Advanced Cryptography — Midterm Exam
Solution

Serge Vaudenay

28.4.2015

- duration: 3h
- any document allowed
- a pocket calculator is allowed
- communication devices are not allowed
- the exam invigilators will **not** answer any technical question during the exam
- readability and style of writing will be part of the grade

The exam grade follows a linear scale in which each question has the same weight.

1 Blind Computing with a DH Oracle

The goal of this exercise is to look at what happens when the discrete logarithm problem is hard but the Diffie-Hellman problem is easy. Let g be an element of a group of prime order q. Computing the discrete logarithm in the group generated by g is assumed to be hard. We assume that we have an oracle function $\mathsf{DH}(X,Y)$ such that when queried with $X = g^x$ and $Y = g^y$ for integers x and y, it returns $\mathsf{DH}(g^x, g^y) = g^{xy}$ with one unit of time complexity.

In this exercise we construct series of algorithms using the oracle DH. These algorithms also know g and q. They can perform a group multiplication and a group inversion within one unit of time complexity. In each question of this exercise except the last one, we define a function by a property based on values which are not always computable. For instance, a blind multiplication defined by $f_0(g^x, g^y) = g^{xy}$ is implemented by the algorithm

$f_0(X,Y)$:
1: $Z \leftarrow \mathsf{DH}(X,Y)$
2: return Z

without being able to compute the logarithms x or y of X and Y.

For each of these questions, define an efficient algorithm to implement the computation of the function and give its complexity. When studying the complexity, separate the number of queries, the number of group multiplications/inversions, and the usual asymptotic complexity of other operations.

Q.1 (blind addition) $f_1(g^x, g^y) = g^{x+y}$.

$f_1(X, Y)$:

$Z \leftarrow X \times Y$
return Z

The complexity is of 1 *group multiplication.*

Q.2 (blind scalar multiplication) $f_2(a, g^x) = g^{ax}$ when a is an integer (positive or negative).

We note that $g^{ax} = X^a$. *So, we just apply the square-and-multiply algorithm to compute* X^a.

$f_2(a, X)$:

if $a = 0$ *return 1*
write $|a|$ *in binary* $|a| = \sum_{i=0}^{\ell-1} a_i 2^i$ *with* $a_i \in \{0, 1\}$ *and* $a_{\ell-1} = 1$.
$Z \leftarrow X$
for $i = \ell - 2$ *down to* 0 ***do***
  $Z \leftarrow Z \times Z$
  if $a_i = 1$ *do* $Z \leftarrow Z \times X$
end for
if $a < 0$ *do* $Z \leftarrow 1/Z$
return Z

We have $\ell = \lfloor \log_2 |a| \rfloor + 1$. *The complexity is of up to* $2\ell - 2$ *group multiplications, up to* 1 *inversion, and* $\mathcal{O}(\ell)$ *other operations.*

Q.3 (blind power) $f_3(e, g^x) = g^{x^e}$ when e is a positive integer.

We apply again the square-and-multiply algorithm, but with the DH *oracle for the multiplications.*

$f_3(e, X)$:

if $a = 0$ *return 1*
write e *in binary* $e = \sum_{i=0}^{\ell-1} e_i 2^i$ *with* $e_i \in \{0, 1\}$ *and* $e_{\ell-1} = 1$.
$Z \leftarrow X$
for $i = \ell - 2$ *down to* 0 ***do***
  $Z \leftarrow \mathsf{DH}(Z, Z)$
  if $e_i = 1$ *do* $Z \leftarrow \mathsf{DH}(Z, X)$
end for
return Z

We have $\ell = \lfloor \log_2 e \rfloor + 1$. *The complexity is of up to* $2\ell - 2$ DH *oracle calls and* $\mathcal{O}(\ell)$ *other operations.*

Q.4 (blind sparse polynomial) $f_4(a_1, e_1, \ldots, a_n, e_n, g^x) = g^{\sum_{i=1}^{n} a_i x^{e_i}}$ when the e_i's are positive integers and the a_i's are nonzero integers.

Using the previous questions, we just compute all $g^{x^{e_i}}$, raise them to the a_i powers, and multiply them all.

$f_4(a_1, e_1, \ldots, a_n, e_n, X)$:

```
Z ← 1
for i = 1 to n do
    compute T = f_3(e_i, X)
    compute T = f_2(a_i, T)
    compute Z = Z × T
end for
return Z
```

The complexity is bounded by $n + 2\sum_{i=1}^{n} \log_2 |a_i|$ multiplications, n inversions, $2\sum_{i=1}^{n} \log_2 e_i$ DH oracle calls, and $\mathcal{O}(n + \sum_{i=1}^{n} \log_2(|a_i| e_i))$ other operations.

Q.5 (blind inversion) $f_5(g^x) = g^{\frac{1}{x} \bmod q}$ when $g^x \neq 1$.

We observe that $\frac{1}{x} \equiv x^{q-2} \pmod q$. So, we just compute g^{x^e} for $e = q - 2$.

$f_5(X)$:

```
Z ← f_3(q − 2, X)
return Z
```

The complexity is bounded by $2\log_2 q$ oracle calls and $\mathcal{O}(\log_2 q)$ other operations.

Q.6 (blind e-th root when e is invertible) $f_6(e, g^{y^e}) = g^y$ when e is a positive integer which is coprime with $q - 1$.

We observe that if $y^e \equiv x \pmod q$, then

$$y \equiv x^{\frac{1}{e} \bmod (q-1)} \pmod q$$

So, we can just compute $\frac{1}{e} \bmod (q-1)$ using the extended Euclid Algorithm in complexity quadratic in $\log q$ then use the previous algorithms.

$f_6(e, X)$:

```
compute t = 1/e mod (q − 1) using the extended Euclid Algorithm
Z ← f_3(t, X)
return Z
```

The complexity is bounded by $2\log_2 q$ oracle calls plus $\mathcal{O}((\log q)^2)$ other operations.

Q.7 (blind square root) $f_7(g^{y^2}) \in \{g^y, g^{-y}\}$. (For simplicity, we assume $q \bmod 4 = 3$.)

We observe that $x^{\frac{q+1}{4}} \bmod q$ is a square root of x whenever such square root exists. So, we use the previous algorithms to compute g^{x^e} for $e = \frac{q+1}{4}$. Note that the two square roots are then x^e and $-x^e$, so $(y^2)^e \in \{y, -y\}$.

$f_7(X)$:

1: *compute $e = \frac{q+1}{4}$*
2: *$Z \leftarrow f_3(e, X)$*
3: *return Z*

The complexity is bounded by $2\log_2 q$ oracle calls and $\mathcal{O}(\log q)$ other operations.

Q.8 With the same notations and assumptions, construct a commitment scheme which is deterministic computationally hiding and perfectly binding on $\mathbf{Z}_q$, with the property that given a rational function $f(x_1, \ldots, x_n)$ and some commitments on x_1, ..., and x_n, it is easy to deduce a commitment to $f(x_1, \ldots, x_n)$ without knowing $x_1, \ldots, x_n$.

The mapping $\mathsf{Com} : x \mapsto g^x$ is such a commitment!
Indeed, it is deterministic. If the discrete logarithm is hard, it is deterministic computationally hiding. Since it is injective on $\mathbf{Z}_q$, it is perfectly binding (there is no collision). From $\mathsf{Com}(x)$ and $\mathsf{Com}(y)$, we can compute $\mathsf{Com}(x+y)$ easily. From $\mathsf{Com}(x)$ and $\mathsf{Com}(y)$, we can compute $\mathsf{Com}(xy)$ if the computational Diffie-Hellman problem is easy. Finally, from $\mathsf{Com}(x)$ we can compute $\mathsf{Com}(1/x)$ using the previous questions. So, we can evaluate any rational function.

2 On the Necessary Number of Samples to Distinguish a Biased Coin

We are given a source of independent random bits following one given distribution. We want to distinguish a given distribution from the uniform one. The goal of this exercise is to prove that if ε is the statistical distance between the two distributions, then ε^{-2} is a necessary and sufficient order of magnitude of number of samples which is needed to reach an advantage of $\frac{1}{2}$ or higher.

Given two random variables X and Y with the same support $\mathcal{Z}$, we define

$$L(X,Y) = \sum_{z\in\mathcal{Z}} |\Pr[X=z] - \Pr[Y=z]|$$

$$D(X\|Y) = \sum_{z\in\mathcal{Z}} \Pr[X=z] \log_2 \frac{\Pr[X=z]}{\Pr[Y=z]}$$

In what follows, some questions are more related to calculus than cryptography. Their results are necessary for the exercise but left as "bonus questions".

Q.1 Let $p = \frac{1}{2}(1+\varepsilon)$ for some $\varepsilon \in [-1,+1]$ and the X_i's be independent boolean random variables with expected value p. Let the Y_i's be independent uniformly distributed boolean random variables. Given a number of samples n, we want to distinguish $X = (X_1,\ldots,X_n)$ from $Y = (Y_1,\ldots,Y_n)$. We assume that the value of ε is known.

Q.1a Given a threshold λ, we propose the distinguisher

get the samples $z_1,\ldots,z_n$
compute $s = z_1 + \cdots + z_n$
if $\frac{s}{n} \geq \lambda$ **then**
  return 1
else
  return 0
end if

Show that for some value λ^* (give the formula) of λ, this distinguisher is optimal among those using n samples.

The optimal distinguisher answers 1 if and only if the likelihood ratio is greater than 1. That it, $p^s(1-p)^{n-s} \geq 2^{-n}$ where $s = z_1 + \cdots + z_n$. This is equivalent to $\left(\frac{p}{1-p}\right)^s \geq (2(1-p))^{-n}$, i.e. to

$$\frac{s}{n} \geq -\frac{\ln(2(1-p))}{\ln\left(\frac{p}{1-p}\right)}$$

So we have

$$\lambda^* = -\frac{\ln(2(1-p))}{\ln\left(\frac{p}{1-p}\right)} = -\frac{\ln(1-\varepsilon)}{\ln\frac{1+\varepsilon}{1-\varepsilon}}$$

Q.1b (Bonus question) Show that λ^* is close to $\frac{1}{2} + \frac{\varepsilon}{4}$ when $|\varepsilon|$ is small.

HINT: for θ close to 0, $\ln(1+\theta) = \theta - \frac{\theta^2}{2} + o(\theta^2)$.

For $p = \frac{1}{2}(1+\varepsilon)$ with ε small, we have

$$\lambda^* = -\frac{\ln(1-\varepsilon)}{\ln\frac{1+\varepsilon}{1-\varepsilon}} \approx \frac{\varepsilon + \frac{\varepsilon^2}{2}}{\ln\left(1 + 2\varepsilon + 2\varepsilon^2\right)} \approx \frac{\varepsilon + \frac{\varepsilon^2}{2}}{2\varepsilon} \approx \frac{1}{2} + \frac{\varepsilon}{4}$$

So, the optimal threshold is close to $\frac{1}{2} + \frac{\varepsilon}{4}$.

Q.1c For $n = 12\varepsilon^{-2}$, show that the advantage of the above distinguisher for $\lambda = \frac{1}{2} + \frac{\varepsilon}{4}$ is greater than $\frac{1}{2}$.

HINT: if $Z_1, \ldots, Z_n$ are i.i.d. boolean random variables of expected value μ, the Chernoff-Hoeffding bound says that $\Pr[Z_1 + \cdots + Z_n < n(\mu - t)] \leq e^{-2nt^2}$.

We can build the following distinguisher:

get n samples $z_1, \ldots, z_n$
if *$\frac{z_1+\cdots+z_n}{n}$ is closer to p than to $\frac{1}{2}$* ***then***
    return 1
else
    return 0
end if

We assume w.l.o.g. that $p > \frac{1}{2}$. When we sample z_i using X, the probability to give a wrong answer is the Type 1 error $\alpha = \Pr[\frac{z_1+\cdots+z_n}{n} < p - \frac{\varepsilon}{4}]$. Using the Chernoff-Hoeffding lemma, we have $\alpha \leq e^{-\frac{n\varepsilon^2}{8}}$. When we sample z_i using Y, the probability to give a wrong answer is the Type 2 error $\beta = \Pr[\frac{z_1+\cdots+z_n}{n} > \frac{1}{2} + \frac{\varepsilon}{4}]$. Using the Chernoff-Hoeffding lemma, we have $\beta \leq e^{-\frac{n\varepsilon^2}{8}}$. So, the advantage is $\mathsf{Adv} = 1 - \alpha - \beta \geq 1 - 2e^{-\frac{n\varepsilon^2}{8}}$. For $n = 12\varepsilon^{-2}$, we obtain $\mathsf{Adv} = 1 - \alpha - \beta \geq 1 - 2e^{-\frac{3}{2}}$ so $\mathsf{Adv} \geq \frac{1}{2}$.

The goal of the next questions is to show that for $n \ll \varepsilon^{-2}$, the best advantage is negligible.

Q.2 If X_1 and X_2 are independent and Y_1 and Y_2 are independent, for $X = (X_1, X_2)$ and $Y = (Y_1, Y_2)$, show that $D(X\|Y) = D(X_1\|Y_1) + D(X_2\|Y_2)$.

Thanks to the independence hypothesis, we have

$$\begin{aligned}
D(X\|Y) &= \sum_{z_1,z_2} \Pr[X_1=z_1, X_2=z_2] \log_2 \frac{\Pr[X_1=z_1, X_2=z_2]}{\Pr[Y_1=z_1, Y_2=z_2]} \\
&= \sum_{z_1,z_2} \Pr[X_1=z_1, X_2=z_2] \log_2 \frac{\Pr[X_1=z_1]\Pr[X_2=z_2]}{\Pr[Y_1=z_1]\Pr[Y_2=z_2]} \\
&= \sum_{z_1,z_2} \Pr[X_1=z_1, X_2=z_2] \log_2 \frac{\Pr[X_1=z_1]}{\Pr[Y_1=z_1]} \\
&\quad + \sum_{z_1,z_2} \Pr[X_1=z_1, X_2=z_2] \log_2 \frac{\Pr[X_2=z_2]}{\Pr[Y_2=z_2]} \\
&= \sum_{z_1} \Pr[X_1=z_1] \log_2 \frac{\Pr[X_1=z_1]}{\Pr[Y_1=z_1]} + \sum_{z_2} \Pr[X_2=z_2] \log_2 \frac{\Pr[X_2=z_2]}{\Pr[Y_2=z_2]} \\
&= D(X_1\|Y_1) + D(X_2\|Y_2)
\end{aligned}$$

Q.3 Given two random boolean variables X' and Y', show that $\frac{L(X',Y')^2}{2\ln 2} \leq D(X'\|Y')$. HINT: express $g(t) = (2\ln 2) \cdot D(X'\|Y') - L(X',Y')^2$ in terms of $t = \Pr[X'=1]$ then derivate $g(t)$ to study the variations of this function.

We have

$$g(t) = (2\ln 2) \cdot D(X'\|Y') - L(X',Y')^2 = 2t \ln \frac{t}{t_0} + 2(1-t) \ln \frac{1-t}{1-t_0} - 4(t-t_0)^2$$

with $t_0 = E(Y')$. *We have*

$$g'(t) = 2\ln \frac{t}{t_0} - 2\ln \frac{1-t}{1-t_0} - 8(t-t_0)$$

and

$$g''(t) = \frac{2}{t} + \frac{2}{1-t} - 8 = \frac{2}{t(1-t)} - 8 \geq 0$$

since $t(1-t) \leq \frac{1}{4}$. *So,* g' *is an increasing function such that* $g'(t_0) = 0$. *Hence,* g *decreases for* $t < t_0$ *then increases for* $t > t_0$. *Besides,* $g(t_0) = 0$. *So,* $g(t) \geq 0$.

Q.4 (Bonus question) Let $p = \frac{1}{2}(1+\varepsilon)$ for some $\varepsilon \in [-1,+1]$ and X_1 be a boolean random variable with expected value p. Let Y_1 be a uniformly distributed boolean random variable. Show that $D(X_1\|Y_1) \leq \frac{\varepsilon^2}{(2\ln 2)\cdot(1-\varepsilon^2)}$.
HINT: the Taylor-Lagrange Theorem states that there exists some t_2 between t_0 and t such that $g(t) = g(t_0) + g'(t_0)(t-t_0) + \frac{1}{2}g''(t_2)(t-t_0)^2$.

With previous notations, we have $t = p$ and $t_0 = \frac{1}{2}$. Furthermore, $L(X_1, Y_1) = |\varepsilon|$ so $g(t) = (2\ln 2) \cdot D(X_1 \| Y_1) - \varepsilon^2$.
We have $g(\frac{1}{2}) = g'(\frac{1}{2}) = 0$. Using the Taylor-Lagrange formula to the order 1, there must exist t_2 between p and $\frac{1}{2}$ such that $g(t) = \frac{1}{2}g''(t_2)(t - \frac{1}{2})^2$. So,

$$D(X_1 \| Y_1) = \frac{g(t) + \varepsilon^2}{2\ln 2} = \frac{\frac{1}{2}g''(t_2)(t - \frac{1}{2})^2 + \varepsilon^2}{2\ln 2} = \frac{\varepsilon^2}{(2\ln 2) \cdot 4t_2(1 - t_2)}$$

We now use $4t_2(1 - t_2) \geq 4t_1(1 - t_1) \geq 4p(1 - p) = 1 - \varepsilon^2$ to obtain

$$D(X_1 \| Y_1) \leq \frac{\varepsilon^2}{(2\ln 2) \cdot (1 - \varepsilon^2)}$$

Q.5 The aim of the next sub-questions is to show that for all function f, $D(f(X) \| f(Y)) \leq D(X \| Y)$.

Q.5a Show $D(g(X) \| g(Y)) = D(X \| Y)$ for all 1-to-1 mapping g.

If g is 1-to-1, we have

$$\begin{aligned} D(g(X) \| g(Y)) &= \sum_z \Pr[g(X) = z] \log_2 \frac{\Pr[g(X) = z]}{\Pr[g(Y) = z]} \\ &= \sum_{z'} \Pr[X = z'] \log_2 \frac{\Pr[X = z']}{\Pr[Y = z']} \\ &= D(X \| Y) \end{aligned}$$

with $z' = g^{-1}(z)$.

Q.5b We say that m is a merging function if every input x except one is a fixed point of m, i.e. $m(x) = x$. Show that an arbitrary f can be written as a composition $f = g \circ m_n \circ \cdots \circ m_1$ of merging functions m_i and a 1-to-1 function g, for some integer n.
HINT: make a proof by induction based on the number of collisions.

We show by induction on the number of collisions of f, we can always write $f = g \circ m_n \circ \cdots \circ m_1$ where each m_i is a merging function and g is 1-to-1.
This is clear when f has no collision since it already 1-to-1.
Next, we assume that f is not 1-to-1. We take a collision $f(a) = f(b)$ with $a \neq b$, define $m_1(x) = x$ for all x except $x = b$ and $m_1(b) = a$, then define f' such that $f = f' \circ m_1$ and $f'(b)$ set to some unreached value. We can see that f' has less collisions and apply the induction to write $f' = g \circ m_n \circ \cdots \circ m_2$. So, $f = g \circ m_n \circ \cdots \circ m_1$.

Q.5c (Bonus question) Show that for all positive $\alpha, \beta, \alpha', \beta'$ real numbers,

$$(\alpha+\beta)\ln\frac{\alpha+\beta}{\alpha'+\beta'} \leq \alpha\ln\frac{\alpha}{\alpha'} + \beta\ln\frac{\beta}{\beta'}$$

Deduce that $D(m_i(X)\|m_i(Y)) \leq D(X\|Y)$ for all merging functions m_i (as defined in Q.5b).

HINT: use the convexity of $x \mapsto x\ln x$ on the two points $\frac{\alpha}{\alpha'}$ and $\frac{\beta}{\beta'}$ and their weighted average $\frac{\alpha+\beta}{\alpha'+\beta'}$.

We notice that $\varphi : x \mapsto x\ln x$ is a convex function. Indeed, its second derivative is $\varphi''(x) = \frac{1}{x}$ which is positive for $x > 0$. Then,

$$\begin{aligned}(\alpha+\beta)\ln\frac{\alpha+\beta}{\alpha'+\beta'} &= (\alpha'+\beta')\varphi\left(\frac{\alpha+\beta}{\alpha'+\beta'}\right)\\ &\leq (\alpha'+\beta')\left(\frac{\alpha'}{\alpha'+\beta'}\varphi\left(\frac{\alpha}{\alpha'}\right) + \frac{\beta'}{\alpha'+\beta'}\varphi\left(\frac{\beta}{\beta'}\right)\right)\\ &= \alpha\ln\frac{\alpha}{\alpha'} + \beta\ln\frac{\beta}{\beta'}\end{aligned}$$

by using the convexity of φ.
Here is another solution: let

$$\Delta = (\alpha+\beta)\ln\frac{\alpha+\beta}{\alpha'+\beta'} - \alpha\ln\frac{\alpha}{\alpha'} - \beta\ln\frac{\beta}{\beta'}$$

We study Δ for α' and β' non-negative of constant sum. If one of them vanishes, then $\Delta = -\infty$. An optimum is reached when $\frac{\partial\Delta}{\partial\alpha'} - \frac{\partial\Delta}{\partial\beta'} = 0$, which is equivalent to $\frac{\alpha}{\alpha'} - \frac{\beta}{\beta'} = 0$. In that case, we have $\Delta = 0$. So, $\Delta \leq 0$.
Then, we take a merging function m_i for which $m_i(a) = m_i(b) = a$. We let $\alpha = \Pr[X = a]$, $\beta = \Pr[X = b]$, $\alpha' = \Pr[Y = a]$, $\beta' = \Pr[Y = b]$. We have

$$D(m_i(X)\|m_i(Y)) - D(X\|Y) = (\alpha+\beta)\log_2\frac{\alpha+\beta}{\alpha'+\beta'} - \alpha\log_2\frac{\alpha}{\alpha'} - \beta\log_2\frac{\beta}{\beta'} \leq 0$$

So, we have $D(m_i(X)\|m_i(Y)) \leq D(X\|Y)$.

Q.5d Show that $D(f(X)\|f(Y)) \leq D(X\|Y)$ for all functions f.

We write $f = g \circ m_n \circ \cdots \circ m_1$ where g is 1-to-1 and each m_i is a merging function. We have

$$\begin{aligned}
D(f(X)\|f(Y)) &= D((g \circ m_n \circ \cdots m_1)(X)\|(g \circ m_n \circ \cdots m_1)(Y)) \\
&\leq D((m_n \circ \cdots m_1)(X)\|(m_n \circ \cdots m_1)(Y)) \\
&\leq D((m_{n-1} \circ \cdots m_1)(X)\|(m_{n-1} \circ \cdots m_1)(Y)) \\
&\leq \cdots \\
&\leq D(m_1(X)\|m_1(Y)) \\
&\leq D(X\|Y)
\end{aligned}$$

Q.6 Show that the best advantage Adv to distinguish the boolean random variables X_i and Y_i, for $E(X_i) = p = \frac{1}{2}(1+\varepsilon)$ and $E(Y_i) = \frac{1}{2}$ satisfies $\mathsf{Adv} \leq \frac{1}{2} \times \sqrt{\frac{n\varepsilon^2}{1-\varepsilon^2}}$. Assuming that $|\varepsilon| \leq \frac{1}{2}$, deduce that for $n \ll \varepsilon^{-2}$, the best advantage is negligible.
HINT: define f the function mapping the vector of n sample bits to the outcome of the distinguisher. Given $X' = f(X_1, \ldots, X_n)$ and $Y' = f(Y_1, \ldots, Y_n)$ for some independent uniformly distributed bits $Y_1, \ldots, Y_n$, express the advantage in terms of $L(X', Y')$ and bound it in terms of $D(X'\|Y')$.

We know that Adv is the statistical distance. So, $\mathsf{Adv} = \frac{1}{2}L(X', Y')$. We follow the hint and have $\mathsf{Adv} \leq \frac{1}{2}\sqrt{(2\ln 2) \cdot D(X'\|Y')}$ due to Q.3. Then, we apply Q.5 and obtain $\mathsf{Adv} \leq \frac{1}{2}\sqrt{(2\ln 2) \cdot D(X\|Y)}$ for $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$. With Q.2, we have $\mathsf{Adv} \leq \frac{1}{2}\sqrt{n(2\ln 2) \cdot D(X_1\|Y_1)}$. Finally, we apply Q.4 and obtain the result.
For $|\varepsilon| \leq \frac{1}{2}$, this shows that $\mathsf{Adv} \leq \frac{1}{\sqrt{3}}\sqrt{n\varepsilon^2}$. Clearly, for $n \ll \varepsilon^{-2}$, we obtain that the best advantage Adv is negligible.