Statistical Distance Is HVSZK

Personal Notes

Ari Biswas
University Of Warwick
aribiswas3@gmail.com

Given two distributions X_0 and X_1, The proof system is quite simple, the verifier samples from either X_0 or X_1 and asks the prover to guess which one. This protocol is complete and sound with constant error, but the problem is that we do not get negligible simulator deviation in the security parameter (just constant). So, the entire work here is to review preliminary material to describe how to send the simulator deviation to negligible. The key trick is the Polarisation lemma, which alternately applies repeated sampling 1 and XOR sampling 2 .

1 Prelims

Fact 1 Given two vectors $\vec{X}$ and $\vec{Y}$, $\vec{X} \otimes \vec{Y}$ represents their tensor product (outer prodct).

$$||\vec{X} \otimes \vec{Y}||_1 = ||\vec{X}||_1 \cdot ||\vec{Y}||_1$$

Fact 2 Let $X = (X_0, X_1)$ be a joint distribution where X_0 and X_1 are independent. Similarly, define $Y = (Y_0, Y_1)$. Then we have

$$d_{\texttt{TV}}(X, Y) = d_{\texttt{TV}}(X_0, Y_0) + d_{\texttt{TV}}(X_1, Y_1) \tag{1}$$

Proof.

$$\begin{aligned} d_{\texttt{TV}}(X, Y) &= \frac{1}{2}\left|X_0 \otimes X_1 - Y_0 \otimes Y_1\right| && (2)\\ &\leq \frac{1}{2}\left|X_0 \otimes X_1 - Y_0 \otimes X_1\right| + \frac{1}{2}\left|Y_0 \otimes X_1 - Y_0 \otimes Y_1\right| && (3)\\ &= \frac{1}{2}\left|X_1 \otimes (X_0 - Y_0)\right| + \frac{1}{2}\left|Y_0 \otimes (X_1 - Y_1)\right| && (4)\\ &= \frac{1}{2}\left|X_1\right|\left|X_0 - Y_0\right| + \frac{1}{2}\left|Y_0\right|\left|X_1 - Y_1\right| && (5)\\ &= d_{\texttt{TV}}(X_0, Y_0) \cdot 1 + 1 \cdot d_{\texttt{TV}}(X_1, Y_1) && (6) \end{aligned}$$

Equation (3) from the triangle inequality and the last equality comes as X_1 and Y_0 are discrete probability distributions.

□

Lemma 1 (Direct Product Lemma) Let X and Y be distributions such that $d_{\texttt{TV}}(X, Y) = \delta$. Then for all $k \in \mathbb{N}$, we have

$$1 - 2\texttt{exp}(-\frac{k\delta^2}{2}) \leq d_{\texttt{TV}}(\otimes^k X, \otimes^k Y) \leq k\delta$$

where $\otimes^k X$ denotes k independent samples of X.

Proof. The upper bound follows directly from Fact 2 by replacing X_0 and X_1 as i.i.d samples from X, likewise with Y.

Let $\delta = d_{\mathtt{TV}}(X, Y)$ and define S^* as $\Pr[X \in S^*] - \Pr[Y \in S^*] = \delta$. Let $p = \Pr[Y \in S^*]$, then $\Pr[X \in S^*] = p + \delta$. Thus, here we have two Bernoulli random variables with mean p and $p + \delta$. By the Chernoff Bound (additive version):

1. The probability that at most $k(p + \frac{\delta}{2})$ fraction of k samples like in S is equivalent to

$$\Pr[\sum_{i \in [k]} X_i \leq k(p + \frac{\delta}{2})] = \Pr[(\frac{1}{k} \sum_{i \in [k]} X_i) - (p + \delta) \leq -\frac{\delta}{2}] \tag{7}$$

$$\leq \mathsf{exp}(-\frac{k}{2}\delta^2) \tag{8}$$

2. Similarly, the probability that at least $k(p + \frac{\delta}{2})$ fraction of k samples lie in S is at most $\mathsf{exp}(-\frac{k}{2}\delta^2)$. (This is just the other tail of the Chernoff bound, writing it out explicitly like above makes it pretty obvious).

$$\Pr[\sum_{i \in [k]} Y_i \geq k(p + \frac{\delta}{2})] = \Pr[(\frac{1}{k} \sum_{i \in [k]} Y_i) - p \geq \frac{\delta}{2}] \tag{9}$$

$$\leq \mathsf{exp}(-\frac{k}{2}\delta^2) \tag{10}$$

Define S' as a set of all k-tuples such that:

$$S' = \left\{(z_1, \ldots, z_k) \in (\{0,1\}^n)^k : \text{at least } k(p + \frac{\delta}{2}) \text{ fraction of samples, lie in } S^*\right\}$$

Thus we have,

$$d_{\mathtt{TV}}(\otimes^k X, \otimes^k Y) \geq \Pr[\otimes^k X \in S'] - \Pr[\otimes^k Y \in S'] \tag{11}$$

$$\geq 1 - 2\mathsf{exp}(-\frac{k}{2}\delta^2) \tag{12}$$

Equation (11) comes from the definition of statistical difference. Equation (12) come from $\Pr[\otimes^k Y \in S'] \leq \mathsf{exp}(-\frac{k}{2}\delta^2)$ as that is the definition of item 2. And $\Pr[\otimes^k X \in S'] \geq 1 - \mathsf{exp}(-\frac{k}{2}\delta^2)$ from the definition given in item (1).

□

The Direct product lemma sends the YES instances to have a distance negligibly far from one at an exponential rate. Still, it also blows up the soundness error for NO instances linearly (in the security parameter) close to 1. We need a way to send the YES instances to have a distance close to 1 and the NO instances to 0, both at exponential rates. This is where the XOR lemma comes in.

Lemma 2 (XOR Lemma) Given a pair of distributions $X_0, X_1 \in \Delta(\Omega_{\mathsf{n}})$ and $k \in \mathbb{N}$ as input, there exists a polynomial time computable function that outputs a pair of distributions (Y_0, Y_1) such that

$$d_{\texttt{TV}}(Y_0, Y_1) = d_{\texttt{TV}}(X_0, X_1)^k$$

Specifically, Y_0, Y_1 are defined as follows:

Y_0: Uniformly select $(b_1, \dots, b_k) \in \{0,1\}^k$ such that $\oplus_{i=1}^k b_i = 0$, and set $Y_0 = (X_{b_1} \otimes \cdots \otimes X_{b_k})$.

Y_1: Uniformly select $(b_1, \dots, b_k) \in \{0,1\}^k$ such that $\oplus_{i=1}^k b_i = 1$, and set $Y_1 = (X_{b_1} \otimes \cdots \otimes X_{b_k})$.

Proof. Prove the base case with $k = 2$. Then the rest follows from induction. Define a_i such that all $i \in [k]$, we have $a_i \in \{0,1\}$.

Let $X_0^{(1)}, X_1^{(1)}, X_0^{(2)}, X_1^{(2)}$ be random variables such that

$Z_0 = X_{a_1}^{(1)} \otimes X_{a_2}^{(2)}$ such that $a_1 \oplus a_2 = 0$

$Z_1 = X_{a_1}^{(1)} \otimes X_{a_2}^{(2)}$ such that $a_1 \oplus a_2 = 1$

$$\begin{aligned}
d_{\texttt{TV}}(Z_0, Z_1) &= \frac{1}{2}|Z_0 - Z_1| \\
&= \frac{1}{2}\left|\frac{1}{2}[(X_0^{(1)} \otimes X_0^{(2)}) + (X_1^{(1)} \otimes X_1^{(2)})] - \frac{1}{2}[(X_1^{(1)} \otimes X_0^{(2)}) + (X_0^{(1)} \otimes X_1^{(2)})]\right| \\
&= \frac{1}{4}\left|(X_0^{(1)} - X_1^{(1)}) \otimes (X_0^{(2)} - X_1^{(2)})\right| \\
&= \frac{1}{2}||X_0^{(1)} - X_1^{(1)}||_1 \cdot \frac{1}{2}||X_0^{(2)} - X_1^{(2)}||_1 \\
&= d_{\texttt{TV}}(X_0^{(1)}, X_1^{(1)})\ d_{\texttt{TV}}(X_0^{(2)}, X_1^{(2)})
\end{aligned}$$

Induction assumption for $k > 2$, we have random variables $\{X_0^{(i)}\}_{i \in [k]}$ and $\{X_1^{(i)}\}_{i \in [k]}$ and that

$$d_{\texttt{TV}}(Z_0, Z_1) = d_{\texttt{TV}}(X_0^{(1)}, X_1^{(1)}) \dots d_{\texttt{TV}}(X_0^{(k)}, X_1^{(k)})$$

When we add variables $X_0^{(k+1)}$ and $X_1^{(k+1)}$, and computing the statistical distance between Z_0 and Z_1 reduces to the base case again as exactly half of the values will have parity 0 and a half will have parity 1. Thus the lemma holds by induction.

□

Lemma 3 (Polarisation Lemma) Let $\alpha, \beta \in [0,1]$ such that $\alpha^2 > \beta$. There is a PPT function $\mathsf{Polarise}_{\alpha,\beta}$, that takes a triple $(X_0, X_1, 1^\kappa)$, where X_0 and X_1 are distributions encoded by circuits, and outputs (Y_0, Y_1) such that

$$d_{\texttt{TV}}(X_0, X_1) \geq \alpha \implies d_{\texttt{TV}}(Y_0, Y_1) \geq 1 - 2^{-k}$$
$$d_{\texttt{TV}}(X_0, X_1) \leq \beta \implies d_{\texttt{TV}}(Y_0, Y_1) \leq 2^{-k}$$

Proof. Let $\lambda = \min\{\alpha^2/\beta, 2\} > 1$, and let $l = \lceil \log_\lambda 4\kappa \rceil = O(\log \kappa)$.

Apply the XOR lemma 2 to $(X_0, X_1, 1^l)$ to get (X_0', X_1') such that

$$\begin{aligned} d_{\mathtt{TV}}(X_0, X_1) \geq \alpha &\implies d_{\mathtt{TV}}(X_0', X_1') \geq \alpha^l \\ d_{\mathtt{TV}}(X_0, X_1) \leq \beta &\implies d_{\mathtt{TV}}(X_0', X_1') \leq \beta^l \end{aligned}$$

Let $m = \frac{\lambda^l}{2\alpha^{2l}} \leq \frac{1}{2\beta^l}$. Notice $m \leq \mathtt{poly}(\kappa)$, since $l = O(\log \kappa)$, and $\lambda \leq 2$, and $\alpha \in [0, 1]$ is constant. Applying the direct product lemma 1 we get $X_0'' = \otimes^m X_0'$ and $X_1'' = \otimes^m X_1'$.

$$\begin{aligned} d_{\mathtt{TV}}(X_0, X_1) \geq \alpha &\implies d_{\mathtt{TV}}(X_0'', X_1'') \geq 1 - 2\mathsf{exp}\left(\frac{\lambda^l}{2\alpha^{2l}} \cdot \frac{(\alpha^l)^2}{2}\right) \geq 1 - 2e^{-\kappa} \\ d_{\mathtt{TV}}(X_0, X_1) \leq \beta &\implies d_{\mathtt{TV}}(X_0', X_1') \leq m\beta^l \leq (\frac{1}{2\beta^l})\beta^l = \frac{1}{2} \end{aligned}$$

Apply the XOR lemma again to the triple $(X_0'', X_1'', 1^\kappa)$ such that

$$\begin{aligned} d_{\mathtt{TV}}(X_0, X_1) \geq \alpha &\implies d_{\mathtt{TV}}(Y_0, Y_1) \geq 1 - 2\kappa e^{-\kappa} > 1 - 2^\kappa \\ d_{\mathtt{TV}}(X_0, X_1) \leq \beta &\implies d_{\mathtt{TV}}(Y_0, Y_1) \leq \frac{1}{2^\kappa} \end{aligned}$$

□

2 Main Result

Next we present the Statistical ZK Proof For SD.

Input: Circuits X_0 and X_1, and security param 1^κ.
Set $\alpha = 2/3$ and $\beta = 1/3$

1. Both prover and verifier compute $(Y_0, Y_1) = \mathsf{Polarise}_{\alpha,\beta}(X_0, X_1, 1^{\kappa-1})$.
2. V: Select $b \xleftarrow{\$} \{0, 1\}$. Then sample $x \xleftarrow{\$} Y_b$ and send x to P.
3. P: If $\Pr[Y_0 = x] > \Pr[Y_1 = x]$ send $c = 0$ to the verifier, otherwise send $c = 1$.
4. V: If $c = b$ accept.

Next, we present the simulator.

Inputs: Polarised circuits Y_0 and Y_1

1. Sample $b \xleftarrow{\$} \{0, 1\}$ and let $x \xleftarrow{\$} Y_b$.
2. Let $c = b$
3. Output (x, c; b)

Fact 3 Let A and B be two discrete distributions on $\mathcal{X}$. Let $S_A = \{x \in \mathcal{X} : \Pr[A = x] > \Pr[B = x]\}$ and $S_B = \{x \in \mathcal{X} : \Pr[B = x] > \Pr[A = x]\}$. Then

$$d_{\mathtt{TV}}(A, B) = \Pr[A \in S_A] - \Pr[B \in S_B] \tag{13}$$

Lemma 4 When two distributions Y_0 and Y_1 have statistical difference δ, then the proof system above makes the verifier accept with probability $\frac{1+\delta}{2}$, and no prover strategy succeeds with greater probability. Thus, we get soundness error $\frac{1+\delta}{2}$. The simulator deviation of the simulator defined above is $\frac{1-\delta}{2}$ which is identical to the completeness error.

Proof. Let
$S_{X_0} = \{x \in \{0,1\}^n : \Pr[X_0 = x] > \Pr[X_1 = x]\}$ and $S_{X_1} = \{x \in \{0,1\}^n : \Pr[X_1 = x] > \Pr[X_0 = x]\}$.

Consider the verifier strategy, which gives the same distribution as the prescribed verifier strategy in the proof system above.

1. Flip a coin d that is 0 with probability $1 - \delta$ and 1 with probability δ.
2. If $d = 0$, pick $x \in \{0,1\}^n$ and $b \overset{\$}{\leftarrow} \{0,1\}$.
3. If $d = 1$, pick $b \overset{\$}{\leftarrow} \{0,1\}$, and then pick $x \overset{\$}{\leftarrow} S_{X_b}$
4. Output (b, x)

When $d = 0$, b is independent of x and provers success of probability is exactly $\frac{1}{2}$ no matter what it does. However, if a verifier sends to the prover $x \in S_{X_0}$ or $x \in S_{X_1}$, it can predict x_b with probability 1. So for the prover gets the verifier to accept with probability

$$\frac{1}{2} \cdot \Pr[d = 0] + 1 \cdot \Pr[d = 1] = \frac{1}{2}(1 - \delta) + \delta = \frac{1 + \delta}{2}$$

Note that simulator deviation and completeness error are the same, as the simulator always guesses right. In the real protocol, the fraction of times the prover guesses the verifiers challenge incorrectly, despite being honest, is exactly the complement of the probability with which it gets the verifier to accept i.e. $1 - \frac{1+\delta}{2} = \frac{1-\delta}{2}$.

Now as the protocol is conducted on Y_0, Y_1 and not X_0, X_1, we are guaranteed that in the YES case, $\delta = 1 - 2^{\kappa}$, thus the simulator deviation is $2^{-\kappa}$. □

References