security for encryption schemes against general 'read-only' adversaries, and we will see that our current constructions can be tweaked slightly to achieve security against such adversaries.

Relevant sections from textbook [Boneh-Shoup]: Section 4.4.3 discusses why a PRP is also a secure PRF. Section 4.5 contains the PRF → PRP transformation. Statistical distance is discussed in Section 3.11.

4 Questions

Question 13.01. Let $G : \{0,1\}^n \to \{0,1\}^{2n}$ be a deterministic function. Consider the following two distributions on $\Omega = \{0,1\}^{2n}$.

$$\mathcal{D}_0 = \{\text{Sample } s \leftarrow \{0,1\}^n \text{ uniformly at random, output } G(s)\}$$
$$\mathcal{D}_1 = \left\{\text{Output } u \leftarrow \{0,1\}^{2n} \text{ uniformly at random}\right\}$$

Show that $\mathsf{SD}(\mathcal{D}_0, \mathcal{D}_1)$ is close to 1 (and therefore, there exists an exponential time algorithm that can distinguish between the two distributions).

Hint: Note that in distribution $\mathcal{D}_0$, there are at most 2^n strings in the sample space with non-zero probability. As a result, there are at least $2^{2n} - 2^n$ strings with probability 0. Let this set be S. Then $\sum_{x \in S} \Pr_{u \leftarrow \mathcal{D}_1}[u = x] - \Pr_{u \leftarrow \mathcal{D}_0}[u = x] = |S|/2^{2n} \geq 1 - 1/2^n$. Therefore, the statistical distance is close to 1.