CS195-7: Introduction to Cryptography DUE: October 31, 2002

Problem Set 5-6

Instructor: Anna Lysyanskaya

Problem 1: Indistinguishability and the hybrid argument

There are five parts to this problem. There are also a few exercises in parentheses. You do not have to turn in the answers to the exercises, but it is a good idea to do them nevertheless.

Recall the notion of indistinguishability. We say that two polynomial-time samplable distributions, $\mathcal{A}$ and $\mathcal{B}$ are indistinguishable if for all probabilistic polynomial-time adversaries $\{A_k\}$ there exists a negligible function $\nu(k)$ such that

$$\Pr[x_0 \leftarrow \mathcal{A}(1^k); x_1 \leftarrow \mathcal{B}(1^k); b \leftarrow \{0,1\}; b' \leftarrow A_k(x_b) : b = b'] = 1/2 + \nu(k)$$

Similarly, we can define indistinguishability if $\mathcal{A}$ and $\mathcal{B}$ are *families* of distributions indexed by some public parameter generated by the key generation procedure G. In that case $(G, \mathcal{A}, \mathcal{B})$ is a family of indistinguishable distributions if for all probabilistic polynomial-time adversaries $\{A_k\}$ there exists a negligible function $\nu(k)$ such that

$$\Pr[PK \leftarrow G(1^k); x_0 \leftarrow \mathcal{A}(PK); x_1 \leftarrow \mathcal{B}(PK); b \leftarrow \{0,1\}; b' \leftarrow A_k(PK, x_b) : b = b'] = 1/2 + \nu(k)$$

Notation: Let $\mathcal{A} \approx \mathcal{B}$ denote that the distributions (or families of distributions) $\mathcal{A}$ and $\mathcal{B}$ are indistinguishable. (Exercise: why is it that $\mathcal{A} \approx \mathcal{B} \Leftrightarrow \mathcal{B} \approx \mathcal{A}$?)

(a) Let $(G, \mathcal{A}, \mathcal{B})$ be families of distributions. Prove that the following statements are equivalent:

1. $\mathcal{A} \approx \mathcal{B}$.
2. For all probabilistic poly-time algorithms $\{A_k\}$, there exists a negligible function $\nu(k)$ such that $|p_{\mathcal{A}}^{\{A_k\}}(k) - p_{\mathcal{B}}^{\{A_k\}}(k)| = \nu(k)$, where by $p_X(k)$, $X = \mathcal{A}, \mathcal{B}$, we mean:

$$p_X(k) = \Pr[PK \leftarrow G(1^k); x \leftarrow X(PK); b' \leftarrow A_k(PK, x) \ : \ b' = 0]$$

In other words, two (families of) distributions are indistinguishable iff no matter what algorithm A_k you run on x drawn from one of the distributions, it will, with all but negligible probability, behave the same as if x was drawn from the other distribution.

In lecture, we have seen the so-called *hybrid argument* for proving that two distributions $\mathcal{A}$ and $\mathcal{B}$ are computationally indistinguishable. The main tool that we used was the following lemma:

Lemma 1 For all polynomial-time samplable (families of) distributions $\mathcal{A}$ and $\mathcal{B}$, if there exists a polynomial-time samplable (family of) distribution(s) $\mathcal{C}$ such that $\mathcal{A} \approx \mathcal{C}$ and $\mathcal{B} \approx \mathcal{C}$, then $\mathcal{A} \approx \mathcal{B}$.

Proof: Let $\{D_k\}$ be an adversary. Using part (a), it is sufficient to show that there exists a negligible function $\nu(k)$ such that $|p_{\mathcal{A}}^{\{D_k\}}(k) - p_{\mathcal{B}}^{\{D_k\}}(k)| = \nu(k)$. Since $\mathcal{A} \approx \mathcal{C}$, and $\mathcal{C} \approx \mathcal{B}$, we know that there exist negligible functions $\nu_1(k)$ and $\nu_2(k)$ such that $|p_{\mathcal{A}}^{\{D_k\}} - p_{\mathcal{C}}^{\{D_k\}}| = \nu_1(k)$ and $|p_{\mathcal{B}}^{\{D_k\}} - p_{\mathcal{C}}^{\{D_k\}}| = \nu_2(k)$. Then

$$|p_{\mathcal{A}}^{\{D_k\}} - p_{\mathcal{B}}^{\{D_k\}}| = |p_{\mathcal{A}}^{\{D_k\}} - p_{\mathcal{C}}^{\{D_k\}} + p_{\mathcal{C}}^{\{D_k\}} - p_{\mathcal{B}}^{\{D_k\}}| \leq |p_{\mathcal{A}}^{\{D_k\}} - p_{\mathcal{C}}^{\{D_k\}}| + |p_{\mathcal{C}}^{\{D_k\}} - p_{\mathcal{B}}^{\{D_k\}}| = \nu_1(k) + \nu_2(k)$$

and we are done since the sum of two negligible functions is negligible.

(b) The claim proved below is false. (Exercise: why?) However, below is a proof of this claim. Find the error in this proof.

> **Claim:** Let $\mathcal{A}(1^k)$ be the uniform distribution on integers in the interval $[1, 2^k]$. Let $\mathcal{B}(1^k)$ be the uniform distribution on integers in the interval $[2^k + 1, 2^{k+1}]$. Then $\mathcal{A} \approx \mathcal{B}$.
>
> **Proof:** For $i \geq 0$, let $C_i(1^k)$ be the uniform distribution on the integers in the interval $[i, 2^k + i]$. It is clear that $C_0 = \mathcal{A}$, and $C_{2^k} = \mathcal{B}$. Also, note that for all i, $C_i \approx C_{i+1}$. (Exercise: why?)
>
> Let us prove that for all j, $C_i \approx C_{i+j}$, by induction on j. The base case, $j = 1$, is given. The inductive step: Suppose $C_i \approx C_{i+j-1}$. We also know that $C_{i+j-1} \approx C_{i+j}$. Therefore, the conditions of Lemma 1 are satisfied, and $C_i \approx C_{i+j}$.
>
> Therefore, in particular, $\mathcal{A} = C_0 \approx C_{2^k} = \mathcal{B}$.

Part (b) shows that care must be taken when applying Lemma 1 to proving indistinguishability of distributions. Let us develop some notation that will allow us to be more careful. Instead of simply talking about two distributions as being indistinguishable, let us specify exactly what the maximum advantage of a distinguisher would be in telling the two distributions apart, as a function of the running time of this distinguisher.

Notation: Let $\mathcal{A}$ and $\mathcal{B}$ be efficiently samplable families of distributions with key generation algorithm G. Let the notation "$\mathcal{A} \approx \mathcal{B}$ with advantage at most $\epsilon(\cdot)$" mean that for all probabilistic polynomial-time adversaries $\{A_k\}$ $|p_{\mathcal{A}}^{\{A_k\}}(k) - p_{\mathcal{B}}^{\{A_k\}}(k)| \leq \epsilon(k)$.

The following lemma is a more precise version of Lemma 1:

Lemma 2 For all polynomial-time samplable (families of) distributions $\mathcal{A}$ and $\mathcal{B}$, if there exists a polynomial-time samplable (family of) distribution(s) $\mathcal{C}$ such that $\mathcal{A} \approx \mathcal{C}$ with advantage at most $\epsilon_1(k)$ and $\mathcal{B} \approx \mathcal{C}$ with advantage at most $\epsilon_2(k)$, then $\mathcal{A} \approx \mathcal{B}$ with advantage at most $\epsilon_1(k) + \epsilon_2(k)$.

Proof idea: Same as the proof of Lemma 1.

Further, we can now consider the meat of the hybrid argument:

Lemma 3 For any integer $i \geq 0$, let $\mathcal{A}_i$ be a family of polynomial-time samplable distributions. Suppose that for all $1 \leq i \leq n$, $\mathcal{A}_{i-1} \approx \mathcal{A}_i$ with advantage at most ϵ_i. Then $\mathcal{A}_0 \approx \mathcal{A}_n$ with advantage at most $\sum_{i=1}^{n} \epsilon_i$.

(c) Prove Lemma 3.

Lemma 4 is the converse of Lemma 3. Namely, it says that if we can distinguish between $\mathcal{A}_0$ and $\mathcal{A}_n$, then we can distinguish between $\mathcal{A}_{i-1}$ and $\mathcal{A}_i$ for some i.

Lemma 4 Let $\mathcal{A}$ and $\mathcal{B}$ be polynomial-time samplable families of distributions. with setup algorithm G. Let a distinguisher $\{D_k\}$ such that $p_{\mathcal{A}}^{\{D_k\}} - p_{\mathcal{B}}^{\{D_k\}} \geq \epsilon(k)$ be given. Let n be any integer. Let $\mathcal{A}_0 = \mathcal{A}$, $\mathcal{A}_n = \mathcal{B}$, and $\mathcal{A}_i$, $1 \leq i < n$, be any families of distributions with setup algorithm G. Then for some $1 \leq i \leq n$, $p_{\mathcal{A}_{i-1}}^{\{D_k\}} - p_{\mathcal{A}_i}^{\{D_k\}} \geq \epsilon(k)/n$.

Proof: Simple averaging argument that we saw in class.

Finally, Lemma 5 states that if we can distinguish between $\mathcal{A}$ and $\mathcal{B}$ using algorithm $\{D_k\}$ with advantage ϵ, then for a random i, $\{D_k\}$ is a good distinguisher with good probability.

Lemma 5 Let $\mathcal{A}_i$ be as in Lemma 4. Let P be a random variable, as follows: $P = p_{\mathcal{A}_i}^{\{D_k\}} - p_{\mathcal{A}_{i-1}}^{\{D_k\}}$ for $1 \leq i \leq n$ chosen uniformly at random. Then $E[P] \geq \epsilon(k)/n$.

(d) Prove Lemma 5.

(e) Apply the Markov inequality to derive a lower bound on the probability that, for a randomly chosen i, the distinguisher will have advantage at least $\epsilon/2n$. Make your bound as tight as you can.

Problem 2: RM-security

Let $\mathcal{M}$ be a set of messages. This set may consist of infinitely many messages, for example, of all binary strings. Let $\mathcal{M}_\ell$ denote the set of all ℓ-bit strings in $\mathcal{M}$. I.e.,

$$\mathcal{M}_\ell = \{m \mid m \in \mathcal{M}, |m| = \ell\}$$

Recall GM security, where the adversary chooses two messages of the same length whose encryptions he intends to tell apart. More precisely:

Definition (GM-Security) A cryptosystem (G, E, D) on a set of messages $\mathcal{M}$ is GM-secure if

$\forall\{A_k\}\ \exists\nu(k)\in \mathbf{neg}$ s.t.

$$\begin{aligned}\Pr[\quad & (PK, SK)\leftarrow G(1^k); (\alpha, m_0, m_1)\leftarrow A_k(PK);\\ & c_0\leftarrow E(PK, m_0); c_1\leftarrow E(PK, m_1);\\ & b\leftarrow\{0,1\}; b'\leftarrow A_k(\alpha, c_b) : \qquad b' = b]\le 1/2+\nu(k)\end{aligned}$$

Consider the following definition of security, which we call random message security (RM security). Here, the adversary chooses just one message. The adversary's goal is to tell an encryption of this message apart from an encryption of a random message of the same length. A cryptosystem is RM-secure if any polynomial-time adversary fails. More precisely:

Definition (RM-Security) A cryptosystem (G, E, D) on a set of messages $\mathcal{M}$ is random-message (RM) secure if

$\forall\{A_k\}\ \exists\nu(k)\in \mathbf{neg}$ s.t.

$$\begin{aligned}\Pr[\quad & (PK, SK)\leftarrow G(1^k); (\alpha, m)\leftarrow A_k(PK);\\ & r\leftarrow\mathcal{M}_{|m|}; c_0\leftarrow E(PK, m); c_1\leftarrow E(PK, r);\\ & b\leftarrow\{0,1\}; b'\leftarrow A_k(\alpha, c_b) : \qquad b' = b]\le 1/2+\nu(k)\end{aligned}$$

Prove or disprove that for all sets of messages $\mathcal{M}$, RM security is equivalent to GM security. (The hybrid argument from Problem 1 may be relevant.)

Problem 3: Pseudorandomness vs. unpredictability

In lecture, we assumed that if a bit is hardcore, i.e., hard to predict, then it is pseudorandom, i.e., hard to distinguish from random. Here, we will prove it and further consider preudorandomness vs. unpredictability.

(a) Prove that a bit is hard to predict if and only if it is pseudorandom. More formally: Let $\{D_{PK}\}$ be an efficiently samplable family of distributions with setup procedure G. Let $f(PK, x)$ be any Boolean function. (For example, $\{D_{PK}\}$ can be the distribution of GM ciphertexts for one bit, and $f(PK, x)$ is the decryption of the ciphertext x under public key PK. Or, $\{D_{PK}\}$ may be a family of one-way permutation, and $f(PK, x)$ may be its hardcore bit.) Prove that the following two conditions are equivalent:

Condition 1 (unpredictability): For all probabilistic polynomial-time adversaries $\{A_k\}$, there exists a negligible function $\nu(k)$, such that

$$\begin{aligned}\Pr[&PK\leftarrow G(1^k); x\leftarrow D_{PK};\\ & b'\leftarrow A_k(PK, x)\ :\ f(PK, x) = b']\le 1/2+\nu(k)\end{aligned}$$

More generally, for any (not necessarily boolean) function f, let $p_{max}(D)$ be the probability of the most likely element of D_{PK}. Then f is unpredictable if for all probabilistic polynomial-time adversaries $\{A_k\}$, there exists a negligible function $\nu(k)$, such that

$$\begin{aligned}\Pr[PK \leftarrow G(1^k); x \leftarrow D_{PK};\\ y' \leftarrow A_k(PK, x) \quad : \quad f(PK, x) = y'] \leq p_{max}(D) + \nu(k)\end{aligned}$$

Condition 2 (pseudorandomness): For all probabilistic polynomial-time adversaries $\{A_k\}$, there exists a negligible function $\nu(k)$, such that

$$\begin{aligned}\Pr[\quad & a_0 \leftarrow \{0,1\};\\ & PK \leftarrow G(1^k); x \leftarrow D_{PK}; a_1 = f(PK, x);\\ & b \leftarrow \{0,1\}; b' \leftarrow A_k(PK, x, a_b) \quad : \; b = b'] \leq 1/2 + \nu(k)\end{aligned}$$

For non-boolean f, the definition of pseudorandomness is essentially the same. Let $\{R_k\}$ denote some polynomial-time samplable family of distributions. For example, $\{R_k\}$ can be the uniform distribution on k-bit strings. Function f is pseudorandom if there exists a polynomial-time samplable family of distributions $\{R_k\}$ such that for all probabilistic polynomial-time adversaries $\{A_k\}$, there exists a negligible function $\nu(k)$, such that

$$\begin{aligned}\Pr[\quad & a_0 \leftarrow R_k;\\ & PK \leftarrow G(1^k); x \leftarrow D_{PK}; a_1 = f(PK, x);\\ & b \leftarrow \{0,1\}; b' \leftarrow A_k(PK, x, a_b) \quad : \; b = b'] \leq 1/2 + \nu(k)\end{aligned}$$

(b) Suppose that $f(PK, x)$ is not a boolean function, but rather a function with the range $\{0,1\}^k$. Will $f(PK, x)$ be unpredictable if it is pseudorandom? Will it be pseudorandom if it is unpredictable? What if the range of the function is $\{0,1\}^{\log k}$? (You don't have to write the proofs in detail for this part of the problem, it is sufficient to just sketch them.)