On Round-By-Round Soundness and State Restoration Attacks

Justin Holmgren*

Abstract

We show that the recently introduced notion of round-by-round soundness for interactive proofs (Canetti et al.; STOC 2019) is equivalent to the notion of soundness against state restoration attacks (Ben-Sasson, Chiesa, and Spooner; TCC 2016). We also observe that neither notion is implied by the random-oracle security of the Fiat-Shamir transform.

1 Introduction

The Fiat-Shamir transform [FS86] is a heuristic methodology for using a hash family $\mathcal{H}$ to convert a public-coin interactive protocol Π (either a proof or argument) into a non-interactive protocol $\mathsf{FS}[\Pi, \mathcal{H}]$. In this protocol, a hash function $H \leftarrow \mathcal{H}$ is first chosen as a public parameter. A proof for a claim x then consists of messages $(\alpha_1, \ldots, \alpha_r)$ such that with $\beta_i = H(\alpha_1, \beta_1, \ldots, \alpha_i)$, the transcript $(\alpha_1, \beta_1, \ldots, \alpha_r, \beta_r)$ is accepted on input x in Π. It is also often convenient to model $\mathcal{H}$ as a random oracle, in which case we will denote the resulting random oracle protocol by $\mathsf{FS}^{\mathsf{RO}}[\Pi]$.

It is known that $\mathsf{FS}^{\mathsf{RO}}[\Pi]$ is sound for all constant-round protocols Π [PS96] and, more generally, for all protocols Π that resist *state restoration attacks* [BCS16]. In a state restoration attack, a malicious prover P^* interacting with a verifier V may at any point reset V to a state that V was previously in. Then, P^* may continue to interact with V, with V using fresh randomness.

Returning our attention to the soundness of Fiat-Shamir in the plain model, the state of the art is that $\mathsf{FS}[\Pi, \mathcal{H}]$ is (computationally) sound if Π is *round-by-round sound* [CCH$^+$19] and $\mathcal{H}$ is correlation intractable [CGH04]. Round-by-round soundness stipulates that there is a way to label certain transcript prefixes as "doomed" relative to an input x such that:

- If x is an input that represents a false claim, then the empty transcript $\emptyset$ is doomed relative to x.
- If τ is any transcript prefix (ending in a verifier message) that is doomed relative to x, then for all choices α of the prover's next messages, it holds with overwhelming probability over β that $\tau|\alpha|\beta$ is also doomed relative to x.
- If τ is a complete transcript that is doomed relative to x, then the verifier on input x will reject the transcript τ.

These two results and two subclasses of public-coin interactive proofs naturally raise the question:

> *What is the relation between soundness against state-restoration attacks and round-by-round soundness?*

It was observed by [CCH$^+$19] that if a protocol Π is round-by-round sound, then Π is also sound against state restoration attacks. Proving the converse (or indeed instantiating Fiat-Shamir by any means for this potentially broader class of protocols) was left as an open question.

In this work, we show that the converse holds.

*Simons Institute. Email: `holmgren@alum.mit.edu`.

Theorem 1.1. *For any public-coin protocol* Π*, if* Π *is sound against state restoration attacks, then* Π *is round-by-round sound.*

We also show that soundness against state restoration attacks is a strictly stronger notion for a protocol Π than the soundness of $\mathsf{FS}^{\mathsf{RO}}[\Pi]$.

Theorem 1.2. *There exists a public-coin interactive proof* Π *such that* Π *is* unsound *against state restoration attacks, but* $\mathsf{FS}^{\mathsf{RO}}[\Pi]$ *is secure.*

Our separation leverages the fact that in a state restoration attack a prover may rewind to the same state multiple times, each time obtaining a freshly random verifier messages. On the other hand, in $\mathsf{FS}^{\mathsf{RO}}[\Pi]$, verifier messages are deterministically generated as a function of the random oracle and the preceding partial transcript.

2 Preliminary Definitions

2.1 Interactive Protocols

It will be convenient for us to consider separately from interactive proofs (which are associated with a language $\mathcal{L}$, involve an input x, and have completeness / soundness properties depending on whether $x \in \mathcal{L}$) a notion of an interactive game, which has no input.

We think of an interactive game as something that is played by a single player in r rounds. At the beginning of the i^{th} round, the player must specify a message $\alpha_i \in \{0,1\}^*$. Then, a message β_i is sampled uniformly from $\{0,1\}^{\ell_i}$ for some ℓ_i that is pre-specified independently of any of the player's choices. At the end of the r^{th} round, a predicate W is applied to $(\alpha_1, \beta_1, \ldots, \alpha_r, \beta_r)$ to determine whether the player wins.

More formally:

Definition 2.1 (Interactive Game)**.** An (r-round) public-coin interactive game is a tuple $(\ell_1, \ldots, \ell_r, W)$, where each $\ell_i \in \mathbb{Z}^+$ and $W \subseteq \{0,1\}^*$ is an "acceptance" set. A strategy is a function $s : \{0,1\}^* \to \{0,1\}^*$.

If $\mathcal{G} = (\ell_1, \ldots, \ell_r, W)$ is a public-coin interactive game and s is a strategy, then the value of $\mathcal{G}$ with respect to s (alternatively the probability with which s wins $\mathcal{G}$) is

$$v[s](\mathcal{G}) \stackrel{\mathsf{def}}{=} \Pr_{\substack{\beta_1 \leftarrow \{0,1\}^{\ell_1} \\ \cdots \\ \beta_r \leftarrow \{0,1\}^{\ell_r}}} \left[(\alpha_1, \beta_1, \ldots, \alpha_r, \beta_r) \in W \right],$$

where each α_i is defined to be $s(\beta_1, \ldots, \beta_{i-1})$. The value of $\mathcal{G}$, denoted $v(\mathcal{G})$, is $\sup_s v[s](\mathcal{G})$.

Definition 2.2 (Interactive Proof)**.** An ($r(\cdot)$-round) public-coin interactive proof for a language $\mathcal{L}$ with soundness error $\epsilon(\cdot)$ is a pair (P, V), where V is a polynomial-time algorithm mapping any string $x \in \{0,1\}^*$ to an $r(|x|)$-round single-player game with the following properties:

- (Completeness) If $x \in \mathcal{L}$, then $P(x)$ is a strategy that wins $V(x)$ with probability 1.
- (Soundness) If $x \notin \mathcal{L}$, then *all* strategies P^* win $V(x)$ with probability at most $\epsilon(|x|)$.

The interactive proof is said to be public-coin if each $V(x)$ is public-coin.

Definition 2.3 (Game Transcript)**.** If $\mathcal{G} = (\ell_1, \ldots, \ell_r, W)$ is a public-coin interactive game, then a (complete) transcript for $\mathcal{G}$ is $\alpha_1|\beta_1|\cdots|\alpha_r|\beta_r$ with each $\beta_i \in \{0,1\}^{\ell_i}$ and $\alpha_i \in \{0,1\}^*$. An accepting transcript is one that is contained in W. A transcript prefix is any $\alpha_1|\beta_1|\cdots|\alpha_i|\beta_i$ for $i \in \{0, \ldots, r\}$.

Definition 2.4 (Game Suffix)**.** If $\mathcal{G} = (\ell_1, \ldots, \ell_r, W)$ is an r-round public-coin interactive game and $\alpha_1|\beta_1|\cdots|\alpha_i|\beta_i$ is a transcript prefix for $\mathcal{G}$, we denote by $\mathcal{G}|_\tau$ the game $(\ell_{i+1}, \ldots, \ell_r, W|_\tau)$, where $W|_\tau$ is the set of strings of the form $\alpha_{i+1}|\beta_{i+1}|\cdots|\alpha_r|\beta_r$ for which $\alpha_1|\beta_1|\cdots|\alpha_r|\beta_r \in W$.

We refer to $\mathcal{G}|_\tau$ as the suffix of $\mathcal{G}$ following τ.

2.2 Notions of Soundness

Let $\mathcal{L}$ be a language and let $\Pi = (P, V)$ be a public-coin interactive proof for $\mathcal{L}$. Recall the following definition from [CCH+19]. Suppose without loss of generality that all verifier messages are of length ℓ.

Definition 2.5 (Round-by-Round Soundness Error [CCH+19])**.** Π has round-by-round soundness error $\epsilon(\cdot)$ if there exists a "doomed set" $\mathcal{D} \subseteq \{0,1\}^*$ such that the following properties hold:

1. If $x \notin L$, then $(x, \emptyset) \in \mathcal{D}$, where $\emptyset$ denotes the empty transcript.
2. If $(x, \tau) \in \mathcal{D}$ for a transcript prefix τ, then for every potential prover next message α, it holds that

$$\Pr_{\beta \leftarrow \{0,1\}^\ell}\Big[\big(x, \tau|\alpha|\beta\big) \notin \mathcal{D}\Big] \leq \epsilon(n)$$

3. For any complete transcript τ, if $(x, \tau) \in \mathcal{D}$ then $V(x, \tau) = 0$.

Definition 2.6 (Asymptotic Round-by-Round Soundness [CCH+19])**.** Π is said to be round-by-round sound if there is a negligible function ϵ such that Π has round-by-round soundness error ϵ.

To define soundness of public-coin interactive proofs against state restoration attacks, we first define corresponding notions for public-coin interactive *games*.

Definition 2.7. For any public-coin interactive game $\mathcal{G} = (\ell_1, \ldots, \ell_r, W)$ and any query-bound q, we define a corresponding q-query state restoration game $\mathsf{SR}^q(\mathcal{G})$. We only informally describe how this game is played:

1. A referee initializes a set $S := \{\emptyset\}$, where $\emptyset$ denotes the empty transcript.
2. Up to q times, P^* may specify a pair (τ, α) where $\tau = \alpha_1|\beta_1|\cdots|\alpha_i|\beta_i \in S$ and $\alpha \in \{0,1\}^*$. The referee samples $\beta \leftarrow \{0,1\}^{\ell_{i+1}}$, and adds $\tau|\alpha|\beta$ to S.
3. P^* wins if S contains any $\tau \in W$.

In our notation, the notion of state restoration soundness from [BCS16] can be formulated as follows.

Definition 2.8 (State Restoration Soundness [BCS16])**.** For functions $q : \mathbb{Z}^+ \to \mathbb{Z}^+$ and $\epsilon : \mathbb{Z}^+ \to \mathbb{R}$, a public-coin interactive proof (P, V) for $\mathcal{L}$ is said to be (q, ϵ)-sound against state restoration attacks if for all n and all $x \in \{0,1\}^n \setminus \mathcal{L}$, the value of $\mathsf{SR}^{q(n)}\big(V(x)\big) \leq \epsilon(n)$.

Π is said simply to be sound against state restoration attacks if for all polynomially bounded $q : \mathbb{Z}^+ \to \mathbb{Z}^+$, there is a negligible function ϵ such that Π is (q, ϵ)-sound against state restoration attacks.

3 Proof of Theorem 1.1

Let $\mathcal{L}$ be a language, and let $\Pi = (P, V)$ be an $r(\cdot)$-round public-coin interactive proof for $\mathcal{L}$. For simplicity suppose that all verifier messages are of length $\ell = \ell(n)$.

Proposition 3.1. *Let $\mathcal{G}$ be a public-coin interactive game, and let $\tau = \alpha_1|\beta_1|\cdots|\alpha_i|\beta_i$ be a transcript prefix for $\mathcal{G}$.*

If $v\big(\mathsf{SR}^q(\mathcal{G}|_\tau)\big) \leq \epsilon$, then for all $q' < q$, all $\epsilon' > \epsilon$, and all $\alpha \in \{0,1\}^*$, it holds that

$$\Pr_{\beta \leftarrow \{0,1\}^{\ell(|x|)}}\Big[v\big(\mathsf{SR}^{q'}(\mathcal{G}|_{\tau|\alpha|\beta})\big) > \epsilon'\Big] \leq -\frac{\ln(\epsilon' - \epsilon)}{q - q'}. \tag{1}$$

Proof. For any α, let p_α denote the left-hand side of Eq. (1). Consider the following (informally specified) strategy for $\mathsf{SR}^q(\mathcal{G}|_\tau)$.

1. Specify (τ, α) repeatedly. Specifically, do so $q - q'$ times. Let S be the set as in the definition of $\mathsf{SR}^q(\mathcal{G}|_\tau)$ (Definition 2.7).

2. Let β be such that $\tau|\alpha|\beta \in S$ and $v\big(\mathsf{SR}^{q'}(\mathcal{G}|_{\tau|\alpha|\beta})\big)$ is maximal.

3. From this point on, P^* plays according to an optimal strategy for $\mathsf{SR}^{q'}(\mathcal{G}|_{\tau|\alpha|\beta})$.

In order for this strategy to not contradict the assumption that $v\big(\mathsf{SR}^q(\mathcal{G}|_\tau)\big) \le \epsilon$, it must hold with probability at least $\epsilon' - \epsilon$ that at the beginning of Step 2, for all β with $\tau|\alpha|\beta \in S$, $v\big(\mathsf{SR}^{q'}(\mathcal{G}|_{\tau|\alpha|\beta})\big) \le \epsilon'$. Because each β is chosen independently, this is equivalent to saying that $(1 - p_\alpha)^{q-q'} \ge \epsilon' - \epsilon$. Thus

$$p_\alpha \le 1 - (\epsilon' - \epsilon)^{\frac{1}{q-q'}} = 1 - e^{\frac{\ln(\epsilon'-\epsilon)}{q-q'}} \le -\frac{\ln(\epsilon' - \epsilon)}{q - q'}. \qquad \square$$

Theorem 3.2. *If Π is (q, ϵ)-sound against state-restoration attacks for $\epsilon < 1$, then it has round-by-round soundness error $\frac{r}{q} \cdot \ln\left(\frac{2r}{1-\epsilon}\right)$.*

Proof. Define $\Delta\epsilon = \frac{1-\epsilon}{2r}$ and $\Delta q = \frac{q}{r}$. Define the set $\mathcal{D} \subseteq \{0,1\}^*$ such that if τ is an i-round transcript prefix for $V(x)$, then $(x, \tau) \in \mathcal{D}$ if and only if $v\big(\mathsf{SR}^{q-i\cdot\Delta q}(V(x)|_\tau)\big) \le \epsilon + i \cdot \Delta\epsilon$.

We now show that $\mathcal{D}$ satisfies the requirements of Definition 2.5.

Claim 3.3. *For $x \notin \mathcal{L}$, $(x, \emptyset) \in \mathcal{D}$ where $\emptyset$ denotes the empty transcript.*

Proof. We have

$$v\big(\mathsf{SR}^{q-0\cdot\Delta q}(V(x)|_\emptyset)\big) = v\big(\mathsf{SR}^q(V(x))\big),$$

which by assumption that Π is (q, ϵ)-sound, must be bounded by ϵ. Thus $(x, \emptyset) \in \mathcal{D}$. $\square$

Claim 3.4. *For all x, τ, if $(x, \tau) \in \mathcal{D}$ then for all α,*

$$\Pr_{\beta \leftarrow \{0,1\}^{\ell(|x|)}} [(x, \tau|\alpha|\beta) \notin \mathcal{D}] \le \frac{r}{q} \cdot \ln\left(\frac{2r}{1-\epsilon}\right).$$

Proof. Suppose that τ is an i-round transcript prefix. Then by definition of $\mathcal{D}$ we have $v\big(\mathsf{SR}^{q-i\cdot\Delta q}(V(x)|_\tau)\big) \le \epsilon + i \cdot \Delta\epsilon$. Then for any α, we have

$$\Pr_{\beta \leftarrow \{0,1\}^{\ell(|x|)}} [(x, \tau|\alpha|\beta) \notin \mathcal{D}] = \Pr_{\beta \leftarrow \{0,1\}^{\ell(|x|)}} \left[v\big(\mathsf{SR}^{q-(i+1)\cdot\Delta q}(V(x)|_{\tau|\alpha|\beta})\big) > \epsilon + (i+1) \cdot \Delta\epsilon\right].$$

By Proposition 3.1, this is bounded by $-\frac{\ln(\Delta\epsilon)}{\Delta q} = \frac{r}{q} \cdot \ln\left(\frac{2r}{1-\epsilon}\right)$. $\square$

Claim 3.5. *For any x and any complete transcript τ, if $(x, \tau) \in \mathcal{D}$, then $V(x, \tau) = 0$.*

Proof. This follows from the fact that for any complete transcript τ, either τ is an accepting transcript for $V(x)$ or it is not, and the definition of $\mathcal{D}$ implies that the probability that τ is accepting for $V(x)$ is at most $\epsilon + r \cdot \Delta\epsilon = \frac{1+\epsilon}{2} < 1$. $\square$

This completes the proof of Theorem 3.2. $\square$

Theorem 1.1 follows as a corollary, also using Proposition 3.6 below.

Proposition 3.6. *If Π is sound against state restoration attacks, then there exists a* super-polynomial *q and a negligible function ϵ such that Π is (q, ϵ)-sound against state restoration attacks.*

Proof. Suppose that Π is sound against state restoration attacks. This implies that there exist $1 = N_0 < N_1 < N_2 < \cdots$ such that for all $n \geq N_c$ and all $x \in \{0,1\}^n \setminus \mathcal{L}$, $v\left(\mathsf{SR}^{n^c}(V(x))\right) \leq n^{-c}$.

Define $q : \mathbb{Z}^+ \to \mathbb{Z}^+$ as follows. For any n, let c be such that $N_c \leq n < N_{c+1}$ and define $q(n) = n^c$. It follows by definition that $q(n) \geq n^{\omega(1)}$ and $\max_{x \in \{0,1\}^n \setminus \mathcal{L}} \left\{ v\left(\mathsf{SR}^{q(n)}(V(x))\right)\right\} \leq n^{-\omega(1)}$. □

We remark that Proposition 3.6 is very similar to an observation of Bellare [Bel02] that there is no difference between the following two types of security definition:

- For every polynomial-time adversary $\mathcal{A}$, there exists a negligible function ϵ bounding $\mathcal{A}$'s advantage in breaking the primitive.
- There exists a negligible function ϵ such that for all polynomial-time adversaries $\mathcal{A}$, ϵ bounds the advantage of $\mathcal{A}$ in breaking the primitive.

4 Proof of Theorem 1.2

Let $r(\cdot)$ be any function with $r(n) = \omega(1)$, and consider the r-round public-coin interactive proof $\Pi = (P, V)$ for the empty language in which all verifier messages are $\log n$-bit strings. The verifier accepts if the prover sent only empty strings, and all of the verifier's messages were the all-zero string. It is easy to see that $\mathsf{FS}[\Pi, \mathcal{H}]$ has soundness error equal to

$$\Pr_{H \leftarrow \mathcal{H}}\left[\forall i \in [r(n)], H(0^{(i-1)\cdot \log n}) = 0^{\log n}\right],$$

which is negligible if $\mathcal{H}$ is replaced by a random oracle.

However, because each verifier message has only $\log n$ bits, Π can only possibly have round-by-round soundness error ϵ if $\epsilon \geq \frac{1}{n}$.

Acknowledgments

We thank Fermi Ma and Ron Rothblum for helpful comments on an early draft of this work.

References

[BCS16] Eli Ben-Sasson, Alessandro Chiesa, and Nicholas Spooner, *Interactive oracle proofs*, TCC (B2), Lecture Notes in Computer Science, vol. 9986, 2016, pp. 31–60.

[Bel02] Mihir Bellare, *A note on negligible functions*, J. Cryptology **15** (2002), no. 4, 271–284.

[CCH+19] Ran Canetti, Yilei Chen, Justin Holmgren, Alex Lombardi, Guy N. Rothblum, Ron D. Rothblum, and Daniel Wichs, *Fiat-shamir: from practice to theory*, STOC, ACM, 2019, pp. 1082–1090.

[CGH04] Ran Canetti, Oded Goldreich, and Shai Halevi, *The random oracle methodology, revisited*, J. ACM **51** (2004), no. 4, 557–594.

[FS86] Amos Fiat and Adi Shamir, *How to prove yourself: Practical solutions to identification and signature problems*, Conference on the Theory and Application of Cryptographic Techniques, Springer, 1986, pp. 186–194.

[PS96] David Pointcheval and Jacques Stern, *Security proofs for signature schemes*, EUROCRYPT, Lecture Notes in Computer Science, vol. 1070, Springer, 1996, pp. 387–398.