Electronic Colloquium on Computational Complexity, Report No. 50 (2026)

Witness-Indistinguishable Arguments of Knowledge and One-Way Functions

Gal Arnon[1], Noam Mazor[2], Rafael Pass[3], and Jad Silbak[4]

[1]Bocconi University, galarnon42@gmail.com
[2]New York University, noammaz@gmail.com
[3]Technion, Cornell Tech, and Tel Aviv University, rafael@cs.cornell.edu
[4]MIT, jadsilbak@gmail.com

April 7, 2026

Abstract

In this paper we study the cryptographic complexity of non-trivial witness-indistinguishable (WI) arguments of knowledge. We establish that:

- Assuming that $\mathsf{NP} \not\subseteq \mathsf{P}/\mathrm{poly}$, the existence of a constant-round computational WI argument of knowledge for NP implies that (infinitely-often) auxiliary-input one-way functions exist.
- Assuming that $\mathsf{NP} \not\subseteq \mathsf{P}^{\mathrm{Sam}}/\mathrm{poly}$, there is no black-box construction of a constant-round (unbounded-verifier) statistical WI argument of knowledge from one-way permutations. Here, Sam is the collision finder oracle of Haitner, Hoch, Reingold, and Segev [FOCS '07].

Moreover, we identify a natural class of knowledge extractors for which stronger versions of the above implications hold (e.g., even if the protocols have many rounds).

ISSN 1433-8092

Contents

1 Introduction

A major task of complexity-theoretic cryptographic research is to identify the minimal assumptions necessary for various cryptographic tasks. A central primitive in this line of research is a *one-way function* [DH76], a function f that can be efficiently computed but cannot be inverted in polynomial time. One-way functions are known to be both necessary [IL89] and sufficient for many of the most central cryptographic primitives and protocols (e.g., pseudorandom generators [BM82; HILL99], pseudorandom functions [GGM86], private-key encryption [GM84], digital signatures [NY89; Rom90], commitment schemes [Nao91; HR07b], identification protocols [FS86], coin-flipping protocols [Blu83], and more). This fact motivated a long line of research that explores the necessity or implication of the nonexistence of one-way functions on various cryptographic primitives (e.g., [IL89; Ost91; OW93; BIKM99; Vad06; HO14; BHT18; PV20; AR21; KMNPRY22; HN24; LMP24; CHK25; CLV26; CHKT26] and many more). In this work we focus on this question through the lens of interactive proof systems [GMR89], where an untrusted prover tries to convince a bounded verifier of the validity of a statement.

The most famous cryptographic variant of interactive proofs is zero-knowledge proofs [GMR89], where the verifier learns nothing about the statement beyond its validity. Ostrovsky and Wigderson [Ost91; OW93] show that auxiliary-input one-way functions[1] are necessary for zero-knowledge assuming worst-case hardness of NP.[2] Recently, Hirahara and Nanashima [HN24] extend these results to show that, assuming worst-case hardness of NP, zero-knowledge proofs imply (standard) one-way functions.

Due to zero-knowledge being a very powerful property, it comes with undesirable drawbacks (e.g., [GO94]), with one notable example being the inability to apply parallel repetition. Feige and Shamir [FS90] observe that, while zero-knowledge is not preserved, a weaker property called "witness indistinguishability" (WI) is. In a WI proof system, the only thing hidden from the verifier is which of a pair of witnesses the prover used. This is formalized by requiring that the verifier's views when interacting with the prover using either witness be computationally indistinguishable. While they are weaker than zero-knowledge, WI proofs have found many uses and have been the focus of a significant body of research (e.g., [FS90; FLS99; Bar01; DN07; BOV07; BP15; KKS18; BKPRV24] to name just a few examples). While all known constructions of WI protocols rely on the existence of one-way functions, it remains open to establish whether this reliance is necessary.

Limitations for Deriving OWFs from WI. In recent work, Baril and Haitner [BH26] establish barriers for showing that WI protocols imply one-way functions. More precisely, they consider fully black-box reductions from one-way functions to witness-indistinguishable interactive arguments and an additional hardness assumption G (e.g., $\mathsf{NP} \not\subseteq \mathsf{P/poly}$). Such a fully black-box reduction guarantees that any one-way function inverter can be used by the reduction to either break soundness of the protocol or to break G. [BH26] prove that the WI argument can be essentially removed from any such reduction, yielding a reduction from one-way functions to G alone. Thus, beyond what is already implied by G, witness-indistinguishable interactive arguments have no black-box implications for the existence of one-way functions. In fact, they show this even if the protocol is succinct, and even if the WI is replaced with perfect zero-knowledge, as long as the reduction does not have access to the simulator.

1.1 Our Contributions

In light of the limitations proven by [BH26], we consider witness indistinguishable arguments with the added property of *knowledge soundness*, known as WI arguments of knowledge ($\mathsf{WI\text{-}AoK}$).

In an argument of knowledge [GMR89; BG92], in addition to the prover P and the verifier V, we assume that there exists a "knowledge extractor" E such that: for every (efficient) malicious prover $\widehat{P}$, if $\widehat{P}$ is capable of convincing V to accept an instance x with high probability, then E, given x and oracle access to $\widehat{P}$ is capable of extracting a valid witness for x. Knowledge soundness is a natural and common property (in fact, the trivial NP proof system where the witness is sent directly satisfies this notion), and is especially useful in conjunction with other properties such as zero-knowledge, witness-indistinguishability, or succinctness.

[1] An auxiliary-input one-way function is a function family $\{f_a\}$ so that for every efficient adversary A there exists a value a so that f_a is hard for A to invert (even knowing a).

[2] [OW93] additionally show that, assuming the *average-case* hardness of NP, *standard* one-way functions are necessary for zero-knowledge.

1.1.1 Cryptography is necessary for nontrivial WI-AoK

Our main result establishes that cryptography is necessary for constructing constant-round computationally witness-indistinguishable arguments of knowledge.

Theorem 1.1 (informal, Corollary 7.3)**.** *Assume that* NP $\not\subseteq$ P/poly*. If there is a* constant-round *WI-AoK for* NP *then infinitely-often auxiliary-input one-way functions exist.*

Recall that one-way functions are also sufficient for constructing WI-AoK.

Toward OWF from honest-verifier WI-AoK. When the knowledge extractor of the protocol is *"weakly non-adaptive"* we strengthen our results to hold even when the protocol: (1) has many rounds and (2) witness-indistinguishability only has to hold for the honest verifier. Moreover, under the mildly stronger assumption that NP $\not\subseteq$ ioP/poly we construct full-fledged auxiliary-input one-way functions, as opposed to infinitely-often. In contrast to a (fully) non-adaptive extractor that chooses all of its queries in advance, roughly, a weakly non-adaptive extractor chooses its rewinding strategy independently of the prover's replies, but its queries in each rewinding branch are allowed to depend on the partial transcript corresponding to this branch.[3]

Theorem 1.2 (informal, Corollary 7.4)**.** *Assume that* NP $\not\subseteq$ ioP/poly*. If there is an honest-verifier WI-AoK for* NP *with a weakly non-adaptive extractor then auxiliary-input one-way functions exist.*

The above result provides evidence that Theorem 1.1 should hold even for many-round protocols satisfying only honest-verifier WI. Indeed, most natural protocols have weakly non-adaptive extractors,[4] such as sigma protocols and, more generally, the important class of special-sound protocols (and their many-round generalization). For these protocols, the extractor first generates a tree-structure of accepting transcripts which is then used to construct a witness. This tree-structure generation can be done in a weakly non-adaptive manner by repeatedly rewinding the prover with fresh and independent verifier messages to realize the tree of transcripts.

1.1.2 Statistical WI and one-way permutations

We next consider black-box constructions of *statistical* WI (SWI) arguments of knowledge, compared to the computational WI (CWI) protocols discussed above. It is known that constant-round SWI can be constructed from constant-round Statistically Hiding Commitments (SHC) or Collision Resistant Hash functions (CRH). We show that one-way permutations alone are unlikely to suffice for such constructions. For this, we let Sam_d be the collision finder oracle of Haitner, Hoch, Reingold and Segev [HHRS07], which, as shown in [HHRS07], is strong enough to break a d-round SHC, but, for any $d = o(n/\log n)$, is not enough to invert random permutations (implying that it is insufficient for breaking one-way permutations in a black-box manner). Following [PV10], we consider the class of polynomial-szie circuits with oracle access to Sam, denoted $\mathsf{P}^{\text{Sam}_d}/\text{poly}$. We show the following barrier on constructing SWI with security against *unbounded* malicious verifiers.

Theorem 1.3 (informal, Lemma 8.6)**.** *If there exists a black-box construction from a one-way permutation of an unbounded malicious verifier* SWI *argument of knowledge for* NP *with d rounds for some constant d, then* $\mathsf{NP} \subseteq \mathsf{P}^{\text{Sam}_d}/\text{poly}$.

That is, establishing the security of protocols with SWI against unbounded malicious verifiers on one-way permutations in a black-box manner would lead to a non-trivial upper-bound on NP with respect to Sam (namely, that $\mathsf{NP} \subseteq \mathsf{P}^{\text{Sam}_d}/\text{poly}$). We remark that the constructions of SWI from constant-round SHC are, in fact, secure against unbounded malicious verifiers, and that the class $\mathsf{P}^{\text{Sam}_d}/\text{poly}$ does not trivially contain NP. For example, as mentioned above, it is possible that one-way functions exist with respect to Sam_d. We point out that in the case that $\mathsf{NP} \not\subseteq \mathsf{P}^{\text{Sam}_d}/\text{poly}$, which seems to be likely, the above results implies that there are no black-box constructions of SWI arguments of knowledge with constant rounds from one-way permutations.

[3] See Definition 2.11 for a formal definition.

[4] Or extractors that can be naturally made weakly non-adaptive. For our result to hold the extractor only needs to work against provers with very high convincing probability, which is a very weak requirement.

Previously, [PV10] showed that only languages in $\mathsf{BPP}^{\mathrm{Sam}}$ have fully black-box constant-round zero-knowledge proofs or constant-round fully black-box zero-knowledge arguments with sublinear verifier communication complexity from one-way permutations.

As in previous results, when the extractor is weakly non-adaptive we can relax the requirement on the verifier and the number of rounds.

Theorem 1.4 (informal)**.** *If there exists a black-box construction from a one-way permutation of an unbounded malicious verifier* SWI *argument of knowledge for* NP *with weakly non-adaptive extractor and* d *rounds for some function* $d(n) = o(n/\log n)$*, then* $\mathsf{NP} \subseteq \mathsf{P}^{\mathrm{Sam}_d}/\mathrm{poly}$.

We note that Theorem 1.4 is tight since an $O(n/\log n)$-round SHC can be constructed from one-way permutations [NOVY98; KS06; HR07a].

1.2 Techniques

In this section we give an overview of the techniques used to prove our results. Let $\mathcal{R}_{\mathsf{SAT}}$ be the canonical relation for SAT where the witness for a formula x is its assignment. We assume the existence of a WI argument (P, V) for $\mathcal{R}_{\mathsf{SAT}}$ satisfying knowledge soundness.[5] We first explain in more detail what we mean when we refer to knowledge soundness, then discuss some simplifying assumptions used in this overview, and then describe the organization of the rest of this section.

Knowledge soundness. The protocol (P, V) satisfies knowledge soundness if there exists a (strict-polynomial time) oracle-aided algorithm E such that for any x and (non-uniform) PPT "malicious prover" $\widehat{P}$ that makes $V(x)$ accept with noticeable probability, then the extractor with black-box access to $\widehat{P}$, denoted $E^{\widehat{P}}(x)$, outputs w so that $(x, w) \in \mathcal{R}_{\mathsf{SAT}}$, with high probability.[6] Black-box access entails that the extractor can make queries of the form $Q = (\mathsf{tr}, a)$ where a is a message and tr is either empty or the result of a previous query. That is $\mathsf{tr} = (\mathsf{tr}'||a'||m')$ where (tr', a') is a previous query and m' is the corresponding answer. In response to the query, the oracle returns a prover message m sampled according to the next-message distribution of P conditioned on P's internal state after producing tr (which was previously stored by the oracle) and on receiving the verifier message a. The prover's updated internal state after generating m is then stored by the oracle for use in subsequent queries. See Section 2.4 for a more formal description.

Temporary simplifying assumptions. For the majority of the overview we will make two simplifying assumptions, which we will later remove:

- *Statistical WI:* The protocol satisfies statistical (transcript-)WI (SWI). That is, for every x, w, w' with $(x, w), (x, w') \in \mathcal{R}_{\mathsf{SAT}}$ and every non-uniform malicious verifier $\widehat{V}$:

$$\mathrm{SD}(\langle P(x, w), \widehat{V}(x)\rangle, \langle P(x, w'), \widehat{V}(x)\rangle) = \mathsf{negl}(|x|)$$

  where SD denotes statistical distance and $\langle P(x, w), \widehat{V}(x)\rangle$ denotes the transcript of a random interaction between $P(x, w)$ and $\widehat{V}$. We use CWI and SWI to refer to computational and statistical WI respectively.

- *Stateless prover:* The next-message function of the honest prover P depends only on its initial inputs x, w, the communication transcript so far tr, and fresh randomness. In particular, it does not keep a state in between rounds of the protocol. Consequently, given inputs x, w and a partial transcript tr, the prover's next message $m \leftarrow P(x, w, \mathsf{tr})$ can be efficiently sampled.

Additionally, for ease of readability, we will be imprecise as to when properties hold for all large enough input lengths compared to only holding infinitely-often.

[5] We assume that the honest prover P is efficient given an instance-witness pair. Note that, unlike zero-knowledge, witness indistinguishability is trivial for inefficient provers who can "canonicalize" the witness.

[6] Knowledge soundness is typically defined with an expected-time knowledge extractor that is always required to output a witness. In our setting (see also Footnote 11), such an extractor can always be made to run in strict polynomial time by outputting $\perp$ if it did not stop after $t = \mathrm{poly}(n)$ time steps. The strict-time extractor fails with probability $O(1/t)$, however such a success probability suffices for our results (moreover, the success probability can be amplified at the cost of a higher query complexity).

Organization. The rest of this section is organized as follows:

- In Section 1.2.1 we show that SWI cannot hold "against the extractor" for non-trivial languages.
- In Section 1.2.2 we show that for constant-round protocols, SWI must hold against the extractor.
- In Section 1.2.3 we discuss how to remove the assumptions of SWI and stateful prover, and put together all of the results towards proving Theorem 1.1 as well as discussing the proof of Theorem 1.3.
- In Section 1.2.4 we show how to extend Theorem 1.1 to the case of many rounds and honest-verifier WI assuming the extractor is weakly non-adaptive, establishing Theorem 1.2.

1.2.1 Knowledge Extractors Can Be Fooled Only For Trivial Languages

Intuitively, there is a fundamental tension between WI and knowledge soundness: while WI posits that it infeasible to detect which of a pair of valid witnesses was used by the honest prover, (black-box) knowledge soundness entails the existence of an extractor that outputs a valid witness when interacting with the same prover. Thus, if WI holds "against the extractor", then the extractor is capable of outputting an alternative witness to the one used by the prover, which is infeasible unless SAT is easily decidable. This is formalized in the following lemma:

Lemma 1.5 (informal, Lemma 3.1)**.** *If for every x, w, w' with $(x, w), (x, w') \in \mathcal{R}_{\mathsf{SAT}}$ it holds that*

$$\mathrm{SD}(E^{P(x,w)}(x), E^{P(x,w')}(x)) \leq 1/\mathrm{poly}(n),$$

then $\mathsf{SAT} \in \mathsf{P}/\mathrm{poly}$.

Proof sketch. Before describing the algorithm, we show a property of the extractor. Let S be some set of satisfiable instances, and for a satisfiable instance x, let $w(x)$ be a witness for x. For a random sample $x_0, x_1 \leftarrow S$ consider the OR statement $x' = (x_0 \vee x_1)$. Observe that for either value of b, $E^{P(x', w(x_b))}(x')$ outputs (with high probability) a witness w' for x', which equates to either a witness for x_0 or a witness for x_1. Let p be the probability (over the choice of $x_0, x_1 \leftarrow S$ and randomness for the extractor) that $(x_0, E^{P(x', w(x_0))}(x')) \in \mathcal{R}_{\mathsf{SAT}}$. Our main observation is that by indistinguishability against the extractor, $(x_0, E^{P(x', w(x_1))}(x')) \in \mathcal{R}_{\mathsf{SAT}}$ with probability $\approx p$ (where the $\approx$ hides at most an inverse-polynomial loss coming from the extractor distinguishing probability). Similarly, both runs of the extractor output a witness for x_1 with probability $\approx 1 - p$. Thus, slightly abusing notation to allow r to be the combined randomness of the extractor and the prover, the probability

$$\Pr_{\substack{x_0, x_1 \leftarrow S \\ b \leftarrow \{0,1\} \\ r}} \left[(x_b, E^{P(x_0 \vee x_1, w(x_{1-b}))}(x_0 \vee x_1; r)) \in \mathcal{R}_{\mathsf{SAT}} \right],$$

which, by symmetry, is equal to

$$\Pr_{\substack{x_0, x_1 \leftarrow S \\ b \leftarrow \{0,1\} \\ r}} \left[(x_1, E^{P(x_b \vee x_{1-b}, w(x_0))}(x_b \vee x_{1-b}; r)) \in \mathcal{R}_{\mathsf{SAT}} \right],$$

is (roughly) lower-bounded by $1/2$. As a result, the expectation

$$\mathop{\mathbf{E}}_{\substack{b \leftarrow \{0,1\} \\ x_0 \leftarrow S \\ r}} \left[\Pr_{x_1 \leftarrow S} \left[(x_1, E^{P(x_b \vee x_{1-b}, w(x_0))}(x_b \vee x_{1-b}; r)) \in \mathcal{R}_{\mathsf{SAT}} \right] \right],$$

is bounded from below by $1/2$.

In other words, for every set S there exist $b = b(S)$, $x_0 = x_0(S)$ and $r = r(S)$ so that it is the case that $(x_1, E^{P(x_b \vee x_{1-b}, w(x_0))}(x_b \vee x_{1-b})) \in \mathcal{R}_{\mathsf{SAT}}$ for $1/2$ of the values $x_1 \in S$. Let C_S be the circuit that, on input x_1, outputs $w = E^{P(x_b \vee x_{1-b}, w(x_0))}(x_b \vee x_{1-b})$ for $b = b(S)$, $x_0 = x_0(S)$ and $r = r(S)$. Observe that, $|C_S|$ has polynomial size since all it needs is $b, x_0, w(x_0), r$ and a circuit representing E. To conclude, we have that for every set S, there exists a polynomial-size circuit C_S so that $(x, C_S(x)) \in \mathcal{R}$ for roughly $1/2$ of the values $x \in S$.

We are now ready to describe the circuit for deciding SAT. Consider a series of m sets (for a polynomial m to be discussed later) $S_1, \ldots, S_m$ where S_1 is the set of all satisfiable formulas of size

n and $S_i = S_{i-1} \setminus \{x | (x, C_{S_{i-1}}(x)) \in \mathcal{R}_{\mathsf{SAT}}\}$ for $C_{S_{i-1}}$ as defined in the above discussion, and let $C_{S_1}, \ldots, C_{S_m}$ be the matching circuits. Then the circuit for deciding SAT on input a formula x outputs 1 if there exists $i \in [m]$ such that $(x, C_{S_i}(x)) \in \mathcal{R}_{\mathsf{SAT}}$ and outputs 0 otherwise.

As we set m to be polynomial, and C_{S_i} has polynomial size, the circuit is polynomial-size. We now show that the circuit decides SAT. If $x \notin \mathsf{SAT}$, then it will always output 0 as no witness for x exists. If $x \in \mathsf{SAT}$ then it will only output 0 if there is no i so that $(x, C_{S_i}(x)) \in \mathcal{R}_{\mathsf{SAT}}$. However, since $C_{S_{i-1}}$ outputs a valid witness for roughly half of the elements in S_{i-1}, we have that $|S_i| \leq |S_{i-1}|/2$. As a result, for a large enough m, we have that $S_m = \emptyset$. This implies that every $x \in S_1$ (i.e., all satisfiable formulas) have some i for which C_{S_i} outputs a valid witness for x. □

We have shown that if WI applies to the extractor, then SAT can be efficiently decided. The next section explores for which protocols the extractor can be fooled in this manner.

1.2.2 Transforming Knowledge Extractors Into Malicious Verifiers

Following Section 1.2.1, we show that for stateless-prover arguments of knowledge with constant round complexity, the distinguishing ability of the knowledge extractor can be bounded. Essentially, we do this by transforming the extractor into a malicious verifier $\widehat{V}$ whose distinguishing probability bounds that of the extractor in the sense of Lemma 1.5, and then use WI to bound the distinguishing probability of $\widehat{V}$.

This establishes the following limitation for stateless-prover SWI arguments of knowledge which may be of independent interest:

Theorem 1.6 (informal, Corollary 4.2)**.** *Assume that* $\mathsf{NP} \not\subseteq \mathsf{P/poly}$. *Then there is no constant-round* stateless-prover SWI *argument of knowledge for* $\mathcal{R}_{\mathsf{SAT}}$.

Proof sketch. Let (P, V) be a k-round SWI argument of knowledge where k is constant, δ is a bound on the statistical-distance in the witness-indistinguishability experiment (for any verifier), and the knowledge extractor E makes q queries. We show that for every x, w, w' with $(x, w), (x, w') \in \mathcal{R}_{\mathsf{SAT}}$:

$$\mathrm{SD}(E^{P(x,w)}, E^{P(x,w')}) = O(kq^{k+1}\delta).$$

This suffices for the theorem by applying Lemma 1.5 assuming $kq^{k+1}\delta = 1/\mathrm{poly}(n)$, which holds if δ is negligible, q is polynomial, and k is constant. Let $\delta' = kq^{k+1}\delta$ and suppose towards contradiction (to witness-indistinguishability) that there exist x, w, w' such that $\mathrm{SD}(E^{P(x,w)}(x), E^{P(x,w')}(x)) \gg \delta'$.

Consider E_i, the extractor that runs with $P(x, w)$ up to and including the i-th query and after this interacts with $P(x, w')$ (observe that this is well defined since the prover is stateless, and so we can run $P(x, w')$ on a transcript generated using $P(x, w)$ without requiring a state). Since E_0 and E_q are equivalent to $E^{P(x,w')}$ and $E^{P(x,w)}$ respectively, by a simple hybrid argument, there exists i^* so that $\mathrm{SD}(E_{i^*}, E_{i^*+1}) \gg \delta'/q$. We henceforth fix this i^*.

The extractor's i-th query is of the form $Q_i = (\mathsf{tr}_i, a_i)$ where a_i is a message and tr_i is a partial transcript whose prefix is either empty or has already been queried by the extractor. Thus, given a set of queries $Q_1, \ldots, Q_q$ made by the extractor, the query Q_{i^*} uniquely defines a list of query indices $j_1 < \cdots < j_t$ leading from the empty transcript up to tr_{i^*}. That is (1) $\mathsf{tr}_{j_1} = \emptyset$, (2) for every s the concatenation $\mathsf{tr}_{j_{s-1}} || a_{j_{s-1}}$ is a prefix of tr_{j_s}, and (3) $j_t = i$ (so that $\mathsf{tr}_{j_t} = i$). Observe that the length t of this list is at most k (but may be shorter) as this is a limit on the length of an interaction transcript.

We next describe the malicious verifier $\widehat{V}$ which has x, w, w', and i^* hard-coded and acts as follows:

1. Guess $t \leftarrow [k]$ and $j_1 < \cdots < j_{t-1} \leftarrow \binom{[i^*]}{t-1}$ and set $j_t = i^*$.
2. Begin running $E(x)$. On the i-th query (tr, a) compute the query answer m as follows: if $i = j_s$ for some s then forward a as a message to the actual prover and let m be its reply. Otherwise, if $i \leq i^*$ then let $m \leftarrow P(x, w, \mathsf{tr})$ and if $i > i^*$ then answer as $a \leftarrow P(x, w', \mathsf{tr})$.
3. Following t rounds of interaction, the verifier aborts and outputs its view.

By witness-indistinguishability of the protocol, we have $\mathrm{SD}(\langle P(x, w), \widehat{V}\rangle, \langle P(x, w'), \widehat{V}\rangle) < \delta$. Our goal is now to show that

$$\begin{aligned}\mathrm{SD}(\langle P(x, w), \widehat{V}\rangle, \langle P(x, w'), \widehat{V}\rangle) &> \frac{1}{kq^k} \cdot \mathrm{SD}(E_{i^*}, E_{i^*+1}) \\ &\gg \frac{\delta'}{kq^{k+1}} = \delta,\end{aligned}$$

as this will lead to a contradiction. Let G be the event that $\widehat{V}$ chooses the correct length and the correct indices on the path leading to tr_{i^*}, and observe that G holds with probability at least $1/k \cdot 1/\binom{i^*}{k-1} > 1/kq^k$. Conditioned on G occurring, E_{i^*} is essentially post-processing of $\langle P(x,w), \widehat{V}\rangle$ by, after the t-th round, continuing to run E using $P(x,w')$ to reply to queries.

At this point we want to show that, conditioned on G, E_{i^*+1} can be described as post-processing of $\langle P(x,w'), \widehat{V}\rangle$. However, this is not immediately obvious. Indeed, while in E_{i^*+1} the queries at indices $j_1, \ldots, j_{t-1}$ are answered according to the witness w (i.e., by $P(x,w)$), in the interaction $\langle P(x,w'), \widehat{V}\rangle$ (conditioned on G) this path is answered using w'. In order to combat this, we crucially recall that, by WI, the distributions $\langle P(x,w'), \widehat{V}\rangle$ and $\langle P(x,w), \widehat{V}\rangle$ are statistically close, and so we can consider the latter distribution in place of the former. We are now done, as E_{i^*+1} *can* be described as post-processing on $\langle P(x,w), \widehat{V}\rangle$ (conditioned on G) by, after the $(t-1)$-th round, continuing to run E using $P(x,w')$ to reply to queries. See Section 4.1 for details. □

1.2.3 Removing Simplifying Assumptions and Putting It All Together

All that remains for proving Theorem 1.1 is to show how to transform a constant-round computational WI argument of knowledge into a (constant-round) *stateless*-prover *statistical* WI argument of knowledge. Once we have established this, we can immediately apply Theorem 1.6 to derive Theorem 1.1.

We thus briefly discuss the CWI to SWI and stateful-to-stateless transformations:

- **Computational to statistical.** It is a well-established fact (see, e.g., [Gol90]) that if there exist efficiently sampleable distributions which are computationally indistinguishable but statistically distinguishable, then one-way functions exist. Since, given x and w the protocol becomes a sampleable distribution, under the assumption that (auxiliary-input) one-way functions do not exist, computational WI becomes statistical WI. To be slightly more precise, if there are no auxiliary-input one-way functions, then for any polynomial p, any protocol satisfying CWI with computational distance δ, satisfies SWI with statistical distance bounded by $\delta + 1/p(n)$. While this leaves us with a protocol for which SWI holds with only inverse-polynomial indistinguishability, all of the steps previously discussed regarding SWI still hold when this polynomial is chosen to be small enough.

- **Stateful to stateless.** In order to make the prover stateless we use techniques inspired by [MV24]. Roughly speaking, given the communication transcript the stateless prover will invert the original stateful prover to sample a relevant state. Assuming that auxiliary-input one-way functions do not exist this can be done efficiently. We identify two distinct transformations from the prover into an alternative stateless honest prover: (1) a transformation that preserves malicious-verifier SWI assuming the protocol is constant-round, and (2) a transformation that works for any number of rounds but only preserves honest-verifier SWI. See Section 5 for more details.

The weakly non-adaptive extractor case differs in that when we make the prover stateless we lose malicious-verifier SWI. Thus we cannot use SWI to bound the distinguishing probability of $\widehat{V}$. We instead go through the transformation described in Lemma 1.7 to show an alternate proof system where extraction still holds but the distinguishing probability of $\widehat{V}$ is small. As discussed earlier, Lemma 1.7 can be used only because in this case $\widehat{V}$ does not depend on the prover. Since we do not need (malicious) SWI now, it suffices to begin with honest-verifier CWI.

Limitations to black-box constructions from one-way permutations. We now discuss the proof of Theorem 1.3. We note that is almost implied by Theorem 1.6, except for the assumption that the prover is stateless. We derive Theorem 1.3 by showing a transformation that makes the prover (almost) stateless, using the collision finding oracle Sam_d of [HHRS07]. Recall that [HHRS07] showed that, for $d = o(n/\log n)$, Sam_d is not enough to break the one-wayness of a random permutation, but can be used to break the binding of Statistically Hiding Commitments (SHC) with d rounds.

We use Sam_d to resample the state of the prover in each round, leading to a prover P' that, in a conversation with the extractor, behaves just like a stateless prover. This almost suffices to go from Theorem 1.6 to Theorem 1.3.[7] However, recall that in the proof of Theorem 1.6, the verifier

[7] We note that in the full proof, we also need to slightly alter Lemma 1.5 to hold with a Sam oracle, but we leave these details for the main body of the paper.

$\widehat{V}$ needed to simulate the prover, which could be done (efficiently) because the prover was stateless. However, since P' is not actually stateless,[8] this cannot be done efficiently. As a result, we simply leave the verifier as inefficient. Consequently, Theorem 1.3 holds when assuming SWI against inefficient malicious verifiers. See Section 8 for details.

1.2.4 Extension to Many Rounds and Honest-Verifier WI

In this section we discuss how to extend the results discussed previously in this section to many-round protocols and honest-verifier witness indistinguishability.

Where did we get stuck? It will first be useful to discuss why our previous approach does require malicious-verifier WI and constant rounds.

- Malicious-verifier WI was used in the proof of Theorem 1.6, since we constructed from the extractor a verifier $\widehat{V}$ and used SWI to bound its distinguishing advantage. As the extractor can choose its messages arbitrarily, $\widehat{V}$ is potentially very different from the honest verifier.
- The constant-round assumption was used twice. The first is, again in the proof of Theorem 1.6, where we could bound the extractor's distinguishing advantage by $\approx kq^{k+1}$ times the advantage of $\widehat{V}$. Recall that this was the case because the verifier has to guess the query indices of the path leading to i^*. The second time that we used the constant-round assumption is when making the prover stateless while preserving SWI.

SWI against prescribed verifiers. We begin by discussing the issue of malicious-verifier WI. A natural approach to rectify this is to transform the protocol to be SWI against malicious verifiers. However, using standard approaches for this (e.g., [Vad06; BKPRV24]) generally requires one-way functions and (even in restricted cases such as public-coin protocols where one-way functions are not necessary) do not seem to preserve knowledge soundness.

Our main observation is that we do not need the full power of witness indistinguishability: we only need to bound the distinguishing advantage of the verifier $\widehat{V}$ constructed in the proof of Theorem 1.6. We prove that if there are no (auxiliary-input) one-way functions then if we start with honest-verifier SWI protocol, we can make it "witness-indistinguishable against $\widehat{V}$":

Lemma 1.7 (informal, Lemma 6.1)**.** *Let (P, V) be an honest-verifier* WI *argument for $\mathcal{R}_{\mathsf{SAT}}$ with stateless prover P and let $\widehat{V}$ be a malicious verifier. If auxiliary-input one-way functions do not exist, then there exists an efficient stateless prover $P' = P'(\widehat{V})$ with the following properties:*

- Completeness preservation: *for every $(x, w) \in \mathcal{R}_{\mathsf{SAT}}$: $P(x, w)$ and $P'(x, w)$ make $V(x)$ accept with roughly the same probability.*
- SWI against $\widehat{V}$: *for every x, w, w' with $(x, w), (x, w') \in \mathcal{R}_{\mathsf{SAT}}$ it holds that*

$$\mathrm{SD}(\langle P'(x, w'), \widehat{V}(x)\rangle, \langle P'(x, w_1), \widehat{V}(x)\rangle \leq 1/\mathrm{poly}(n).$$

A proof sketch of Lemma 1.7 follows some additional discussion (see Section 6 for the full proof).

While at first glance it seems that Lemma 1.7 immediately gets the malicious-verifier WI assumption, a subtle issue remains. The verifier $\widehat{V}$ in Section 1.2.2 needs to internally run P', but the prover P' itself depends on $\widehat{V}$. We thus have a circularity issue where both the prover and the verifier depend on each other. In the case of weakly non-adaptive extraction we can bypass this dependency cycle.

Weakly non-adaptive extraction. More precisely, we assume that the extractor is *weakly non-adaptive*, meaning that it chooses its rewinding strategy independently of the prover's replies, and its queries in each rewinding branch depend only on the partial transcript corresponding to this branch. That is, the structure of its interaction with the prover is determined prior to receiving any information from the prover (in particular, before it has any information about which witness is being used). Since the extractor's queries to the interaction path chosen by $\widehat{V}$ depend only on the prover's replies in this path, $\widehat{V}$ does not need to run the P. That is, (after a minor change to choose i^* at random) the verifier $\widehat{V}$ is independent of P and its inputs.

[8] Roughly speaking, the state of the prover contains only the transcript of the protocol so far, but it will not answer to any query which is not a continuation of this partial transcript.

Now that we have gotten rid of the requirement of malicious-verifier SWI, we can use the many-round stateful-to-stateless transformation, getting rid of one of the elements limiting us to constant rounds. Luckily, the second reason, the multiplicative overhead of $\approx kq^{k+1}$, also comes "for free" with weakly non-adaptive extractors: since the structure of the rewinding strategy is fixed prior to any interaction with the prover, the verifier $\widehat{V}$ does not need to guess it (i.e., after i^* is chosen, t and $j_1, \ldots, j_{t-1}$ are fixed and known). By using this observation we derive

$$\mathrm{SD}(E^{P(x,w)}(x), E^{P(x,w')}(x)) = O(q \cdot \mathrm{SD}(\langle P(x,w), \widehat{V}\rangle, \langle P(x,w'), \widehat{V}\rangle)),$$

which now suffices for any polynomial round complexity.

Proof sketch of Lemma 1.7. With all of this established, we go back to discuss the proof of Lemma 1.7. Suppose that we could construct a prover P' that answers only messages that were sent by the honest verifier, but aborts given any message that does not come from this distribution. This would suffice: either $\widehat{V}$ acts similarly to the honest verifier, or it sees only aborts, yielding no information about the witness. This is the basic idea that underlies our construction.[9]

The proof consists of two steps. First, we construct an *inefficient* prover P'' that preserves completeness and is SWI against $\widehat{V}$, and then we show how to emulate P'' with an efficient prover P'.

Notation. Before describing the proof in more detail, we will need some notation: $\langle P(x,w), V(x)\rangle_{<i,V}$ is the random variable representing the first $i-1$ full rounds of the transcript and ending in the i-th verifier message. For a partial transcript tr containing i full rounds, let $\mathcal{V}_i(x, \mathsf{tr})$ be the distribution of the verifier's next message conditioned on tr being the transcript so far. Let $\widehat{\mathcal{V}}_i(x, \mathsf{tr})$ be the same distribution, but for $\widehat{V}$. Finally, letting a be a verifier message, let $\mathcal{P}_i(x, w, \mathsf{tr}, a)$ be defined similarly as the next-message distribution of P conditioned on tr and verifier message a.

Inefficient prover. Before describing the prover, we define an important ratio. For an instance x, witness w and transcript $\mathsf{tr} = (a_1, m_1, \ldots, a_{i-1}, m_{i-1}, a_i)$ let[10]

$$\rho_{x,w}(\mathsf{tr}) = \frac{\mathbf{Pr}\left[\langle P(x,w), \widehat{V}(x)\rangle_{<i,V} = \mathsf{tr}\right]}{\mathbf{Pr}\left[\langle P(x,w), V(x)\rangle_{<i,V} = \mathsf{tr}\right]},$$

be the multiplicative difference between the likelihood of the transcript tr appearing in an interaction between P and $\widehat{V}$ compared to a conversation with V. Intuitively, what we want is for the prover to abort if the ratio is too high (i.e., tr is much more likely in a conversation with $\widehat{V}$ than with the honest verifier). Thus, a verifier that does not follow the honest-verifier strategy will receive only aborts.

Let $Q \in \mathrm{poly}(n)$ be a parameter. The inefficient prover $P''(x,w)$ with interaction transcript tr prior to the prover's move acts as follows: If $\rho_{x,w}(\mathsf{tr}) > Q$ then abort. Otherwise sample and output message $m \leftarrow P(x, w, \mathsf{tr})$. (Note that since P is stateless, so is P''.)

We show that P'' preserves completeness of the original protocol and is WI against $\widehat{V}$:

- *Completeness preservation.* The only difference between P and P'' is on transcripts for which P'' aborts. Thus, to show that completeness is preserved, it suffices to show that P'' will abort with small probability when interacting with V. Let S be the set of all transcripts on which P'' aborts (and for which P'' does not abort on any of their prefixes). Since the prover only aborts when $\rho_{x,w}(\mathsf{tr}) > Q$, the set S is characterized precisely by all transcripts tr for which, letting $i(\mathsf{tr})$ be the length of tr, it holds that:

  $$Q < \rho_{x,w}(\mathsf{tr}) = \frac{\mathbf{Pr}\left[\langle P(x,w), \widehat{V}(x)\rangle_{<i(\mathsf{tr}),V} = \mathsf{tr}\right]}{\mathbf{Pr}\left[\langle P(x,w), V(x)\rangle_{<i(\mathsf{tr}),V} = \mathsf{tr}\right]}$$

  Thus, the probability that a transcript in S is sampled (and hence that the prover aborts) in interaction with V is at most:

  $$\sum_{\mathsf{tr}\in S} \mathbf{Pr}\left[\langle P(x,w), V(x)\rangle_{<i(\mathsf{tr}),V} = \mathsf{tr}\right] < \sum_{\mathsf{tr}\in S} \frac{1}{Q} \cdot \mathbf{Pr}\left[\langle P(x,w), \widehat{V}(x)\rangle_{<i(\mathsf{tr}),V} = \mathsf{tr}\right] \leq \frac{1}{Q}.$$

[9] A similar high-level idea, but with different implementation details and context (e.g., dealing with interaction), was recently used by [CHK25] in the setting of non-interactive zero-knowledge.

[10] For simplicity, in this overview we ignore the case that $\mathbf{Pr}\left[\langle P(x,w), V(x)\rangle_{<i,V} = \mathsf{tr}\right] = 0$.

Thus, P'' causes V to accept with the same probability as the honest prover, up to an error of $1/Q = 1/\mathrm{poly}(n)$.

- WI *against* $\widehat{V}$. While it is not hard to show that whenever $\widehat{V}$ leads the prover to a transcript that is unlikely in an interaction with the honest verifier, P'' will abort with high probability, this alone does not suffice to show WI against $\widehat{V}$. The reason is that $\rho_{x,w}(\mathsf{tr})$ depends on the witness used by the prover, and so it may abort on different transcripts in either case, which can be used to distinguish which witness was used. Luckily, we show that this is not an issue by making the crucial observation that $\rho_{x,w}(\mathsf{tr})$ *is independent of the prover* and, as a result, of the witness:

$$\begin{aligned}\rho_{x,w}(\mathsf{tr}) &= \frac{\mathbf{Pr}\left[\langle P(x,w), \widehat{V}(x)\rangle_{<i,V} = \mathsf{tr}\right]}{\mathbf{Pr}\left[\langle P(x,w), V(x)\rangle_{<i,V} = \mathsf{tr}\right]} \\ &= \frac{\mathbf{Pr}\left[\widehat{\mathcal{V}}_i(x, \mathsf{tr}_{<i}) = a_i\right]}{\mathbf{Pr}\left[\mathcal{V}_i(x, \mathsf{tr}_{<i}) = a_i\right]} \cdot \prod_{j<i} \frac{\mathbf{Pr}\left[\widehat{\mathcal{V}}_j(x, \mathsf{tr}_{<j}) = a_j\right] \mathbf{Pr}\left[\mathcal{P}_j(x, w, \mathsf{tr}_{<j}, a_j) = m_j\right]}{\mathbf{Pr}\left[\mathcal{V}_j(x, \mathsf{tr}_{<j}) = a_j\right] \mathbf{Pr}\left[\mathcal{P}_j(x, w, \mathsf{tr}_{<j}, a_j) = m_j\right]} \\ &= \prod_{j\le i} \frac{\mathbf{Pr}\left[\widehat{\mathcal{V}}_j(x, \mathsf{tr}_{<j}) = a_j\right]}{\mathbf{Pr}\left[\mathcal{V}_j(x, \mathsf{tr}_{<j}) = a_j\right]},\end{aligned}$$

  where $\mathsf{tr}_{<j}$ denotes the first j full rounds contained in tr. The implication of this is that given partial transcript tr, $P''(x, w_0)$ and $P''(x, w_1)$ abort with the same probability. Thus, since $\widehat{V}$ must roughly follow the honest verifier's messages, honest-verifier SWI implies SWI against $\widehat{V}$.

Making the prover efficient. The prover P'' is inefficient only in its computation of $\rho_{x,w}$. The main observation towards emulating P'' by an efficient prover P' is that, assuming that there are no auxiliary-input one-way functions, the ratio $\rho_{x,w}(\mathsf{tr})$ can be efficiently approximated. Thus, the efficient P' will be nearly identical to P'' except that it uses this approximated value. Indeed, consider the function $f_{x,w,i}(b, r_P, r_V) = (|T|, T, b)$ where $|T|$ is the bit length of T and:

- If $b = 0$ then $\mathsf{tr} \leftarrow \langle P(x,w), V(x)\rangle_{<i,V}$.
- If $b = 1$ then $\mathsf{tr} \leftarrow \langle P(x,w), \widehat{V}(x)\rangle_{<i,V}$.

Observe that f is efficiently computable given x, w, and i. By standard techniques, assuming there are no auxiliary-input one-way functions, there exists a predictor $\mathsf{Pred}_{x,w,i}$ so that with high probability $\mathsf{Pred}_{x,w,i}(|\mathsf{tr}|, \mathsf{tr}) \in (1 \pm \epsilon) \cdot \mathbf{Pr}\left[b = 1 |\ f_{x,w,i} \text{ outputs } \mathsf{tr}\right]$. Observe that

$$\begin{aligned}\mathbf{Pr}\left[b = 1 |\ f_{x,w,i} \text{ outputs } \mathsf{tr}\right] &= \frac{\mathbf{Pr}\left[f_{x,w,i} \text{ outputs } \mathsf{tr} | b = 1\right] \mathbf{Pr}\left[b = 1\right]}{\mathbf{Pr}\left[f_{x,w,i} \text{ outputs } \mathsf{tr}\right]} \\ &= \frac{\mathbf{Pr}\left[\mathsf{tr} \leftarrow \langle P(x,w), \widehat{V}(x)\rangle_{<i,V}\right]}{2\,\mathbf{Pr}\left[f_{x,w,i} \text{ outputs } \mathsf{tr}\right]}\end{aligned}$$

Similarly,

$$1 - \mathbf{Pr}\left[b = 1 |\ f_{x,w,i} \text{ outputs } \mathsf{tr}\right] = \mathbf{Pr}\left[b = 0 |\ f_{x,w,i} \text{ outputs } \mathsf{tr}\right] = \frac{\mathbf{Pr}\left[\mathsf{tr} \leftarrow \langle P(x,w), V(x)\rangle_{<i,V}\right]}{2\,\mathbf{Pr}\left[f_{x,w,i} \text{ outputs } \mathsf{tr}\right]}$$

and so

$$\frac{\mathbf{Pr}\left[b = 1 |\ f_{x,w,i} \text{ outputs } \mathsf{tr}\right]}{1 - \mathbf{Pr}\left[b = 1 |\ f_{x,w,i} \text{ outputs } \mathsf{tr}\right]} = \frac{\mathbf{Pr}\left[\langle P(x,w), \widehat{V}(x)\rangle_{<i,V} = \mathsf{tr}\right]}{\mathbf{Pr}\left[\langle P(x,w), V(x)\rangle_{<i,V} = \mathsf{tr}\right]} = \rho_{x,w}(\mathsf{tr}).$$

Thus, $\rho_{x,w}(\mathsf{tr})$ can be approximated by running $p \leftarrow \mathsf{Pred}_{x,w,i}(\mathsf{tr})$ and computing $\widetilde{\rho}_{x,w}(\mathsf{tr}) = \frac{p}{1-p}$. We remark that in order to take into account that $\widetilde{\rho}_{x,w}(\mathsf{tr})$ may not be precisely equal to $\rho_{x,w}(\mathsf{tr})$, we need to slightly alter the construction of P'', introducing a small amount of noise in its decision whether to abort. Finally, note that computing $\widetilde{\rho}_{x,w}$ does not depend on any previous state, and so by computing it rather than $\rho_{x,w}$ we retain statelessness.

1.3 Additional Related Works

As mentioned in Section 1, [OW93; HN24] prove that one-way functions are necessary for zero-knowledge arguments for NP, and [LMP24] extended this result to the uniform settings. More recently, [CHK25] prove that weak Non-Interactive Zero-Knowledge (NIZK) proofs imply one-way functions in some parameter regimes, and used it for NIZK amplification. Very recently, [CLV26; CHKT26] improve the above result, showing that any non-trivial constant-round zero-knowledge protocol implies one-way functions. [MV24] study *instance-hiding* proofs, and show that the existence of such a constant-round proof for a hard language implies non-uniform one-way functions. If the proof has many rounds but it is additionally simulatable, [MV24] showed it implies one-way functions.

[MPS25] defined and study witness pseudo-canonicalization (WPC), a notion closely related to Non-Interactive WI proofs, and showed that assuming one-way functions, Levin NP-complete languages do not admit WPC. WPC for a Levin NP-complete language can be seen as a NIWI with a weak form of knowledge extractor.

Lastly, the recent work of Bitansky et al. [BKPRV24] gives another motivation to study WI. It shows that BARGs and SNARGs imply a weak form of WI. It is plausible that their results may be strengthened to preserve knowledge soundness, and to achieve stronger forms of WI. In particular, proving that BARGs/SNARGs for NP with knowledge soundness imply WI of knowledge with good enough parameters, combined with our theorem, would mean that BARGs/SNARGs imply auxiliary-input one-way functions.

2 Preliminaries

General notation. We use the following conventions throughout the paper:

- All logarithms are taken in base 2.
- Calligraphic letters denote sets and distributions. Uppercase letters denote random variables; lowercase letters denote values and functions.
- poly denotes the set of polynomially bounded functions. PPT denotes probabilistic polynomial-time algorithms. A non-uniform polynomial-time algorithm A is equipped with a fixed polynomial-size advice sequence $\{z_n\}_{n\in\mathbb{N}}$ (typically omitted from notation), and we write A_n for A with advice z_n on inputs of length n.
- For a randomized algorithm A, we write $A(\cdot; r)$ for A with randomness fixed to $r \in \{0,1\}^*$. We write negl for a negligible function.
- For $n \in \mathbb{N}$, we let $[n] \coloneqq \{1, \ldots, n\}$. For $i \leq j$, we let $[i : j] \coloneqq \{i, \ldots, j\}$.
- For a vector $v \in \Sigma^n$, we write v_i for its i-th entry, $v_{<i} = (v_1, \ldots, v_{i-1})$, and $v_{\leq i} = (v_1, \ldots, v_i)$. For $\mathcal{I} \subseteq [n]$, we write $v_{\mathcal{I}}$ for the ordered tuple $(v_i)_{i\in\mathcal{I}}$.
- For strings $x, y \in \{0,1\}^*$, we write $x||y$ for their concatenation.

Probability notation. When unambiguous, we will naturally identify a random variable with its distribution. The support of a finite distribution $\mathcal{P}$ is defined by $\mathrm{Supp}(\mathcal{P}) \coloneqq \{x\colon \mathbf{Pr}_{\mathcal{P}}[x] > 0\}$. For a (discrete) distribution $\mathcal{P}$, let $x \leftarrow \mathcal{P}$ denote that x was sampled according to $\mathcal{P}$. Similarly, for a set $\mathcal{S}$, let $x \leftarrow \mathcal{S}$ denote that x is drawn uniformly from $\mathcal{S}$. For $m \in \mathbb{N}$, we use $\mathbf{U}_m$ to denote a uniform random variable over $\{0,1\}^m$ (that is independent from other random variables in consideration). A collection of distributions $\{\mathcal{D}_a\}_{a\in\{0,1\}^{\lambda(n)},n\in\mathbb{N}}$ is samplable if there exists a poly(n)-time randomized algorithm S such that for every $n \in \mathbb{N}$ and every $a \in \{0,1\}^{\lambda(n)}$ $S(a) \equiv \mathcal{D}_a$.

Statistical distance. The statistical distance of two distributions $\mathcal{P}$ and $\mathcal{Q}$ over a discrete domain $\mathcal{X}$ is defined by $\mathrm{SD}(\mathcal{P}, \mathcal{Q}) \coloneqq \max_{\mathcal{S}\subseteq\mathcal{X}} |\mathcal{P}(\mathcal{S}) - \mathcal{Q}(\mathcal{S})| = \frac{1}{2}\sum_{x\in\mathcal{X}} |\mathcal{P}(x) - \mathcal{Q}(x)|$.

2.1 NP Relations

For a language $\mathcal{L} \subseteq \{0,1\}^*$, let $\mathcal{L}_n = \mathcal{L} \cap \{0,1\}^n$, and for $\mathcal{L} \in \mathsf{NP}$ let $\mathcal{R}(\mathcal{L}) \subseteq \mathcal{L} \times \{0,1\}^*$ be a canonical NP relation. Similarly, for an NP relation $\mathcal{R} \subseteq \{0,1\}^* \times \{0,1\}^*$, we let $\mathcal{L}(\mathcal{R}) = \{x|\ \exists w : (x,w) \in \mathcal{R}\}$. For a string x, we let $\mathcal{L}(x)$ be the indicator function that outputs 1 if $x \in \mathcal{L}$ and 0 otherwise. The "or-relation" associated with $\mathcal{R}$, is defined as $\mathcal{R} \vee \mathcal{R} \coloneqq \{((x_0, x_1), w) \mid (x_0, w) \in \mathcal{R} \vee (x_1, w) \in \mathcal{R}\}$ (we sometimes denote instances (x_0, x_1) for $\mathcal{R} \vee \mathcal{R}$ as $\vee$ statements, e.g., $x_0 \vee x_1$).

The SAT language and relation. We define the canonical relation for the NP complete language SAT to be $\mathcal{R}_{\mathsf{SAT}} = \{(\varphi, v) \colon \varphi \text{ is a Boolean formula and } v \text{ is an assignment such that } \varphi[v] = 1\}$. Note that $\mathcal{R}_{\mathsf{SAT}} \vee \mathcal{R}_{\mathsf{SAT}}$ is equivalent to $\mathcal{R}_{\mathsf{SAT}}$.

2.2 Auxiliary-Input One-Way Functions

Definition 2.1. *An efficiently computable function family* $\mathbf{F} = \{f_a \colon \{0,1\}^{m(n)} \to \{0,1\}^{m(n)}\}_{a \in \{0,1\}^{\lambda(n)}}$ *is an* auxiliary-input one-way function (AI-OWF) *if for every PPT* $\mathcal{A}$ *there exists a negligible function* μ *such that for every* $n \in \mathbb{N}$*, there exists some* $a \in \{0,1\}^{\lambda(n)}$ *such that*

$$\Pr_{x \leftarrow \{0,1\}^{m(n)}} \left[\mathcal{A}(a, f_a(x)) \in f_a^{-1}(f_a(x))\right] \leq \mu(n) \tag{1}$$

If a *is* $0^{\lambda(n)}$*, we refer to the family as simply a* one-way function (OWF)*. We say that* $\mathbf{F}$ *is an* infinitely often auxiliary-input one-way function (ioAI-OWF) *if for every* $\mathcal{A}$ *the above holds for infinitely many* $n \in \mathbb{N}$*. We sometimes use membership notation: e.g.,* $\mathbf{F} \in$ ioAI-OWF *denotes that* $\mathbf{F}$ *is an* ioAI-OWF*.*

We now define the notion of universal extrapolation:

Theorem 2.2 (Universal Extrapolation, [IL89])**.** *Let* $m, \lambda \in$ poly*, and let* $\mathcal{D} = \{\mathcal{D}_a\}_{a \in \{0,1\}^{\lambda(n)}, n \in \mathbb{N}}$ *be a collection of samplable distributions over* $\{0,1\}^{m(n)}$*. Then for every* $\epsilon \in 1/\text{poly}$ *there exists an efficiently computable function family* $\mathbf{F} = \{f_a\}_a$*, such that if* $\mathbf{F}$ *is not an* ioAI-OWF*, then there exists an efficiently computable algorithm* Samp *such that the following holds for any* $n \in \mathbb{N}$*, any* $a \in \{0,1\}^{\lambda(n)}$*, and any* $i \in [m(n)]$*: Let* X *be a random variable taking values from* $\mathcal{D}_a$*. Then*

$$\mathrm{SD}((X_{\leq i}, \mathsf{Samp}(a, X_{\leq i})), X) \leq \epsilon(n).$$

Moreover, the proof is black-box: For every $p \in$ poly *there exists an oracle-aided algorithm* Samp*, such that for every* A *and any large enough* n*, the following holds. If* A *inverts* f_a *for* $a \in \{0,1\}^{\lambda(n)}$ *with probability* $1/p(n)$*, then* $\mathrm{SD}((X_{\leq i}, \mathsf{Samp}^A(a, X_{\leq i})), X) \leq \epsilon(n)$*.*

Almost directly from Theorem 2.2, we get that, assuming no OWFs, any samplable distribution has a next-bit predictor.

Definition 2.3 (Next-Bit Predictor)**.** *For an distribution* $Z \in \{0,1\}^m$*, we say that a function* $\mathsf{Pred} \colon \{0,1\}^* \times \{0,1\}^{\ell_{\mathsf{Pred}}} \to \mathbb{R}$ *is a* next-bit ε-predictor *for* Z*, if for every* $i \in [m]$

$$\Pr_{\substack{z_{<i} \leftarrow Z_{<i} \\ r_{\mathsf{Pred}} \leftarrow \{0,1\}^{\ell_{\mathsf{Pred}}}}} [\mathsf{Pred}(z_{<i}; r_{\mathsf{Pred}}) \in (p(z_{<i}) \cdot (1-\varepsilon), p(z_{<i}) \cdot (1+\varepsilon))] \geq 1 - \varepsilon$$

for $p(z_{<i}) = \mathbf{Pr}\left[Z_i = 1 |_{Z_{<i} = z_{<i}}\right]$*.*

For a collection of distributions $Z = \{Z_a\}_{a \in \{0,1\}^{\lambda(n)}, n \in \mathbb{N}}$*, we say that* Pred *is a next-bit* ε*-predictor for the collection* Z *if there exists a* $q \in$ poly *such that for every sequence* $\{a_n \in \{0,1\}^{\lambda(n)}\}_n$ *and every sufficiently large* $n \in \mathbb{N}$*, the function* $\mathsf{Pred}_{a_n}(\cdot) \coloneqq \mathsf{Pred}(a_n, \cdot)$ *is a next-bit* $\varepsilon(n)$*-predictor for* Z_{a_n} *and the runtime of* Pred_{a_n} *is at most* $q(n)$*.*

Theorem 2.4 (AI-OWF and next-bit predictor, [IL89])**.** *Let* $\mathcal{D} = \{\mathcal{D}_a\}_{a \in \{0,1\}^{\lambda(n)}, n \in \mathbb{N}}$ *be a collection of samplable distributions. Then for every* $\epsilon \in 1/\text{poly}$ *there exists a function family* $\mathbf{F} = \{f_a\}_a$*, such that if* $\mathbf{F}$ *is not a* ioAI-OWF*, then there exists an* ϵ*-next bit predictor for* $\mathcal{D}$*.*

Moreover, the proof is black-box: For every $p \in$ poly *there exists an oracle-aided algorithm* Pred*, such that for every PPT* A *and any large enough* n*, the following holds. If* A *inverts* f_a *for* $a \in \{0,1\}^{\lambda(n)}$ *with probability* $1/p(n)$*, then* Pred^A *is an* ϵ*-next-bit predictor for* $\mathcal{D}_a$*.*

Finally, we will make use of the following generalization of [Gol90] for AI-OWFs.

Theorem 2.5 ([Gol90])**.** *Let* $\mathcal{D} = \{\mathcal{D}_a\}_{a \in \{0,1\}^{\lambda(n)}}$ *be a collection of efficiently samplable distribution over* $\{0,1\}^{m(n)}$*. Then for every* $\epsilon \in 1/\text{poly}$ *there exists a function family* $\mathbf{F} = \{f_{a,a'}\}_{a,a'}$*, such that if* $\mathbf{F}$ *is not a* ioAI-OWF*, then the following holds.*

Assume that there are two ensembles $\{a_n \in \{0,1\}^{\lambda(n)}\}_{n \in \mathbb{N}}$*,* $\{a'_n \in \{0,1\}^{\lambda(n)}\}_{n \in \mathbb{N}}$ *such that* $\{\mathcal{D}_{a_n}\}_{n \in \mathbb{N}}$ *and* $\{\mathcal{D}_{a'_n}\}_{n \in \mathbb{N}}$ *are* δ*-indistinguishable. Then for every* $p \in$ poly*,*

$$\mathrm{SD}(\mathcal{D}_{a_n}, \mathcal{D}_{a'_n}) \leq \delta(n) + \epsilon(n).$$

Moreover, the proof is black-box: For every $p \in \text{poly}$ there exists an oracle-aided algorithm D, such that for every PPT A, and any large enough n, the following holds. If A inverts $f_{a,a'}$ for $a, a' \in \{0,1\}^{\lambda(n)}$ with probability $1/p(n)$, then D^A distinguishes $\mathcal{D}_a$ and $\mathcal{D}_{a'}$ with advantage $\epsilon + \text{SD}(\mathcal{D}_a, \mathcal{D}_{a'})$.

2.3 Interactive Arguments

Interactive protocols. Let $P = (P_1, \ldots, P_k)$ and $V = (V_1, \ldots, V_k, V_{dec})$ be algorithms (the prover and verifier, respectively). We assume $k = k(|x|) \in \text{poly}(|x|)$ where x is the input to V.

For inputs x, y and k-length vectors of random coins r_P, r_V, we denote by

$$\mathsf{tr} \leftarrow \langle P(y; r_P), V(x; r_V) \rangle$$

the transcript of their interaction. When r_P, r_V are omitted, they are assumed to be sampled uniformly and independently. The transcript has the form $\mathsf{tr} = (a_1, m_1, \ldots, a_k, m_k)$, where for $i \in [k]$:

$$(a_i, \mathsf{st}_{V,i}) = V_i(x, \mathsf{tr}_{<i}, \mathsf{st}_{V,i-1}; r_{V,i}),$$
$$(m_i, \mathsf{st}_{P,i}) = P_i(y, \mathsf{tr}_{<i}, a_i, \mathsf{st}_{P,i-1}; r_{P,i}),$$

where $\mathsf{tr}_{<i} = (a_1, m_1, \ldots, a_{i-1}, m_{i-1})$ and with $\mathsf{st}_{V,i}$, $\mathsf{st}_{P,i}$ being the states of V and P, respectively, at the i-th round, where $\mathsf{st}_{V,0} = \mathsf{st}_{P,0} = \emptyset$ and $m_0 = \emptyset$. We sometimes also use the notation: $\mathsf{tr}_{\leq i} = (a_1, m_1, \ldots, a_i, m_i)$, and sometimes omit the subscript i from V_i and P_i when the round index is clear from context. We additionally use the notation $\mathsf{tr}_{<i,V}$ to mean the concatenation of $\mathsf{tr}_{<i}$ with a_i. We use similar notations with $\langle P(y; r_P), V(x; r_V) \rangle$ to denote the distribution of partial transcripts. We define the verifier's output distribution as

$$\mathsf{Out}(P(y), V(x)) \coloneqq V_{dec}(x, \mathsf{tr}, \mathsf{st}_{V,k}),$$

where tr, is the transcript following the entire interaction, and $\mathsf{st}_{V,k}$ is the verifier's last outputted state as in the process described above.

We say that the protocol is public-coin if for every i, the verifier directly sends its random coins, so that $a_i = r_{V,i}$. Without loss of generality if the protocol is public-coin we can assume that $\mathsf{st}_{V,i} = \emptyset$ as the state can be recovered from the transcript.

Definition 2.6 (Interactive Argument)**.** *An* interactive argument *(IA) for a relation $\mathcal{R} \subseteq \{0,1\}^* \times \{0,1\}^*$ is a pair of PPT interactive machines (P, V) such that:*

- ***Completeness:*** *For every $(x, w) \in \mathcal{R}$,*

$$\mathbf{Pr}\left[\mathsf{Out}(P(x, w), V(x)) = 1\right] \geq 1 - \alpha(|x|).$$

- ***Computational Soundness:*** *For every $x \notin L_{\mathcal{R}}$ and every non-uniform PPT prover $\widehat{P}$,*

$$\mathbf{Pr}\left[\mathsf{Out}(\widehat{P}, V(x)) = 1\right] \leq \beta(|x|).$$

Whenever α and β are not explicitly specified, they are assumed to be negligible functions in $|x|$. Note that we will not actually use soundness in this paper.

2.4 Knowledge Soundness and Weakly Non-Adaptive Extractors

We begin by defining what it means for an algorithm to have black-box access to an interactive algorithm:

Definition 2.7. *Let P be a k-round interactive algorithm. An algorithm E has* black-box access *to P, denoted E^P, if E can be described as interacting with the oracle defined as follows:*

- *The oracle maintains a list of interaction pairs $(\mathsf{tr}, \mathsf{st})$ where tr is a transcript ending in a P-message (empty transcripts are considered as such), and st is a state of P corresponding to the transcript tr.*
- *E queries with a transcript/message pair (tr, a).*

- *On receiving a query* (tr, a)*, the oracle checks whether* tr *belongs to a pair in the maintained list. If it is not, it returns* $\perp$*. If it does, let* st *be the accompanying state. Sample* $(m, \mathsf{st}') \leftarrow P(\mathsf{tr}, a, \mathsf{st})$ *(using random coins if required). Add the tuple* $(\mathsf{tr}||a||m, \mathsf{st}')$ *to the list. Return* m *as the oracle answer to* E*.*

Notice that the algorithm E *may rewind by querying from any previously recorded transcript prefix, but it has no access to the internal state of* P *beyond what is revealed through its messages.*

We can now define knowledge soundness of a protocol:

Definition 2.8. *An interactive argument satisfies (black-box strict-time)* (η, μ)-computational knowledge soundness *if there exists a polynomial-time oracle machine* E *such that for every* x *and every non-uniform PPT prover* $\widehat{P}$*:*

$$\mathbf{Pr}\left[\mathsf{Out}(\widehat{P}, V(x)) = 1\right] > \eta(|x|) \quad \Longrightarrow \quad \mathbf{Pr}\left[(x, E^{\widehat{P}}(x)) \in \mathcal{R}\right] \geq 1 - \mu(|x|).$$

The extractor runs in strict polynomial time (i.e., not expected polynomial time).

If η *and* μ *are not specified, then it is assumed that* $\eta = 1/2$ *and* $\mu = \mathsf{negl}(|x|)$*. In this case, we may simply call the protocol an "argument of knowledge".*[11]

In this paper we always assume that if the protocol has completeness error α, then there exists a polynomial p such that $1 - \alpha - \eta > 1/p$. That is, there is a noticeable gap between the honest-prover success probability and the bound where the extractor begins acting.

Remark 2.9. *Knowledge soundness is sometimes defined where the extractor runs on* expected *time and its success probability is related to that of the prover, e.g.*

$$\mathbf{Pr}\left[(x, E^{\widehat{P}}(x)) \in \mathcal{R}\right] > \mathbf{Pr}\left[\mathsf{Out}(\widehat{P}, V(x)) = 1\right] - \epsilon(|x|).$$

We note that for every fixed polynomial p *this notion implies ours with* $\eta = 1/p$ *and* $\mu = \mathrm{poly}(1/p)$ *by considering only provers with* $\mathbf{Pr}\left[\mathsf{Out}(\widehat{P}, V(x)) = 1\right] > 1/p$ *and getting rid of the expected time by running the extractor for sufficient (polynomial) time before aborting (and using Markov's inequality to bound the success probability).*

Remark 2.10 (Amplification of extraction probability)**.** *Observe that membership in the relation is efficiently testable. Therefore, by repeating the extractor* t *times, it is possible to improve the probability that* $(x, E^{\widehat{P}}(x)) \in \mathcal{R}$ *from* $1 - \mu$ *to* $1 - \mu^t$*. The extractor's query complexity goes from* q *to* tq*.*

This preserves whether the extractor is weakly non-adaptive (see Definition 2.11 below).

We now define a specific class of extractors for knowledge soundness:

Definition 2.11 (Weakly non-adaptive extractor)**.** *A* q*-query knowledge extractor* E *is* weakly non-adaptive *if the following holds: On input* x*, the extractor samples randomness* r *and determines (based only on* x *and* r*) a sequence of oracle-query specifications* $(\ell_1, V^{(1)}), \ldots, (\ell_q, V^{(q)})$*, where:*

- *each* ℓ_j *specifies which previously obtained transcript prefix to query, and*
- *each* $V^{(j)}$ *is a polynomial-size circuit.*[12]

During execution, the extractor keeps a list of all transcripts previously obtained (including the empty transcript) and, in round j*, submits as its query, the* ℓ_j*-th transcript prefix* tr_{ℓ_j} *along with* $a = V^{(j)}(\mathsf{tr})$ *(without modifying the* ℓ_j *or* $V^{(j)}$ *based on the prover's responses).*

In particular, note that the sequence $(\ell_1, V^{(1)}), \ldots, (\ell_q, V^{(q)})$ *is "committed to" (i.e., fixed) before interacting with the prover.*

Remark 2.12 (Relation to other notions of extraction)**.** *We clarify the differences between fully non-adaptive extractors, (tree) special soundness, and weakly non-adaptive extractors.*

[11] We note that our notion only makes sense when η is a noticeable function as otherwise E is unlikely to see an accepting transcript when running in polynomial time. Thus, our "default" setting is $\eta = 1/2$ but could be inverse-polynomial for any polynomial. See also [BL04].

[12] Note that since extractor is probabilistic, it can sample and plug in randomness for the circuit in advance and so we can consider the circuits as having random inputs.

- *A* (fully) non-adaptive extractor *fixes its entire sequence of oracle queries in advance as a function only of the instance and its internal randomness. In particular, neither the number, structure, nor content of oracle queries depends on the prover's responses.*
- Special soundness *is a structural property of the protocol: from a fixed combinatorial set of accepting transcripts (e.g., for 3-round protocols, transcripts sharing a prefix and differing in verifier challenges, and more generally, a tree) one can efficiently compute a witness. The corresponding extractor rewinds the prover according to a strategy of verifier challenges until this tree structure is produced, and then uses special-soundness to extract a witness. Commonly, and especially for public-coin protocols, the extractor's challenges can be described as independent of the prover's replies in rewinding branches.*
- *A* weakly non-adaptive extractor *fixes in advance its entire rewinding schedule (i.e., the sequence of rewind indices) and the associated verifier-response functions, based only on the instance and internal randomness. While this schedule does not depend on the prover's behavior, the actual verifier messages are obtained by applying the fixed response functions to the dynamically evolving transcript, and therefore depend on the prover's messages.*

Thus, non-adaptive extractors and the extractors induced by special soundness when the extractor's rewinding schedule is fixed are special cases of weakly non-adaptive extractors.

2.5 Witness Indistinguishability

We define witness indistinguishability. Specifically, we define WI applied a weaker notion than standard, where indistinguishability applies to the transcript distributions rather than the verifier view. Using this weaker definition only strengthens our results (see Remark 2.14).

Definition 2.13. *An interactive argument (P, V) for a relation $\mathcal{R}$ satisfies* computational witness (transcript-)indistinguishability *with distinguishing advantage δ (δ-*CWI*) if for every x and every w_0, w_1 with $(x, w_0), (x, w_1) \in \mathcal{R}$, and every non-uniform PPT verifier $\widehat{V}$, and every non-uniform PPT distinguisher D:*

$$\left|\mathbf{Pr}\left[D\left(\langle P(x, w_0), \widehat{V}(x)\rangle\right) = 1\right] - \mathbf{Pr}\left[D\left(\langle P(x, w_1), \widehat{V}(x)\rangle\right) = 1\right]\right| < \delta(|x|).$$

We further define:

- *A protocol satisfies* statistical witness indistinguishability *δ-*SWI *as above, except it must hold against any (possibly inefficient) distinguisher (equivalently, the two distributions above have statistical distance at most δ).*
- *A protocol is δ-honest-*CWI *(resp. δ-honest-*SWI*) if indistinguishability holds when $\widehat{V}$ is the honest verifier V (so that δ-*CWI *implies δ-honest-*CWI*, but the opposite is not necessarily true).*
- *When δ is a negligible function, we sometimes omit δ from all of the above notation (e.g., δ-*CWI *becomes* CWI*).*

Remark 2.14. *Witness indistinguishability is more commonly defined with respect to the* view *of the verifier, as opposed to the transcript. The transcript-indistinguishability variant is a strictly weaker (i.e., more easily achievable) notion as any distinguisher can always restrict itself to looking only at the transcript. Our results show that this weaker notion is already sufficient to imply cryptography.*

We additionally note that for SWI *the two notions are equivalent: since the verifier's randomness is independent of the witness used by the prover, its view can be post-sampled given the transcript.*

We will make use of the following fact on WI when there are no one-way functions.

Lemma 2.15. *Let (P, V) be a δ-honest-*CWI *interactive argument for a relation $\mathcal{R}$.*

- *For any $\epsilon \in 1/\text{poly}$, there exists a function family $\mathbf{F}$ such that if $\mathbf{F}$ is not an* ioAI-OWF*, then (P, V) is $(\delta + \epsilon)$–honest-*SWI*.*
- *If there is no* ioAI-OWF *and (P, V) is a δ-*CWI*, then (P, V) is $(\delta + \epsilon)$-*SWI *for any $\epsilon \in 1/\text{poly}$.*

Moreover, the proof is black-box: For every efficient (potentially malicious and non-uniform) verifier $\widehat{V}_a$ and any $\epsilon > 1/\text{poly}$, there exists a function family $\mathbf{F} = \{f_{a,x,w_0,w_1}\}_{a,x,w_0,w_1}$ such that the following holds. For any $p \in \text{poly}$ there exists an oracle-aided algorithm D, such that for every PPT A and any large enough n,if A inverts f_{a,x,w_0,w_1} for $x \in \{0,1\}^n$ with probability $1/p(n)$, then D^A distinguishes $\langle P(x,w_0), \widehat{V}_a(x)\rangle$ and $\langle P(x,w_1), \widehat{V}_a(x)\rangle$ with advantage $\epsilon + \text{SD}(\langle P(x,w_0), \widehat{V}_a(x)\rangle, \langle P(x,w_1), \widehat{V}_a(x)\rangle)$.

Proof. For an efficient, non-uniform verifier $\widehat{V}$ with non-uniform advice a, consider the distribution collection $\mathcal{D} = \{\mathcal{D}_{x,w,a}\}_{(x,w)\in\mathcal{R},a}$, where $\mathcal{D}_{x,w,a} = \langle P(x,w), \widehat{V}_a(x)\rangle$. Fix some ensemble $\{(x_n, w_n^0, w_n^1)\}_{n\in\mathbb{N}}$ with $(x_n, w_n^b) \in \mathcal{R}_n$ for $b \in \{0,1\}$. By the assumed computational WI it holds that $\{\mathcal{D}_{x_n,w_n^0,a}\}_{n\in\mathbb{N}}$ is $\delta(n)$-indistinguishable from $\{\mathcal{D}_{x_n,w_n^1,a}\}_{n\in\mathbb{N}}$. By Theorem 2.5 it follows that

$$\text{SD}(\mathcal{D}_{x_n,w_n^0,a}, \mathcal{D}_{x_n,w_n^1,a}) \leq \delta + 1/p(n)$$

for every large enough $n \in \mathbb{N}$, which implies that the protocol is statistical-WI.

The moreover part follows by construction and by Theorem 2.5. □

2.6 Useful Lemmas

In this section we prove some useful lemmas. The first two lemmas dealing with statistical distance conditioned on some prefix.

Lemma 2.16. *Let (X_0, Y_0) and (X_1, Y_1) be two pairs of jointly distributed random variables with $\text{SD}((X_0,Y_0),(X_1,Y_1)) \leq \epsilon$. Then*

$$\Pr_{x\leftarrow X_0}\left[\text{SD}(Y_0|_{X_0=x}, Y_1|_{X_1=x}) \geq \alpha\right] \leq 2\epsilon/\alpha$$

(Also by symmetry, $\Pr_{y\leftarrow Y_0}\left[\text{SD}(X_0|_{Y_0=y}, X_1|_{Y_1=y}) \geq \alpha\right] \leq 2\epsilon/\alpha$.)

Proof. Let $\widehat{Y}_1$ be a random variable jointly distributed with X_0, defined by $\widehat{Y}_1|_{X_0=x} \equiv Y_1|_{X_1=x}$. First note that $\text{SD}(X_0, X_1) \leq \epsilon$, and thus by data-processing also $\text{SD}((X_0, \widehat{Y}_1), (X_1, Y_1)) \leq \epsilon$. By the triangle inequality we get that

$$\text{SD}((X_0,Y_0),(X_0,\widehat{Y}_1)) \leq \text{SD}((X_0,\widehat{Y}_1),(X_1,Y_1)) + \text{SD}((X_0Y_0),(X_1Y_1)) \leq 2\epsilon,$$

and we are interested in bounding

$$\Pr_{x\leftarrow X_0}\left[\text{SD}(Y_0|_{X_0=x}, \widehat{Y}_1|_{X_0=x}) \geq \alpha\right].$$

Observe that

$$\mathop{\mathbf{E}}_{x\leftarrow X_0}\left[\text{SD}(Y_0|_{X_0=x}, \widehat{Y}_1|_{X_0=x})\right] = \text{SD}((X_0,Y_0),(X_0,\widehat{Y}_1)) \leq 2\epsilon.$$

Thus, by Markov's inequality,

$$\Pr_{x\leftarrow X_0}\left[\text{SD}(Y_0|_{X_0=x}, \widehat{Y}_1|_{X_0=x}) \geq \alpha\right] \leq 2\epsilon/\alpha.$$

□

Lemma 2.17 (Bad prefixes)**.** *Let A, B be random variables over Σ^n such that $\text{SD}(A,B) \leq \epsilon$ and for $\alpha \in (0,1]$, define*

$$\text{BAD} = \{x \in \Sigma^i : i \in [n], \text{SD}(A|_{A_{<i}=x}, B|_{B_{<i}=x}) \geq \alpha\}.$$

Then, for every $i \in [n]$:

$$\Pr_{x\leftarrow A}\left[x_{<i} \in \text{BAD}\right] \leq 2\epsilon/\alpha.$$

(By symmetry it also holds that $\Pr_{x\leftarrow B}\left[x_{<i} \in \text{BAD}\right] \leq 2\epsilon/\alpha$.)

Proof. Fix α, and $i \in [n]$, and define $\mathrm{BAD}^i = \{x \in \Sigma^i \colon \mathrm{SD}(A|_{A_{<i}=x}, B|_{B_{<i}=x}) \geq \alpha\}$. By applying Lemma 2.16 on the pairs $(X_0, Y_0) = (A_{<i}, A_{\geq i})$ and $(X_1, Y_1) = (B_{<i}, B_{\geq i})$, we get that

$$\Pr_{x_{<i} \leftarrow A_{<i}} [\mathrm{SD}(A_{\geq i}|_{A_{<i}=x_{<i}}, B_{\geq i}|_{B_{<i}=x_{<i}}) \geq \alpha] = \Pr_{x \leftarrow X_0} [\mathrm{SD}(Y_0|_{X_0=x}, Y_1|_{X_1=x}) \geq \alpha] \leq 2\epsilon/\alpha.$$

Finally, since $\{x_{<i}\} \cap \mathrm{BAD} = \{x_{<i}\} \cap \mathrm{BAD}^i$ and $A = (A_{<i}, A_{\geq i})$ we have that

$$\Pr_{x \leftarrow A} [x_{<i} \in \mathrm{BAD}] = \Pr_{x \leftarrow A} \left[x_{<i} \in \mathrm{BAD}^i\right] = \Pr_{x_{<i} \leftarrow A_{<i}} [\mathrm{SD}(A_{\geq i}|_{A_{<i}=x_{<i}}, B_{\geq i}|_{B_{<i}=x_{<i}}) \geq \alpha] \leq 2\epsilon/\alpha$$

□

The last lemma shows how to bound the statistical distance with a bound on the probability of some bad event.

3 Fooling Extractors Implies Trivial Languages

In this section, we show that an interactive argument of knowledge has "witness indistinguishability against the extractor", then the language is trivial. This formalizes the intuitive tension between witness indistinguishability (hiding the witness) and knowledge soundness (extractability of the witness).

Lemma 3.1. *Let $\mathcal{R}$ be an NP relation and (P, V) be an interactive argument for $\mathcal{R} \vee \mathcal{R}$ with completeness error α and (η, μ) computational knowledge soundness with extractor E. If $1 - \alpha > \eta$ and there exist δ and $\varepsilon \in 1/\mathrm{poly}$ so that $2\mu + \delta \leq 1 - \varepsilon$ and for every x_0, x_1, w_0, w_1 with $(x_0, w_0) \in \mathcal{R}$ and $(x_1, w_1) \in \mathcal{R}$ it holds that*

$$SD\left(E^{P(x_0 \vee x_1, w_0)}(x_0 \vee x_1), E^{P(x_0 \vee x_1, w_1)}(x_0 \vee x_1)\right) \leq \delta,$$

then $\mathcal{L}(\mathcal{R}) \in \mathsf{P}/\mathrm{poly}$.

Proof. We begin by proving that since the extractor is unable to distinguish which witness the prover uses it will output a witness for the second statement with high probability.

Claim 3.2. *For every x_0, x_1, w_0, w_1 with $(x_0, w_0) \in \mathcal{R}$ and $(x_1, w_1) \in \mathcal{R}$,*

$$\Pr_{b \leftarrow \{0,1\}} \left[\left(x_{1-b}, E^{P(x_0 \vee x_1, w_b)}(x_0 \vee x_1)\right) \in \mathcal{R}\right] \geq \frac{1}{2} - \frac{2\mu + \delta}{2}.$$

Proof. Let $x = (x_0 \vee x_1)$. Suppose towards contradiction (of indistinguishability of the extractor distributions) that

$$\Pr_{b \leftarrow \{0,1\}} \left[\left(x_{1-b}, E^{P(x, w_b)}(x)\right) \in \mathcal{R}\right] < \frac{1}{2} - \frac{2\mu + \delta}{2}$$

By completeness of the interactive argument, for every b:

$$\mathbf{Pr}\left[\mathsf{Out}(P(x, w_b), V(x)) = 1\right] \geq 1 - \alpha > \eta,$$

where the final inequality is by definition, that knowledge soundness applying to the honest prover (Definition 2.8). Thus, by knowledge soundness, we have

$$\Pr_{\substack{b \leftarrow \{0,1\} \\ w \leftarrow E^{P(x_0 \vee x_1, w_b)}(x_0 \vee x_1)}} [(x_b, w) \in \mathcal{R} \vee (x_{1-b}, w) \in \mathcal{R}] = \Pr_{b \leftarrow \{0,1\}} \left[\left(x, E^{P(x, w_b)}(x)\right) \in \mathcal{R} \vee \mathcal{R}\right] \geq 1 - \mu$$

Inspecting this probability:

$$\begin{aligned} &\Pr_{\substack{b \leftarrow \{0,1\} \\ w \leftarrow E^{P(x_0 \vee x_1, w_b)}(x_0 \vee x_1)}} [(x_b, w) \in \mathcal{R} \vee (x_{1-b}, w) \in \mathcal{R}] \\ = &\Pr_{\substack{b \leftarrow \{0,1\} \\ w \leftarrow E^{P(x_0 \vee x_1, w_b)}(x_0 \vee x_1)}} [(x_{1-b}, w) \in \mathcal{R}] + \Pr_{\substack{b \leftarrow \{0,1\} \\ w \leftarrow E^{P(x_0 \vee x_1, w_b)}(x_0 \vee x_1)}} [(x_b, w) \in \mathcal{R} \wedge (x_{1-b}, w) \notin \mathcal{R}] \end{aligned}$$

Putting this together:

$$\begin{aligned}
&\mathop{\mathbf{Pr}}_{\substack{b\leftarrow\{0,1\}\\ w\leftarrow E^{P(x,w_b)}(x)}}[(x_b,w)\in\mathcal{R}\ \wedge\ (x_{1-b},w)\notin\mathcal{R}]\\
&=\mathop{\mathbf{Pr}}_{\substack{b\leftarrow\{0,1\}\\ w\leftarrow E^{P(x,w_b)}(x)}}[((x,w)\in\mathcal{R}\vee\mathcal{R})]-\mathop{\mathbf{Pr}}_{\substack{b\leftarrow\{0,1\}\\ w\leftarrow E^{P(x,w_b)}(x)}}[(x_{1-b},w)\in\mathcal{R}]\\
&>(1-\mu)-\left(\frac{1}{2}-\frac{2\mu+\delta}{2}\right)=\frac{1}{2}+\frac{\delta}{2}
\end{aligned}$$

We now build a distinguisher D between the distributions $E^{P(x,w_0)}(x)$ and $E^{P(x,w_1)}(x)$:

1. Receive as input witness w.
2. If for a bit b' it holds that $(x_{b'},w)\in\mathcal{R}$ and $(x_{1-b'},w)\notin\mathcal{R}$ then output b'. Otherwise output a uniformly random bit.

By definition,

$$\mathop{\mathbf{Pr}}_{\substack{b\leftarrow\{0,1\}\\ w\leftarrow E^{P(x,w_b)}(x)}}[D(w)=b]\geq\mathop{\mathbf{Pr}}_{\substack{b\leftarrow\{0,1\}\\ w\leftarrow E^{P(x,w_b)}(x)}}[(x_b,w)\in\mathcal{R}\ \wedge\ (x_{1-b},w)\notin\mathcal{R}]>1/2+\delta/2$$

Finally,

$$\begin{aligned}
\mathop{\mathbf{Pr}}_{\substack{b\leftarrow\{0,1\}\\ w\leftarrow E^{P(x,w_b)}(x)}}[D(w)=b]&=\tfrac{1}{2}\mathop{\mathbf{Pr}}_{w\leftarrow E^{P(x,w_0)}(x)}[D(w)=0]+\tfrac{1}{2}\mathop{\mathbf{Pr}}_{w\leftarrow E^{P(x,w_1)}(x)}[D(w)=1]\\
&=\tfrac{1}{2}+\tfrac{1}{2}\left(\mathop{\mathbf{Pr}}_{w\leftarrow E^{P(x,w_0)}(x)}[D(w)=1]-\mathop{\mathbf{Pr}}_{w\leftarrow E^{P(x,w_1)}(x)}[D(w)=1]\right).
\end{aligned}$$

Thus,

$$\left|\mathop{\mathbf{Pr}}_{w\leftarrow E^{P(x,w_0)}(x)}[D(w)=1]-\mathop{\mathbf{Pr}}_{w\leftarrow E^{P(x,w_1)}(x)}[D(w)=1]\right|>\delta,$$

which contradicts the assumption that the extractor outputs have statistical distance at most δ. □

We are now ready to prove the lemma. For any instance $x\in\mathcal{L}_\mathcal{R}$, define $w(x)$ to be an arbitrarily chosen witness such that $(x,w(x))\in\mathcal{R}$. By Claim 3.2, we have that for every $x_0,x_1\in\mathcal{L}_\mathcal{R}$:

$$\mathop{\mathbf{Pr}}_{b\leftarrow\{0,1\}}\left[(x_{1-b},E^{P(x_0\vee x_1,w(x_b))}(x_0\vee x_1))\in\mathcal{R}\right]\geq 1/2-\gamma,$$

for $\gamma=\frac{2\mu+\delta}{2}$. By symmetry, it holds that for every $S\subseteq\mathcal{L}_n$

$$\mathop{\mathbf{Pr}}_{\substack{x_0,x_1\leftarrow S\\ b\leftarrow\{0,1\}}}\left[(x_1,E^{P(x_b\vee x_{1-b},w(x_0))}(x_b\vee x_{1-b}))\in\mathcal{R}\right]\geq 1/2-\gamma,$$

or equivalently,

$$\mathop{\mathbf{E}}_{b\leftarrow\{0,1\}}\left[\mathop{\mathbf{Pr}}_{x_0,x_1\leftarrow S}\left[(x_1,E^{P(x_b\vee x_{1-b},w(x_0))}(x_b\vee x_{1-b}))\in\mathcal{R}\right]\right]\geq 1/2-\gamma.$$

Let b^* be the value of b that maximizes the above expectation. Then

$$\mathop{\mathbf{Pr}}_{x_0,x_1\leftarrow S}\left[(x_1,E^{P(x_{b^*}\vee x_{1-b^*},w(x_0))}(x_{b^*}\vee x_{1-b^*}))\in\mathcal{R}\right]\geq 1/2-\gamma.$$

Again, this is equivalent to

$$\mathop{\mathbf{E}}_{\substack{x_0\leftarrow S\\ r}}\left[\mathop{\mathbf{Pr}}_{x_1\leftarrow S}\left[(x_1,E_r^{P(x_{b^*}\vee x_{1-b^*},w(x_0))}(x_{b^*}\vee x_{1-b^*}))\in\mathcal{R}\right]\right]\geq 1/2-\gamma.$$

Above, r is the combined randomness for the extractor and all relevant executions of P, and we let $E_r^{P(x_{b^*} \vee x_{1-b^*}, w(x_0))}$ be the extractor when r is plugged in as the randomness of the extractor and the randomnesses of the prover as appropriate.

Thus, for every set $S \subseteq \mathcal{L}_n$, there is a choice of x_0 and randomness r for the extractor, such that

$$A_{x_0, w(x_0), r, b^*}(x_1) := E_r^{P(x_{b^*} \vee x_{1-b^*}, w(x_0))}(x_{b^*} \vee x_{1-b^*})$$

finds a witness for a $(1/2 - \gamma)$-fraction of the set S. We next find $q = O(n/(1/2 - \gamma)) \in \text{poly}$ tuples $(x^1, w^1, r^1, b^1), \ldots, (x^q, w^q, r^q, b^q)$, such that for every $x \in \mathcal{L}_n$, it holds that for some $i \in [q]$, A_{x^i, w^i, r^i, b^i} finds a witness for x.

We do this iteratively: we start by choosing $S_1 = \mathcal{L}_n$, and taking (x^1, w^1, r^1, b^1) to be the choice promised above so that A_{x^1, w^1, r^1, b^1} finds a witness for at least $(1/2 - \gamma)$-fraction of S_1. Next, for each i, let S_{i+1} be the set of size at most $(1/2 + \gamma)|S_i|$ of inputs x such that $(x, A_{x^i, w^i, r^i, b^i}(x)) \notin \mathcal{R}$. We next let $(x^{i+1}, w^{i+1}, r^{i+1}, b^{i+1})$ to be the choice promised above such that $A_{x^{i+1}, w^{i+1}, r^{i+1}, b^{i+1}}$ finds a witness for at least $(1/2 - \gamma)$-fraction of S_{i+1}. Since

$$|S_i| \leq (1/2 + \gamma)^i |S_0| \leq (1/2 + \gamma)^i \cdot 2^n = (1/2 + \gamma)^i |S_0| \leq (1 - (1/2 - \gamma))^i \cdot 2^n,$$

we get that for $q = O(n/(1/2 - \gamma)) \in \text{poly}$ (recall that by assumption $1 - 2\gamma$ is inverse-polynomial) this is polynomial by assumption), it must be that the set S_q is empty and the process halts.

Finally, the circuit that decides $\mathcal{L}(\mathcal{R})$ is the circuit that, on input x, simulates the executions $A_{x^1, w^1, r^1, b^1}, \ldots, A_{x^q, w^q, r^q, b^q}$ and for every execution checks whether A returns a witness attesting to $x \in \mathcal{L}(\mathcal{R})$ (if it does, then the circuit outputs 1 and otherwise it outputs 0). Observe that this can be done by hard-coding $\{(x^i, w^i, r^i, b^i) \colon i \in [q(|x|)]\}$ into the circuit. □

4 Limitations for Stateless-Prover SWI Arguments of Knowledge

A structural property which will be of special importance to this paper is *statelessness*. Intuitively, an interactive algorithm is stateless if its next-message function depends only on its initial input, all communication so far and fresh randomness. In particular, it does not keep a state in between rounds of the protocol.

Definition 4.1. *We say that an interactive algorithm $A = (A_1, \ldots, A_k)$ is* stateless *if for every input x, round index $i \in [k]$, partial transcript $\mathsf{tr} = (a_1, m_1, \ldots, a_{i-1}, m_{i-1})$, and randomness r, letting*

$$(a_i, \mathsf{st}_i) = A(x, \mathsf{tr}, \mathsf{st}_{i-1}; r),$$

it holds that $\mathsf{st}_i = \bot$.

Consequently, when an interactive algorithm is stateless, we sometimes omit the state st_{i-1} from its input.

In this section we prove that for certain knowledge-sound SWI protocols with a stateless prover, the extractor cannot determine which witness was used by the honest prover. The section is organized into two sections, each proving such a lemma for a separate case:

- In Section 4.1 we prove Lemma 4.3, which shows indistinguishability for the extractor when the round-complexity of the protocol is constant.
- In Section 4.2 we prove Lemma 4.5, which shows that the extractor can be fooled so long as it is weakly non-adaptive (regardless of the prover).

While these results are primarily a preliminary step towards our main theorems for CWI, as a corollary, when combined with Lemma 3.1 we derive limitations for non-trivial stateless-prover SWI. Recall that if not stated otherwise, SWI arguments of knowledge have negligible completeness, knowledge soundness and WI errors:

Corollary 4.2. *Let $\mathcal{R}$ be an* NP *relation with $\mathcal{L}(\mathcal{R}) \notin \text{P/poly}$. Then there is no stateless-prover* SWI *argument of knowledge for the OR relation $\mathcal{R} \vee \mathcal{R}$ such that one of the following holds:*

- *the protocol is constant-round, or*
- *the knowledge extractor is weakly non-adaptive.*

The proof of Corollary 4.2 is given in Section 4.3.

4.1 Constant Rounds

Lemma 4.3. *Let (P, V) be a k-round interactive argument with a stateless prover for an* NP *relation $\mathcal{R}$ with δ-*SWI*, and let E be a q-query extractor. Then for every x, w_0, w_1 with $(x, w_0) \in \mathcal{R}$ and $(x, w_1) \in \mathcal{R}$ it holds that*

$$\mathrm{SD}\left(E^{P(x,w_0)}(x), E^{P(x,w_1)}(x)\right) \leq 2kq^{k+1}\delta.$$

Proof. Denote $\delta' := 2kq^{k+1} \cdot \delta$. Assume towards contradiction (to δ-SWI of the protocol) that for some x, w_0, w_1, the extractor E can distinguish between $P(x, w_0)$ and $P(x, w_1)$. That is

$$\mathrm{SD}\left(E^{P(x,w_0)}(x), E^{P(x,w_1)}(x)\right) \geq \delta'.$$

Recall that E can rewind and use P adaptively. We construct a malicious verifier $\widehat{V}$ so that the interaction between $\widehat{V}$ and $P(x, w_0)$ is distinguishable from its interaction with $P(x, w_1)$, contradicting the δ-SWI of the protocol.

We start with a simple hybrid argument. For $i \in [q+1]$ and $b \in \{0, 1\}$ let $E_i(w_b)$ be the following algorithm (recalling that P is stateless):

1. Emulate $E(x)$ answering its j-th query, which has the form (tr, a), with prover message m sampled follows:
   (a) If $j < i$: $m \leftarrow P(x, w_0, \mathsf{tr}, a)$.
   (b) If $j = i$: $m \leftarrow P(x, w_b, \mathsf{tr}, a)$.
   (c) If $j > i$: $m \leftarrow P(x, w_1, \mathsf{tr}, a)$.
2. Output whatever $E(x)$ outputs.

Observe that $E_1(w_1) \equiv E^{P(x,w_1)}(x)$, that $E_i(w_0) \equiv E_{i+1}(w_1)$, and that $E_{q+1}(w_1) \equiv E^{P(x,w_0)}(x)$. Thus, we have

$$\begin{aligned}\sum_{i=1}^{q} \mathrm{SD}\left(E_i(w_0), E_i(w_1)\right) &= \sum_{i=1}^{q} \mathrm{SD}\left(E_{i+1}(w_1), E_i(w_1)\right) \\ &\geq \mathrm{SD}\left(E_{q+1}(w_1), E_1(w_1)\right) \\ &= \mathrm{SD}\left(E^{P(x,w_0)}(x), E^{P(x,w_1)}(x)\right) > \delta'.\end{aligned}$$

By an averaging argument, there exists i with:

$$\mathrm{SD}\left(E_i(w_0), E_i(w_1)\right) \geq \delta'/q.$$

We henceforth fix i such that the above holds. For $t \in [k]$, let $E_{i,t}$ be E_i, but where E_i outputs $\perp$ if the transcript tr given during the i-th query is not a $(t-1)$-round partial transcript (i.e., tr is not of the form $(a_1, m_1, \ldots, a_{t-1}, m_{t-1})$). By a further averaging argument, there exists t such that E_i chooses this transcript length with probability at least $1/k$. That is, there exists $t \leq \min\{i, k\}$ such that:

$$\mathrm{SD}\left(E_{i,t}(w_0), E_{i,t}(w_1)\right) > \delta'/kq,$$

Fix t such that the above holds. We next design a malicious verifier $\widehat{V}$ which receives as advice w_0, w_1, i, and t and input x, and interacts with a prover $\bar{P}$:

1. Sample $(j_1 < j_2 < \cdots < j_{t-1}) \leftarrow \binom{[i-1]}{t-1}$. Set $j_t = i$.
2. Initialize an indicator bit $\mathsf{Ind} = 0$, counter $c = 1$ and initial transcript $\mathsf{tr}_0 = \emptyset$.
3. Emulate $E(x)$, answering its ℓ-th query, which has the form (tr, a) with prover message m as follows:

(a) If $\ell = j_c$ then:

i. If $\mathsf{tr}_{c-1} \neq \mathsf{tr}$ set $\mathsf{Ind} = 1$ and abort.

ii. Send a to $\bar{P}$. Receive back m in return.

iii. Set $\mathsf{tr}_c = (\mathsf{tr}||a||m)$ and update $c = c + 1$.

(b) Otherwise if $\ell < i$: sample $m \leftarrow P(x, w_0, \mathsf{tr}, a)$.

(c) If $\ell > i$, sample $m \leftarrow P(x, w_1, \mathsf{tr}, a)$.

4. Output whatever $E(x)$ outputs.

That is, $\widehat{V}$ simulates E while guessing the indices of the calls E made to P that lead to the partial transcript which is asked on the i-th query. In these indices, $\widehat{V}$ sends each query to the real prover. The answer to other queries is simulated by $\widehat{V}$. Note that $\widehat{V}$ only interacts with the prover for t rounds.

Observe that since P is efficient and stateless, $\widehat{V}$ is efficient. By the assumed δ-SWI of (P, V) we get that

$$\mathrm{SD}\left(\langle P(x, w_0), \widehat{V}(x)\rangle, \langle P(x, w_1), \widehat{V}(x)\rangle\right) < \delta. \tag{2}$$

Let $P_t(w_b)$ be the prover that runs as $P(x, w_0)$ for the first $t-1$ rounds (regardless of its input), and then, in the t-th round, given transcript tr and verifier message a answers with $m \leftarrow P(x, w_b, \mathsf{tr}, a)$. Following round t it outputs $\perp$. By the triangle inequality for statistical distance,

$$\begin{aligned}&\mathrm{SD}\left(\langle P_t(w_1), \widehat{V}(x)\rangle, \langle P(x, w_1), \widehat{V}(x)\rangle\right)\\ &\quad\geq \mathrm{SD}\left(\langle P(x, w_0), \widehat{V}(x)\rangle, \langle P_t(w_1), \widehat{V}(x)\rangle\right) - \mathrm{SD}\left(\langle P(x, w_0), \widehat{V}(x)\rangle, \langle P(x, w_1), \widehat{V}(x)\rangle\right)\end{aligned} \tag{3}$$

The following claim shows that $\widehat{V}$ can distinguish between an interaction between $P(x, w_0)$ and $P_t(w_1)$:

Claim 4.4. $\mathrm{SD}\left(\langle P(x, w_0), \widehat{V}(x)\rangle, \langle P_t(w_1), \widehat{V}(x)\rangle\right) > \frac{\delta'}{kq^{k+1}}$.

Proof. We begin by noting that, since $\widehat{V}$ does not interact with the prover beyond round t, it holds that

$$\mathrm{SD}\left(\langle P(x, w_0), \widehat{V}(x)\rangle, \langle P_t(w_0), \widehat{V}(x)\rangle\right) = 0.$$

Thus, it suffices to bound $\mathrm{SD}\left(\langle P_t(w_0), \widehat{V}(x)\rangle, \langle P_t(w_1), \widehat{V}(x)\rangle\right)$. Consider the indicator bit Ind. It switches from 0 to 1 only if $\mathsf{tr}_{c-1} \neq \mathsf{tr}$ for some c. Recall that by definition of black-box access to an interactive algorithm, that every query (tr, a) made by E has tr that is either empty or its prefix when removing the final prover message has been previously queried. Thus, if $j_1, \ldots, j_{t-1}$ is chosen such that this is the query-path of transcripts leading to the i-th query, Ind will remain 0. It follows that Ind remains 0 with probability at least $1/\binom{i-1}{t-1} > 1/\binom{q}{k} > 1/q^k$(regardless of w_b used by P_t). Finally, notice that the value of Ind at the end of the protocol is independent of the witness w_b used by P_t. Conditioned on Ind remaining 0, the output distribution of $\widehat{V}$ in $\langle P_t(w_b), \widehat{V}(x)\rangle$ is identical to the output distribution of $E_{i,t}(w_b)$. To conclude,

$$\begin{aligned}\mathrm{SD}\left(\langle P_t(w_0), \widehat{V}(x)\rangle, \langle P_t(w_1), \widehat{V}(x)\rangle\right) &\geq \frac{1}{q^k} \cdot \mathrm{SD}\left(\mathsf{Out}(P_t(w_0), \widehat{V}(x))|_{\mathsf{Ind}=0}, \mathsf{Out}(P_t(w_1), \widehat{V}(x))|_{\mathsf{Ind}=0}\right)\\ &= \frac{1}{q^k} \cdot \mathrm{SD}\left(E_{i,t}(w_0), E_{i,t}(w_1)\right)\\ &\geq \frac{\delta'}{kq^{k+1}},\end{aligned}$$

where the first inequality holds by data processing, since the randomness (and thus the output) of $\widehat{V}$ can be sampled according to the right distribution given the transcript, independently of the witness. □

By plugging Equation (2) and Claim 4.4 into Equation (3) we derive

$$\mathrm{SD}\left(\langle P_t(w_1), \widehat{V}(x)\rangle, \langle P(x, w_1), \widehat{V}(x)\rangle\right) \geq \frac{\delta'}{kq^{k+1}} - \delta = \delta.$$

Finally, observe that the prover message at the last round in both $\langle P_t(w_1), \widehat{V}(x)\rangle$ and $\langle P(x,w_1), \widehat{V}(x)\rangle$ is sampled according to $P(x,w_1)$. Moreover, all other prover messages in $\langle P_t(w_1), \widehat{V}(x)\rangle$ are sampled according to $P(x,w_0)$. Thus, by data processing (using the post-processing that resample the last message according to $P(x,w_1)$) it holds that

$$\mathrm{SD}\left(\langle P(x,w_0), \widehat{V}(x)\rangle, \langle P(x,w_1), \widehat{V}(x)\rangle\right) \geq \mathrm{SD}\left(\langle P_t(w_1), \widehat{V}(x)\rangle, \langle P(x,w_1), \widehat{V}(x)\rangle\right) \geq \delta,$$

which contradicts Equation (2) (i.e., contradicts δ-SWI). □

4.2 Weakly Non-Adaptive Extractor

Lemma 4.5. *Let E be a weakly non-adaptive extractor that makes q queries. There exists $\widehat{V}$ that runs in time polynomial in the running time of E such that for every stateless prover P' and every x, w_0,w_1 with $(x,w_0),(x,w_1) \in \mathcal{R}$:*

$$\mathrm{SD}(E^{P'(x,w_0)}(x), E^{P'(x,w_1)}(x)) \leq 2q \cdot \mathrm{SD}(\langle P'(x,w_0), \widehat{V}(x)\rangle, \langle P'(x,w_1), \widehat{V}(x)\rangle).$$

The proof of Lemma 4.5 is similar to the proof of Lemma 4.3, where the main difference is that the verifier we construct from the extractor does not need to guess which queries of the extractors to send to the prover, or to simulate the prover on the other queries of the extractor. We give a full proof below.

Proof. We assume without loss of generality that the list output by the extractor always contains q queries and is sorted: $E(x)$ always outputs a list $(\ell_1, V^{(1)}), \ldots, (\ell_q, V^{(q)})$ with $\ell_j \leq \ell_{j+1}$ for every j.

Fix x, w_0, w_1 as in the lemma statement.

We design a malicious verifier $\widehat{V}$ which receives input x, and interacts with a prover $\bar{P}$:

1. Run $E(x)$ until it outputs the oracle-query specifications $(\ell_1, V^{(1)}), \ldots, (\ell_q, V^{(q)})$.
2. Choose $i^* \leftarrow [q]$ uniformly at random.
3. Let $j_1 < \cdots < j_t$ be such that: $\ell_1 = 0$, $j_{a-1} = \ell_{j_a}$, and $j_t = i^*$. That is, $j_1, \ldots, j_t$ are the query indices of the path leading from the empty transcript up to the transcript sampled at the i^*-th query. Such a path always exists by definition of a weakly non-adaptive extractor.
4. In round i of the interaction, let $\mathsf{tr}_{<i}$ be the transcript so far. If $i > t$ then send $\perp$ and abort. Otherwise, send $a_i = V^{(j_i)}(\mathsf{tr}_{<i})$.

That is, $\widehat{V}$ simulates E, forwarding to the real prover the messages corresponding to the partial transcript asked in the i^*-th query. Note that $\widehat{V}$ receives no advice, does not need to simulate the prover, and it interacts only over t rounds.

We additionally consider an inefficient verifier $\widehat{V}'$ that acts identically to $\widehat{V}$, but additionally simulates answers for each query $(\ell_i, V^{(i)})$ made by $E(x)$ that was not set in the interaction with the prover, where it uses $P'(x,w_0)$ on every $i < i^*$ and $P'(x,w_1)$ for every $i > i^*$ (note that i^* is always answered by the real prover). It finally outputs E's output.

Lastly, for any $t \in [k]$, let $P'_t(w_b)$ be the prover that runs as $P'(x,w_0)$ for the first $t-1$ rounds (regardless of its input), and then, in the t-th round, given transcript tr and verifier message a answers with $m \leftarrow P'(x, w_b, \mathsf{tr}, a)$. Following round t it outputs $\perp$.

In the following, we let T (T' resp.) be the random variable taking the value of t in a random execution of $\widehat{V}$ ($\widehat{V}'$ resp). We use $\langle P'_T(w_b), \widehat{V}(x)\rangle$ to denote the (transcript of the) interaction between $\widehat{V}$ and $P'_T(w_b)$, where T takes the value of t chosen by $\widehat{V}$ during the execution. Note the value of T is determined by the internal randomness of $\widehat{V}$ and so P'_T is not formally a "prover". Similarly, we use $\langle P'_{T'}(w_b), \widehat{V}'(x)\rangle$ for the same experiment with respect to $\widehat{V}'$.

By the triangle inequality for statistical distance,

$$\begin{aligned}&\mathrm{SD}\left(\langle P'_T(w_1), \widehat{V}(x)\rangle, \langle P'(x,w_1), \widehat{V}(x)\rangle\right)\\ &\quad\geq \mathrm{SD}\left(\langle P'(x,w_0), \widehat{V}(x)\rangle, \langle P'_T(w_1), \widehat{V}(x)\rangle)\right) - \mathrm{SD}\left(\langle P'(x,w_0), \widehat{V}(x)\rangle, \langle P'(x,w_1), \widehat{V}(x)\rangle\right) \qquad (4)\end{aligned}$$

The following claim bounds the extractor's distinguishing probability with that of the first expression on the second line of Equation (4):

Claim 4.6. $\mathrm{SD}\left(\langle P'(x,w_0),\widehat{V}(x)\rangle,\langle P'_T(w_1),\widehat{V}(x)\rangle\right) \geq 1/q \cdot \mathrm{SD}(E^{P'(x,w_0)}(x),E^{P'(x,w_1)}(x))$.

Proof. Since $\widehat{V}$ does not interact with the prover beyond round t, it holds that $\langle P'(x,w_0),\widehat{V}(x)\rangle \equiv \langle P'_T(w_0),\widehat{V}(x)\rangle)$, and that,

$$\langle P'_T(w_b),\widehat{V}(x)\rangle \equiv \langle P'_{T'}(w_b),\widehat{V}'(x)\rangle$$

Thus, it suffices to bound $\mathrm{SD}\left(\langle P'_{T'}(w_0),\widehat{V}'(x)\rangle,\langle P'_{T'}(w_1),\widehat{V}'(x)\rangle\right)$. Notice that $P'_{T'}$ and $\widehat{V}'$ are independent given the value of T, which can be computed from the length of the transcript (as the verifier aborts after t rounds of interaction). Thus, the views of $P'_{T'}$ and $\widehat{V}'$ are independent given the transcript. It follows that, given the transcript, it is possible to sample the view of $\widehat{V}'$ according to the right distribution, without knowing the input w_b of $P'_{T'}$. By data-processing we get that

$$\mathrm{SD}\left(\langle P'_{T'}(w_0),\widehat{V}'(x)\rangle,\langle P'_{T'}(w_1),\widehat{V}'(x)\rangle\right) \geq \mathrm{SD}\left(\mathsf{Out}(P'_{T'}(w_0),\widehat{V}'(x)),\mathsf{Out}(P'_{T'}(w_1),\widehat{V}'(x))\right). \tag{5}$$

We analyze the right-hand side of the above expression by inspecting the extractor. For $i \in [q+1]$ and $b \in \{0,1\}$ let $E_i(w_b)$ be the following algorithm (recalling that P is stateless):

1. Emulate $E(x)$ answering its j-th query, which has the form (tr,a), with prover message m sampled follows:
   (a) If $j<i$: $m \leftarrow P'(x,w_0,\mathsf{tr},a)$.
   (b) If $j=i$: $m \leftarrow P'(x,w_b,\mathsf{tr},a)$.
   (c) If $j>i$: $m \leftarrow P'(x,w_1,\mathsf{tr},a)$.
2. Output whatever $E(x)$ outputs.

Observe that $E_1(w_1) \equiv E^{P'(x,w_1)}(x)$, that $E_i(w_0) \equiv E_{i+1}(w_1)$, and that $E_{q+1}(w_1) \equiv E^{P'(x,w_0)}(x)$. Thus, we have

$$\begin{aligned}\sum_{i=1}^{q}\mathrm{SD}\left(E_i(w_0),E_i(w_1)\right) &= \sum_{i=1}^{q}\mathrm{SD}\left(E_{i+1}(w_1),E_i(w_1)\right)\\ &\geq \mathrm{SD}\left(E_{q+1}(w_1),E_1(w_1)\right)\\ &= \mathrm{SD}\left(E^{P'(x,w_0)}(x),E^{P'(x,w_1)}(x)\right).\end{aligned}$$

It follows that, for $I \leftarrow [q]$,

$$\mathop{\mathbf{E}}_{I}\left[\mathrm{SD}\left(E_I(w_0),E_I(w_1)\right)\right] \geq 1/q \cdot \mathrm{SD}\left(E^{P'(x,w_0)}(x),E^{P'(x,w_1)}(x)\right).$$

We now connect the inspection of the extractor back to the main task at hand, by observing that $\mathsf{Out}(P'_{T'}(w_b),\widehat{V}'(x))$ is distributed identically to $E_I(w_b)$, and so we have

$$\mathrm{SD}\left(\mathsf{Out}(P'_{T'}(w_0),\widehat{V}'(x)),\mathsf{Out}(P'_{T'}(w_1),\widehat{V}'(x))\right) \geq 1/q \cdot \mathrm{SD}\left(E^{P'(x,w_0)}(x),E^{P'(x,w_1)}(x)\right).$$

The claim is proved by plugging the above derivation into Equation (5). □

By plugging Claim 4.6 into Equation (4) we derive

$$\begin{aligned}&\mathrm{SD}\left(\langle P'_T(w_1),\widehat{V}(x)\rangle,\langle P'(x,w_1),\widehat{V}(x)\rangle\right)\\ &\quad\geq 1/q \cdot \mathrm{SD}(E^{P'(x,w_0)}(x),E^{P'(x,w_1)}(x)) - \mathrm{SD}\left(\langle P'(x,w_0),\widehat{V}(x)\rangle,\langle P'(x,w_1),\widehat{V}(x)\rangle\right)\end{aligned} \tag{6}$$

Finally, observe that the prover message at the last round in both $\langle P'_T(w_1),\widehat{V}(x)\rangle$ and $\langle P'(x,w_1),\widehat{V}(x)\rangle$ is sampled according to $P'(x,w_1)$. Moreover, all other prover messages in $\langle P'_T(w_1),\widehat{V}(x)\rangle$ are sampled according to $P(x,w_0)$. Thus, by data processing (using the post-processing that resample the last message according to $P(x,w_1)$) it holds that

$$\mathrm{SD}\left(\langle P'_T(w_1),\widehat{V}(x)\rangle,\langle P'(x,w_1),\widehat{V}(x)\rangle\right) \leq \mathrm{SD}\left(\langle P'(x,w_0),\widehat{V}(x)\rangle,\langle P'(x,w_1),\widehat{V}(x)\rangle\right)$$

which, together with Eq. (6) implies that

$$\mathrm{SD}(E^{P(x,w_0)}(x), E^{P(x,w_1)}(x)) \leq 2q \cdot \mathrm{SD}(\langle P'(x,w_0), \widehat{V}(x)\rangle, \langle P'(x,w_1), \widehat{V}(x)\rangle),$$

as we wanted to show.

□

4.3 Proving Corollary 4.2

We are now ready to prove Corollary 4.2.

Proof of Corollary 4.2. Suppose towards contradiction that there exists such an interactive argument (P, V) and let E be the extractor.We first show that for every $(x_0, w_0), (x_1, w_1) \in \mathcal{R}$:

$$\mathrm{SD}\left(E^{P(x_0 \vee x_1, w_0)}(x_0 \vee x_1), E^{P(x_0 \vee x_1, w_1)}(x_0 \vee x_1)\right) = \mathsf{negl}(|x_0 \vee x_1|).$$

If the protocol is constant-round, then this follows immediately by Lemma 4.3 (using $q = \mathrm{poly}(|x|)$, $k \in \mathbb{N}$, and $\delta = \mathsf{negl}(|x|)$ (where x is the combined input size). If the extractor is weakly non-adaptive, then by Lemma 4.5 there exists a verifier $\widehat{V}$ where

$$\begin{aligned}&\mathrm{SD}\left(E^{P(x_0 \vee x_1, w_0)}(x_0 \vee x_1), E^{P(x_0 \vee x_1, w_1)}(x_0 \vee x_1)\right)\\ &\leq \mathrm{poly}(|x_0 \vee x_1|) \cdot \mathrm{SD}(\langle P(x_0 \vee x_1, w_0), \widehat{V}(x_0 \vee x_1)\rangle, \langle P(x_0 \vee x_1, w_1), \widehat{V}(x_0 \vee x_1)\rangle),\end{aligned}$$

which is, in turn bounded by a negligible function by SWI of the protocol. Finally, we apply Lemma 3.1 to derive that $\mathcal{L}(\mathcal{R}) \in \mathsf{P}/\mathrm{poly}$ in contradiction to the assumption that $\mathcal{L}(\mathcal{R}) \not\subseteq \mathsf{P}/\mathrm{poly}$. □

5 Stateless Interactive Algorithms and AI-OWFs

In this section we show that the non-existence of ioAI-OWFs implies that interactive algorithms can be made stateless. In Section 5.1, we show that this holds for any algorithm A so long as the second party in the protocol is fixed in advance. In Section 5.2 we show that if A interacts for only a constant number of rounds, this can be extended to work universally for any second party. Our techniques bear similarities to previous transformations in similar contexts and for private-to-public coins for proof systems [AR21; MV24; CHKT26].

5.1 Prescribed Second Party

In this section we show that (under invertability of a specific function family) for every fixed pair of algorithms A and B, the algorithm A can be transformed into a stateless algorithm A' so that the interaction transcripts of A and A' with B are statistically close.

Theorem 5.1. *Let A and B be polynomial-time interactive algorithms and $\epsilon \in 1/\mathrm{poly}$. There exists an efficiently computable function family $\mathbf{F}$, such that if $\mathbf{F}$ is not an ioAI-OWF then there exists a polynomial-time stateless interactive algorithm A' such that for all $n \in \mathbb{N}$ and all $x, y \in \{0,1\}^n$,*

$$\mathrm{SD}(\langle A(x), B(y)\rangle, \langle A'(x,y), B(y)\rangle) \leq \epsilon(n).$$

By plugging in an interactive proof we immediately derive the following corollary:

Corollary 5.2. *Let $\epsilon \in 1/\mathrm{poly}$. Suppose that $\mathcal{R}$ has a k-round IA with completeness error α, (η, μ) computational knowledge soundness, and δ-honest-SWI. There exists an efficiently computable function family $\mathbf{F}$ such that if $\mathbf{F}$ is not an ioAI-OWF, then $\mathcal{R}$ has a k-round stateless-prover IA with completeness error $\alpha + \epsilon$, (η, μ) computational knowledge soundness, and $(\delta + \epsilon)$-honest-SWI.*

Proof of Theorem 5.1. We will use a_i to denote the i-th message of A and b_i to denote the i-th message of B. Here we assume that B is the first to communicate in the protocol. This is without loss of generality by starting with B sending an empty/dummy initial message.

We describe an auxiliary-input function f':

- $f'_{x,y,i}(\vec{r}_A, \vec{r}_B)$: Compute $\mathsf{tr} = \langle A(x;\vec{r}_A), B(y;\vec{r}_B)\rangle$ where, during the computation take note of st_{i-1}, the state of A prior to sending its i-th message. Output $(\mathsf{tr}_{<i}, b_i, \mathsf{st}_{i-1})$.[13]

For every x, y, i denote by $\mathcal{F}_{x,y,i}$ the distribution of $f'_{x,y,i}(U_A, U_B)$ where U_A, U_B are uniform strings of the appropriate lengths for the randomness of A and B. Write $\mathcal{F}_{x,y,i} = (\mathcal{F}^{(1)}_{x,y,i}, \mathcal{F}^{(2)}_{x,y,i})$ where $\mathcal{F}^{(1)}_{x,y,i}$ represents the $(\mathsf{tr}_{\leq i-1}, b_i)$ part of the output and $\mathcal{F}^{(2)}_{x,y,i}$ represents the st_{i-1} part.

By Theorem 2.2 applied to the samplable distributions $\{\mathcal{F}_{x,y,i}\}_{x,y,i}$ with parameter ϵ/k, there exists an auxiliary-input function family $\mathbf{F}$ such that if $\mathbf{F}$ is not an $\mathsf{ioAI\text{-}OWF}$, then there exists a polynomial-time algorithm Samp such that for any large enough $n \in \mathbb{N}$, any $x, y \in \{0,1\}^n$, and any $i \in [k]$:

$$\mathrm{SD}\left(\left(\mathcal{F}^{(1)}_{x,y,i}, \mathsf{Samp}(x, y, i, \mathcal{F}^{(1)}_{x,y,i})\right), \left(\mathcal{F}^{(1)}_{x,y,i}, \mathcal{F}^{(2)}_{x,y,i}\right)\right) \leq \epsilon/k. \tag{7}$$

We denote $\mathsf{Samp}_{x,y,i}(\mathsf{tr}, b) = \mathsf{Samp}(x, y, i, (\mathsf{tr}, b))$. Finally, we describe the stateless algorithm A' at round i:

- $A'(x, y, \mathsf{tr}, b)$: sample $\mathsf{st}_{i-1} \leftarrow \mathsf{Samp}_{x,y,i}(\mathsf{tr}, b)$ and sample $(a, \mathsf{st}_i) \leftarrow A(x, \mathsf{tr}, b, \mathsf{st}_{i-1})$. Output a.

It is clear by definition that A' is stateless, and it is polynomial time due to Samp being a polynomial-time machine. We now prove that A' faithfully emulates A:

Claim 5.3. *For any x, y:* $\mathrm{SD}(\langle A(x), B(y)\rangle, \langle A'(x, y), B(y)\rangle) \leq \epsilon$.

Proof. Fix x and y, and for every $1 \leq i \leq k$ consider the following experiment $\mathcal{H}_i$, in which A is stateless in the first i rounds (i.e., sample its state using Samp before sending each message, and stateful after the ith round:

1. Let $\mathsf{st}'_0, \mathsf{st}_{B,0}$, tr_0 and a_0 all equal the empty string.
2. For $j = 1$ to k:
   (a) Sample $(b_j, \mathsf{st}_{B,j}) \leftarrow B_j(\mathsf{tr}_{<j}, a_{j-1}, \mathsf{st}_{B,j-1})$.
   (b) If $1 < j \leq i$: sample $\mathsf{st}_{j-1} \leftarrow \mathsf{Samp}_{x,y,j}(\mathsf{tr}_{<j}, b_j)$.
   (c) Otherwise, let $\mathsf{st}_{j-1} = \mathsf{st}'_{j-1}$.
   (d) Sample $(a_j, \mathsf{st}'_j) \leftarrow A(x, \mathsf{tr}_{<j}, b_j, \mathsf{st}_{j-1})$.
   (e) Update $\mathsf{tr}_{<j+1} = (\mathsf{tr}_{<j} || b_j || a_j)$.
3. Output $\mathsf{tr} = \mathsf{tr}_{<k+1}$.

Observe that $\mathcal{H}_1 \equiv \langle A(x), B(y)\rangle$ and that $\mathcal{H}_k \equiv \langle A'(x, y), B(y)\rangle$ and so

$$\mathrm{SD}(\langle A(x), B(y)\rangle, \langle A'(x, y), B(y)\rangle) = \mathrm{SD}(\mathcal{H}_1, \mathcal{H}_k) \leq \sum_{i=2}^{k} \mathrm{SD}(\mathcal{H}_{i-1}, \mathcal{H}_i).$$

We conclude the claim by showing that for every $1 < i \leq k$ it holds that $\mathrm{SD}(\mathcal{H}_{i-1}, \mathcal{H}_i) \leq \epsilon/k$.

Fix $1 < i \leq k$. The only difference between $\mathcal{H}_i$ and $\mathcal{H}_{i-1}$ is the distribution used to generate st_{i-1}. By Equation (7) with x, y, and index i, for $(\mathsf{tr}_{<i}, b_i)$ sampled from $\mathcal{F}^{(1)}_{x,y,i}$, the distribution $\mathsf{Samp}_{x,y,i}(\mathsf{tr}_{<i}, b_i)$ is ϵ/k-close to distribution of st_{i-1} induced by running the real interaction and conditioning on $(\mathsf{tr}_{<i}, b_i)$. Therefore, by replacing the sampling of st_{i-1} in $\mathcal{H}_i$ with an exact sample from this conditional distribution, we change the overall transcript distribution by at most ϵ/k.

Under that replacement, st_{i-1} and st'_{i-1} are identically distributed. Hence, after the replacement, the experiment is identical to $\mathcal{H}_{i-1}$. We thus have $\mathrm{SD}(\mathcal{H}_{i-1}, \mathcal{H}_i) \leq \epsilon/k$. □

□

[13] Formally, we pad the output of $f_{x,y,i}$ to have the same output length regardless of x, y, i but we disregard this notational discrepancy throughout.

5.2 Any Second Party

In this section we show that if ioAI-OWFs do not exist, then every constant-round algorithm A can be transformed into a stateless algorithm A' so that the interaction transcripts of A and A' with B are statistically close for every B.

Theorem 5.4. *Assume that there is no* ioAI-OWF*. Let A be a polynomial-time constant-round interactive algorithm and let $\epsilon \in 1/\text{poly}$. There exists a polynomial-time stateless interactive algorithm A' such that for all $n \in \mathbb{N}$, all $x \in \{0,1\}^n$, and every (possibly inefficient) interactive algorithm B with input y,*[14]

$$\text{SD}(\langle A(x), B(y)\rangle, \langle A'(x), B(y)\rangle) \leq \epsilon(n).$$

Plugging in the prover of an interactive argument into Theorem 5.4, yields the following corollary that the prover of an interactive argument can be made stateless without changing the verifier:

Corollary 5.5. *Assume that there is no* ioAI-OWF *and let $\epsilon \in 1/\text{poly}$. Suppose that $\mathcal{R}$ has a constant-round IA with completeness error α, (η, μ) computational knowledge soundness with q queries, and δ-*SWI*. Then $\mathcal{R}$ has a stateless-prover constant-round IA with completeness error $\alpha + \epsilon$, (η, μ) computational knowledge soundness with q queries, and $(\delta + \epsilon)$-*SWI*.*

Proof of Theorem 5.4. Let $k \in \mathbb{N}$ be the number of rounds in the protocol. We will use a_i to denote the i-th message of A and b_i to denote the second party's messages. Here we assume that the other party is the first to communicate in the protocol. This is without loss of generality by starting with the other party always sending an empty/dummy initial message.

We describe auxiliary-input functions $f^{(1)}, \ldots, f^{(k)}$ and $\mathsf{Samp}^{(1)}, \ldots, \mathsf{Samp}^{(k)}$ inductively:[15]

- $f^{(i)}_{x,\mathsf{tr},b}(r, r')$:
  1. If $i = 1$ let st_{i-1} be the empty initial state.
  2. Otherwise, parse $\mathsf{tr} = \mathsf{tr}'||b_{i-1}||a_{i-1}$ and compute $\mathsf{st}_{i-1} = \mathsf{Samp}^{(i-1)}_{x,\mathsf{tr}',b_{i-1}}(a_{i-1}, r')$.
  3. Compute and output $(a, \mathsf{st}_i) = A(x, \mathsf{tr}, b, \mathsf{st}_{i-1}; r)$.

  Denote by $f^{(i),1}_{x,\mathsf{tr},b_i}$ and $f^{(i),2}_{x,\mathsf{tr},b_i}$ the first and second part of the outputs of $f^{(i)}_{x,\mathsf{tr},b_i}$ respectively, and let $(\mathcal{F}^{(i),1}_{x,\mathsf{tr},b_i}, \mathcal{F}^{(i),2}_{x,\mathsf{tr},b_i})$ be the joint distribution of the output of $f^{(i)}_{x,\mathsf{tr},b_i}$ on uniformly random inputs.

- $\mathsf{Samp}^{(i)}_{x,\mathsf{tr},b_i}(a_i)$: a next-block sampler of $f^{(i)}$ so that for every x, tr, and b_i:

$$\text{SD}\left(\left(x, \mathsf{tr}, b_i, \mathcal{F}^{(i),1}_{x,\mathsf{tr},b_i}, \mathsf{Samp}^{(i)}_{x,\mathsf{tr},b_i}(\mathcal{F}^{(i),1}_{x,\mathsf{tr},b_i})\right), \left(x, \mathsf{tr}, b_i, \mathcal{F}^{(i),1}_{x,\mathsf{tr},b_i}, \mathcal{F}^{(i),2}_{x,\mathsf{tr},b_i}\right)\right) \leq \epsilon/k. \tag{8}$$

That is, for every i, x and a partial transcript tr, b_i, a_i, $\mathsf{Samp}^{(i)}_{x,\mathsf{tr},b_i}(a_i)$ samples a state st_i for A that is aligned with the messages of A in the transcript. $f^{(i)}_{x,\mathsf{tr},b}$ then uses the state sampled by $\mathsf{Samp}^{(i-1)}$ to compute A's answer to the message b, together with A's new state. The new interactive algorithm A' is now defined as follows at round i:

- $A'(x, \mathsf{tr}, b; r, r')$: Compute $(a, \mathsf{st}) = f^{(i)}_{x,\mathsf{tr},b_i}(r, r')$ and output a.

Now that we have defined all of the algorithms, we can begin to prove their properties. Firstly, notice that each A'_i is stateless by construction. We prove that all of the above algorithms are efficiently computable:

Claim 5.6. *The algorithm A' runs in time* $\text{poly}(n)$.

Proof. Let A'_i be the algorithm of A' at round i. We show that there exist constants $c_1, \ldots, c_k \in \mathbb{N}$ such that for every i the function $f^{(i)}$ runs in time $n^{\prod_{j \leq i} c_j}$. Since k is a constant, this implies that all $f^{(i)}$-s, and therefore by construction also the A'_i-s, run in time $n^{\prod_{j \leq k} c_j} = \text{poly}(n)$.

We prove this claim by induction. As the basis, by construction it is immediate that there exists a constant c_1 so that $f^{(1)}$ runs in time n^{c_1}. Assume now that $f^{(i-1)}$ runs in time $t_{i-1} = n^{\prod_{j \leq i-1} c_j}$.

[14] As the theorem holds for any B, the additional input y is unnecessary, but added here for convenience.

[15] We use superscript notation here for these functions to resolve any potential confusion with their auxiliary inputs.

Then, since no ioAI-OWF exist, by Theorem 2.2, there exists a constant c_i' so that $\mathsf{Samp}^{(i-1)}$ can be implemented in time $t_{i-1}^{c_i'}$. Now, $f^{(i)}$ simply runs $\mathsf{Samp}^{(i-1)}$ and then A_i which is a polynomial-time machine. Therefore, there exists c_i such that $f^{(i)}$ is computable in time $t_{i-1}^{c_i} = n^{\prod_{j \le i} c_j}$ as required. □

We now prove that A' faithfully emulates A.

Claim 5.7. *For every inputs x, y and every (possibly inefficient) B:*

$$\mathrm{SD}(\langle A(x), B(y)\rangle, \langle A'(x), B(y)\rangle) \le \epsilon.$$

Proof. Fix x, y, and B, and, for every $1 \le i \le k$ consider the following experiment $\mathcal{H}_i$ in which A is stateless in the first i rounds (i.e., sample its state using Samp before sending each message, and stateful after the i-th round:

1. Let $\mathsf{st}_0', \mathsf{st}_{B,0}$, tr_0 and a_0 all equal the empty string.
2. For $j = 1$ to k:
   (a) Sample $(b_j, \mathsf{st}_{B,j}) \leftarrow B_j(y, \mathsf{tr}_{j-1}, a_{j-1}, \mathsf{st}_{B,j-1})$.
   (b) If $1 < j \le i$: sample $\mathsf{st}_{j-1} \leftarrow \mathsf{Samp}^{(j-1)}_{x,\mathsf{tr}_{j-1},b_j}(a_{j-1})$.
   (c) Otherwise, let $\mathsf{st}_{j-1} = \mathsf{st}'_{j-1}$.
   (d) Sample $(a_j, \mathsf{st}'_j) \leftarrow A(x, \mathsf{tr}_{j-1}, b_j, \mathsf{st}_{j-1})$.
   (e) Update $\mathsf{tr}_j = (\mathsf{tr}_{j-1}||b_j||a_j)$.
3. Output $\mathsf{tr} = \mathsf{tr}_k$.

Observe that $\mathcal{H}_1 \equiv \langle A(x), B\rangle$ and that $\mathcal{H}_k \equiv \langle A'(x), B\rangle$ and so using this notation

$$\mathrm{SD}(\langle A(x), B\rangle, \langle A'(x), B\rangle) = \mathrm{SD}(\mathcal{H}_1, \mathcal{H}_k) \le \sum_{i=2}^{k} \mathrm{SD}(\mathcal{H}_{i-1}, \mathcal{H}_i).$$

We conclude the claim by showing that for every $1 < i \le k$ it holds that $\mathrm{SD}(\mathcal{H}_{i-1}, \mathcal{H}_i) \le \epsilon/k$, so that

$$\mathrm{SD}(\mathcal{H}_1, \mathcal{H}_k) \le (k-1) \cdot \epsilon/k < \epsilon.$$

By Equation (8), the output of $\mathsf{Samp}^{(i-1)}_{x,\mathsf{tr}_{i-1},b_i}(a_{i-1})$ is ϵ/k-close to a sample from $\mathcal{F}^{(i-1),2}_{x,\mathsf{tr}_{i-1},b_i}$ conditioned on $\mathcal{F}^{(i-1),1}_{x,\mathsf{tr}_{i-1},b_i} = a_{i-1}$. Thus $\mathcal{H}_i$ has statistical distance at most ϵ/k from the distribution described by the following process:

1. Let $\mathsf{st}_0', \mathsf{st}_{B,0}$, tr_0 and a_0 all equal the empty string.
2. For $j = 1$ to k:
   (a) Sample $(b_j, \mathsf{st}_{B,j}) \leftarrow B_j(y, \mathsf{tr}_{j-1}, a_{j-1}, \mathsf{st}_{B,j-1})$.
   (b) If $1 < j \le i-1$: sample $\mathsf{st}_{j-1} \leftarrow \mathsf{Samp}^{(j-1)}_{x,\mathsf{tr}_{j-1},b_i}(a_{j-1})$.
   (c) If $j = i$: sample $\mathsf{st}_{i-1} \leftarrow \mathcal{F}^{(i-1),2}_{x,\mathsf{tr}_{i-1},b_i}|_{\mathcal{F}^{(i-1),1}_{x,\mathsf{tr}_{i-1},b_i} = a_{i-1}}$.
   (d) Otherwise, let $\mathsf{st}_{i-1} = \mathsf{st}'_{i-1}$.
   (e) Sample $(a_j, \mathsf{st}'_j) \leftarrow A(x, \mathsf{tr}_{j-1}, b_j, \mathsf{st}_{j-1})$.
   (f) Update $\mathsf{tr}_j = (\mathsf{tr}_{j-1}||b_j||a_j)$.
3. Output $\mathsf{tr} = \mathsf{tr}_k$.

Observe that st_{i-1} and st'_{i-1} come from exactly the same distribution. Thus, the above distribution is identical to the one described next:

1. Let $\mathsf{st}_0', \mathsf{st}_{B,0}$, tr_0 and a_0 all equal the empty string.
2. For $j = 1$ to k:
   (a) Sample $(b_j, \mathsf{st}_{B,j}) \leftarrow B_j(y, \mathsf{tr}_{j-1}, a_{j-1}, \mathsf{st}_{B,j-1})$.
   (b) If $1 < j \le i-1$: sample $\mathsf{st}_{j-1} \leftarrow \mathsf{Samp}^{(j-1)}_{x,\mathsf{tr}_{j-1},b_i}(a_{j-1})$.
   (c) If $j = i$: let $\mathsf{st}_{i-1} = \mathsf{st}'_{i-1}$.
   (d) Otherwise, let $\mathsf{st}_{i-1} = \mathsf{st}'_{i-1}$.
   (e) Sample $(a_j, \mathsf{st}'_j) \leftarrow A(x, \mathsf{tr}_{j-1}, b_j, \mathsf{st}_{j-1})$.

(f) Update $\mathsf{tr}_j = (\mathsf{tr}_{j-1}||b_j||a_j)$.

3. Output $\mathsf{tr} = \mathsf{tr}_k$.

This is equivalent to $\mathcal{H}_{i-1}$. We thus conclude that $\mathrm{SD}(\mathcal{H}_{i-1}, \mathcal{H}_i) \leq \epsilon/k$ as required. □

□

6 SWI Against a Prescribed Malicious Verifier

In this section we prove the following lemma, which states that for any SWI protocol (P, V), and for any efficient malicious verifier $\widehat{V}$, there exists an efficient prover P' such that P' is SWI against both V and $\widehat{V}$.

Lemma 6.1. *Let (P, V) be a k-round δ-honest-SWI interactive argument for a relation $\mathcal{R}$, and let $\widehat{V}$ be a PPT algorithm. There is a function family $\mathbf{F}$ such that if $\mathbf{F}$ is not an ioAI-OWF then for every $\epsilon \in 1/\mathrm{poly}$ and $Q \in \mathrm{poly}$ there exists a prover P' such that:*

- *For every x, w_0, w_1 with $(x, w_0) \in \mathcal{R}$ and $(x, w_1) \in \mathcal{R}$:*

$$\mathrm{SD}(\langle P'(x, w_0), \widehat{V}(x)\rangle, \langle P'(x, w_1), \widehat{V}(x)\rangle) \leq 20k(\epsilon + Q^2\sqrt{\delta}).$$

- *For every $(x, w) \in \mathcal{R}$: $\mathrm{SD}(\langle P(x, w), V(x)\rangle, \langle P'(x, w), V(x)\rangle) \leq 10k(\epsilon + 1/Q)$.*

The running time of P' is polynomial in $1/\epsilon$, Q, and the running times of P, V, and $\widehat{V}$. Moreover, if P is stateless, then so is P'.

Proof. We begin by describing an *inefficient* prover P''. In round $i \in [k]$, it does the following:

1. On input x, w, a_i, tr and state st_{i-1} (if $i = 1$ then st_{i-1} is empty).
2. Let $\mathsf{tr}' = (\mathsf{tr}||a_i)$.
3. Compute[16]

$$\rho_w(\mathsf{tr}') = \begin{cases} \frac{\mathbf{Pr}\left[\langle P(x,w), \widehat{V}(x)\rangle_{<i,V} = \mathsf{tr}'\right]}{\mathbf{Pr}\left[\langle P(x,w), V(x)\rangle_{<i,V} = \mathsf{tr}'\right]} & \mathbf{Pr}\left[\langle P(x,w), V(x)\rangle_{<i,V} = \mathsf{tr}'\right] > 0 \\ Q^2 + 1 & o.w. \end{cases}$$

4. Sample a number $\tau \leftarrow [0, 1]$. If $\tau < \frac{\rho_w(\mathsf{tr}')}{Q^2}$, set $\mathsf{Ind}_i = 1$, abort and output $\perp$ (in which case the protocol ends). Otherwise, set $\mathsf{Ind}_i = 0$.[17]
5. Sample $(m_i, \mathsf{st}_i) \leftarrow P(x, w, a_i, \mathsf{tr}', \mathsf{st}_{i-1})$.
6. Output message m_i and state st_i.

The lemma follows by the triangle inequality and the following two claims (and not being very tight in the bounds). The first claim shows that P'' is WI against $\widehat{V}$ but preserves the transcript distribution when interacting with the honest verifier:

Claim 6.2. *For every x, w_0, w_1 such that $(x, w_0) \in \mathcal{R}$ and $(x, w_1) \in \mathcal{R}$:*

- $\mathrm{SD}(\langle P''(x, w_0), \widehat{V}(x)\rangle, \langle P''(x, w_1), \widehat{V}(x)\rangle) \leq 6kQ^2\sqrt{\delta}$, *and*
- $\mathrm{SD}(\langle P(x, w_0), V(x)\rangle, \langle P''(x, w_0), V(x)\rangle) \leq 2k/Q$.

In the second claim we show that there exists an *efficient* prover P' that, when interacting with either V or $\widehat{V}$, produces transcripts that are statistically close:

Claim 6.3. *Assume there is no io-AI-OWF. There exists an algorithm P' whose running time is polynomial in $1/\epsilon$, Q, and the running times of P, V, and $\widehat{V}$ such that for every $(x, w) \in \mathcal{R}$:*

- $\mathrm{SD}(\langle P'(x, w), \widehat{V}(x)\rangle, \langle P''(x, w), \widehat{V}(x)\rangle) \leq 10k\epsilon$, *and*
- $\mathrm{SD}(\langle P'(x, w), V(x)\rangle, \langle P''(x, w), V(x)\rangle) \leq 10k\epsilon$.

We prove Claim 6.2 in Section 6.1 and Claim 6.3 in Section 6.2. □

[16] Recall that $\langle P(x, w), \widehat{V}(x)\rangle_{<i,V}$ denotes the transcript that contains $i - 1$ back-and-forth rounds of interaction appended with the verifier's i-th message.

[17] The variables Ind_i are used as indicator bits for convenience in the proof but do not affect the protocol.

6.1 Inefficient Prover - Proving Claim 6.2

Proof of Claim 6.2. We first prove that P'' has WI against $\widehat{V}$ and then show that it preserves the transcript distribution with the honest verifier.

Witness-indistinguishability against $\widehat{V}$. Fix x, w_0, w_1 with $(x, w_0) \in \mathcal{R}$ and $(x, w_1) \in \mathcal{R}$ and let $\delta' = \sqrt{\delta}$. We show that

$$\mathrm{SD}(\langle P''(x, w_0), \widehat{V}(x)\rangle, \langle P''(x, w_1), \widehat{V}(x)\rangle) \leq 2(k\delta' + 2kQ^2\delta/\delta') < 6kQ^2\sqrt{\delta}. \tag{9}$$

We begin by defining a number of random variables. For round index $i \in [k]$, a transcript $\mathsf{tr}_{<i}$, and verifier message a_i:

- $\mathcal{V}_i(x, \mathsf{tr}_{<i})$ be the random variable of the verifier's i-th message a_i in a random interaction of $\langle P(x, w), V(x)\rangle$ conditioned on $\mathsf{tr}_{<i}$ being the transcript of the first $i-1$ rounds of the protocol. Similarly, $\mathcal{P}_i(x, w, \mathsf{tr}_{<i}, a_i)$ is the prover's i-th message m_i conditioned on $\mathsf{tr}_{<i}$ being the first $i-1$ rounds and a_i being the i-th message of V.
- $\widehat{\mathcal{V}}_i(x, \mathsf{tr}_{<i})$ and $\widehat{\mathcal{P}}(x, w, \mathsf{tr}_{<i}, a_i)$ are defined as above, except that V is replaced by $\widehat{V}$.

Observe that since $\mathcal{P}_i(w, \mathsf{tr}_{<i}, a_i)$ and $\widehat{\mathcal{P}}_i(w, \mathsf{tr}_{<i}, a_i)$ do not depend on the verifier (only on $\mathsf{tr}_{<i}$ and a_i), they are identically distributed.

Using this notation, for every fixed partial transcript $\mathsf{tr} = (a_1, m_1, \ldots, a_{i-1}, m_{i-1}, a_i)$ for which $\mathbf{Pr}\left[\langle P(x, w), V(x)\rangle_{<i,V} = \mathsf{tr}\right] > 0$, the following holds:

$$\begin{aligned}
\rho_w(\mathsf{tr}) &= \frac{\mathbf{Pr}\left[\langle P(x, w), \widehat{V}(x)\rangle_{<i,V} = \mathsf{tr}\right]}{\mathbf{Pr}\left[\langle P(x, w), V(x)\rangle_{<i,V} = \mathsf{tr}\right]} \\
&= \frac{\mathbf{Pr}\left[\widehat{\mathcal{V}}_i(x, \mathsf{tr}_{<i}) = a_i\right]}{\mathbf{Pr}\left[\mathcal{V}_i(x, \mathsf{tr}_{<i}) = a_i\right]} \cdot \prod_{j<i} \frac{\mathbf{Pr}\left[\widehat{\mathcal{V}}_j(x, \mathsf{tr}_{<j}) = a_j\right] \mathbf{Pr}\left[\widehat{\mathcal{P}}_j(x, w, \mathsf{tr}_{<j}, a_j) = m_j\right]}{\mathbf{Pr}\left[\mathcal{V}_j(x, \mathsf{tr}_{<j}) = a_j\right] \mathbf{Pr}\left[\mathcal{P}_j(x, w, \mathsf{tr}_{<j}, a_j) = m_j\right]} \\
&= \frac{\mathbf{Pr}\left[\widehat{\mathcal{V}}_i(x, \mathsf{tr}_{<i}) = a_i\right]}{\mathbf{Pr}\left[\mathcal{V}_i(x, \mathsf{tr}_{<i}) = a_i\right]} \cdot \prod_{j<i} \frac{\mathbf{Pr}\left[\widehat{\mathcal{V}}_j(x, \mathsf{tr}_{<j}) = a_j\right] \mathbf{Pr}\left[\mathcal{P}_j(x, w, \mathsf{tr}_{<j}, a_j) = m_j\right]}{\mathbf{Pr}\left[\mathcal{V}_j(x, \mathsf{tr}_{<j}) = a_j\right] \mathbf{Pr}\left[\mathcal{P}_j(x, w, \mathsf{tr}_{<j}, a_j) = m_j\right]} \\
&= \prod_{j \leq i} \frac{\mathbf{Pr}\left[\widehat{\mathcal{V}}_j(x, \mathsf{tr}_{<j}) = a_j\right]}{\mathbf{Pr}\left[\mathcal{V}_j(x, \mathsf{tr}_{<j}) = a_j\right]}.
\end{aligned}$$

On the other hand, if $\mathbf{Pr}\left[\langle P(x, w), V(x)\rangle_{<i,V} = \mathsf{tr}\right] = 0$ then $\rho_w(\mathsf{tr}) = Q^2 + 1$ always. Putting this together, the ratio ρ_w is *independent* of w used by P'', i.e., the functions ρ_{w_0} and ρ_{w_1} are identical. As a result, for every tr, st, and verifier message a:

$$\mathbf{Pr}\left[P''(x, w_0, \mathsf{tr}, a, \mathsf{st}) = \bot\right] = \mathbf{Pr}\left[P''(x, w_1, \mathsf{tr}, a, \mathsf{st}) = \bot\right].$$

Next, letting $\mathcal{T}^0 \equiv \langle P(x, w_0), V(x)\rangle$ and $\mathcal{T}^1 \equiv \langle P(x, w_1), V(x)\rangle$ define a bad set of partial transcripts ending in a verifier message

$$\mathrm{BAD}' = \{\mathsf{tr} \colon i \in [k],\ \mathrm{SD}(\mathcal{T}^0|_{\mathcal{T}^0_{<i}=\mathsf{tr}}, \mathcal{T}^1|_{\mathcal{T}^1_{<i}=\mathsf{tr}}) \geq \delta'\}$$

to be the set of transcript prefixes for which the suffix distribution when interacting with $P(x, w_0)$ is α-far from interacting with $P(x, w_1)$. We now consider only those bad transcripts which will not be immediately rejected by P'':

$$\mathrm{BAD} = \mathrm{BAD}' \cap \{\mathsf{tr} \colon \rho_{w_0}(\mathsf{tr}) < Q^2\}.$$

Observe that for every partial transcript $\mathsf{tr} = (a_1, m_1, \ldots, a_{i-1}, m_{i-1}, a_i) \notin \mathrm{BAD}$ ending in a verifier message one of the following holds:

1. $\rho_{w_0}(\mathsf{tr}) \geq Q^2$: In this case, since $\rho_{w_0}(\mathsf{tr})/Q^2 \geq 1$, it holds that for any st, $\mathbf{Pr}\left[P''(x, w_1, \mathsf{tr}, \mathsf{st}) = \bot\right] = \mathbf{Pr}\left[P''(x, w_0, \mathsf{tr}, \mathsf{st}) = \bot\right] = 1$, or,

2. $\mathsf{tr} \notin \mathrm{BAD}'$: In this case, by definition of BAD' and data-processing,

$$\mathrm{SD}((\mathsf{tr}, \mathcal{P}''(x, w_0, \mathsf{tr})), (\mathsf{tr}, \mathcal{P}''(x, w_1, \mathsf{tr}))) < \delta'.$$

where $\mathcal{P}''(x, w, \mathsf{tr})$ is the next-message distribution of P'' using x, witness w conditioned partial transcript tr.

Putting together both cases, for every transcript $\mathsf{tr} \notin \mathrm{BAD}$,

$$\mathrm{SD}((\mathsf{tr}, \mathcal{P}''(x, w_0, \mathsf{tr})), (\mathsf{tr}, \mathcal{P}''(x, w_1, \mathsf{tr}))) < \delta'.$$

The following claim uses the fact above to bound the distinguishing probability of P'' with $\widehat{V}$.

Claim 6.4.

$$\mathrm{SD}(\langle P''(x, w_0), \widehat{V}(x)\rangle, \langle P''(x, w_1), \widehat{V}(x)\rangle) \leq 2k \left(\delta' + \max_{i \in [k]} \left\{ \mathbf{Pr}\left[\langle P''(x, w_0), \widehat{V}(x)\rangle_{<i,V} \in \mathrm{BAD}\right]\right\}\right).$$

Proof. Denote:

$$\beta := \max_{i \in [k]} \left\{ \mathbf{Pr}\left[\langle P''(x, w_0), \widehat{V}(x)\rangle_{<i,V} \in \mathrm{BAD}\right]\right\}.$$

The proof is by an hybrid argument. Let $\widehat{V}'$ be the (possibly inefficient) interactive algorithm that simulates $\widehat{V}$, but before sending a message in the i-th round, $\widehat{V}'$ checks if the partial transcript (with the i-th message) is in BAD, and if so $\widehat{V}'$ aborts (without sending the i-th message to the prover). By definition, the probability of $\widehat{V}'$ to abort in each round i is at most β. Thus, by the union bound, the probability of $\widehat{V}'$ to abort is at most $k \cdot \beta$.

Moreover, whenever $\widehat{V}'$ does not abort, it behaves exactly like $\widehat{V}$. Thus we have that

$$\mathrm{SD}(\langle P''(x, w_0), \widehat{V}(x)\rangle, \langle P''(x, w_0), \widehat{V}'(x)\rangle) \leq k \cdot \beta. \tag{10}$$

We next show that

$$\mathrm{SD}(\langle P''(x, w_0), \widehat{V}'(x)\rangle, \langle P''(x, w_1), \widehat{V}'(x)\rangle) \leq k \cdot \delta'. \tag{11}$$

Eq. (11) holds by a second hybrid argument. Indeed, for each $i \in \{0, \ldots, k\}$, let H_i be the hybrid in which the prover uses w_0 it the first i rounds and w_1 in the rest of the rounds. Then H_k is distributed according to $\langle P''(x, w_0), \widehat{V}'(x)\rangle$ and H_0 is distributed according to $\langle P''(x, w_1), \widehat{V}'(x)\rangle$. It thus enough to show that $\mathrm{SD}(H_i, H_{i+1}) \leq \delta'$ for every i. Let $\mathcal{T}_{<i,V}$ be a random variable distributed according to $\langle P''(x, w_0), \widehat{V}'(x)\rangle_{<i,V}$. Observe that by data processing,

$$\mathrm{SD}(H_i, H_{i+1}) \leq \mathrm{SD}((\mathcal{T}_{<i,V}, \mathcal{P}''(x, w_0, \mathcal{T}_{<i,V})), (T_{<i,V}, \mathcal{P}''(x, w_1, \mathcal{T}_{<i,V}))) \leq \delta'$$

Where the last inequality holds by the definition of $\widehat{V}'$ and BAD.

Finally, we conclude that by Eq. (11) together with the fact that $\widehat{V}'$ aborts with probability at most $k\beta$ in the interaction with $P''(x, w_0)$, that V' aborts with probability at most $k(\beta + \delta')$ in the interaction with $P''(x, w_1)$. It follows that

$$\mathrm{SD}(\langle P''(x, w_1), \widehat{V}(x)\rangle, \langle P''(x, w_1), \widehat{V}'(x)\rangle) \leq k \cdot (\delta' + \beta). \tag{12}$$

The claim now follows by Eqs. (10) to (12) and the triangle inequality. □

Thus, to achieve Equation (9) it is enough to bound

$$\mathbf{Pr}\left[\langle P''(x, w_0), \widehat{V}(x)\rangle_{<i,V} \in \mathrm{BAD}\right] \leq \mathbf{Pr}\left[\langle P(x, w_0), \widehat{V}(x)\rangle_{<i,V} \in \mathrm{BAD}\right]. \tag{13}$$

Toward this, fix $i \in [k]$. We first bound the probability of BAD when interacting with the *honest* verifier. By δ-honest-SWI, the definitions of the sets BAD and BAD' and by Lemma 2.17, it holds that

$$\mathbf{Pr}\left[\langle P(x, w_0), V(x)\rangle_{<i,V} \in \mathrm{BAD}\right] \leq \mathbf{Pr}\left[\langle P(x, w_0), V(x)\rangle_{<i,V} \in \mathrm{BAD}'\right] \leq 2\delta/\delta'.$$

Moreover, by the definition of P'', for every $\mathsf{tr} \in \text{BAD}$ we have that

$$\rho_{w_0}(\mathsf{tr}) = \frac{\mathbf{Pr}\left[\langle P(x,w_0), \widehat{V}(x)\rangle_{<i,V} = \mathsf{tr}\right]}{\mathbf{Pr}\left[\langle P(x,w_0), V(x)\rangle_{<i,V} = \mathsf{tr}\right]},$$

which implies that

$$\mathbf{Pr}\left[\langle P(x,w_0), \widehat{V}(x)\rangle_{<i,V} = \mathsf{tr}\right] = \rho_{w_0}(\mathsf{tr}) \cdot \mathbf{Pr}\left[\langle P(x,w_0), V(x)\rangle_{<i,V} = \mathsf{tr}\right] \leq Q^2 \cdot \mathbf{Pr}\left[\langle P(x,w_0), V(x)\rangle_{<i,V} = \mathsf{tr}\right].$$

Finally, it follows that

$$\mathbf{Pr}\left[\langle P(x,w_0), \widehat{V}(x)\rangle_{<i,V} \in \text{BAD}\right] \leq Q^2 \cdot \mathbf{Pr}\left[\langle P(x,w_0), V(x)\rangle_{<i,V} \in \text{BAD}\right] \leq 2Q^2\delta/\delta'. \tag{14}$$

Equation (9) now follows by combining Claim 6.4 and Eqs. (13) and (14).

Preserving distance with honest verifier. We now show that for every $(x,w) \in \mathcal{R}$:

$$\text{SD}(\langle P(x,w), V(x)\rangle, \langle P''(x,w), V(x)\rangle) \leq 2k/Q.$$

Let $\overline{P}$ be the prover that runs P but additionally computes and records the values Ind_i as in P''. Observe that given a transcript sampled from $\langle P(x,w), V(x)\rangle$ and the bits $\mathsf{Ind}_1, \ldots, \mathsf{Ind}_k$ it is possible to precisely calculate the transcript when interacting with P'': if there is an i such that $m_i = 1$, then take the first such index, output $\perp$ as its message, and indicate the prover aborted. If there is no such index, then simply output the transcript as is. In other words, the transcript and recorded values of an interaction with $\overline{P}$ are enough to compute the transcript when interacting with P''. Thus, by the triangle inequality and data-processing:

$$\begin{aligned}
&\text{SD}(\langle P(x,w), V(x)\rangle, \langle P''(x,w), V(x)\rangle) \\
&= \text{SD}((\langle \overline{P}(x,w), V(x)\rangle, 0, \ldots, 0), (\langle \overline{P}(x,w), V(x)\rangle, \mathsf{Ind}_1, \ldots, \mathsf{Ind}_k)) \\
&\leq \mathbf{Pr}\left[\exists i \in [k] : \mathsf{Ind}_i = 1\right].
\end{aligned}$$

Let S be the set of all partial transcripts $\mathsf{tr}_{<i,V}$ so that $\rho_w(\mathsf{tr}_{<i,V}) \geq Q$ and $\rho_w(\mathsf{tr}_{<j,V}) < Q$ for all $j < i$ (such that each transcript tr has at most one prefix $\mathsf{tr}_{<i,V}$ in the set S). Then

$$\begin{aligned}
\mathop{\mathbf{Pr}}_{\mathsf{tr} \leftarrow \langle P(x,w), V(x)\rangle}\left[\exists i : \mathsf{tr}_{<i,V} \in S\right] &= \sum_{\mathsf{tr} \in S} \mathbf{Pr}\left[\exists i \colon \langle P(x,w), V(x)\rangle_{<i,V} = \mathsf{tr}\right] \\
&\leq \sum_{\mathsf{tr} \in S} 1/Q \cdot \mathbf{Pr}\left[\exists i \colon \langle P(x,w), \widehat{V}(x)\rangle_{<i,V} = \mathsf{tr}\right] \\
&= 1/Q \cdot \sum_{\mathsf{tr} \in S} \mathbf{Pr}\left[\exists i \colon \langle P(x,w), \widehat{V}(x)\rangle_{<i,V} = \mathsf{tr}\right] \\
&\leq 1/Q.
\end{aligned}$$

Furthermore, for every i and $\mathsf{tr} \notin S$ containing $i-1$ rounds, Ind_i becomes 1 when $\tau < 1/Q$, which occurs with probability $1/Q$. Thus,

$$\begin{aligned}
\mathbf{Pr}\left[\exists i \in [k] : \mathsf{Ind}_i = 1\right] &\leq \mathop{\mathbf{Pr}}_{\mathsf{tr} \leftarrow \langle P(x,w), V(x)\rangle}\left[\exists i \colon \mathsf{tr}_{<i,V} \in S\right] + \mathop{\mathbf{Pr}}_{\mathsf{tr} \leftarrow \langle P(x,w), V(x)\rangle}\left[\exists i \in [k] \colon \mathsf{Ind}_i = 1 \mid \forall i \colon \mathsf{tr}_{<i,V} \notin S\right] \\
&\leq 1/Q + \sum_{i \in [k]} \mathop{\mathbf{Pr}}_{\mathsf{tr} \leftarrow \langle P(x,w), V(x)\rangle}\left[\mathsf{Ind}_i = 1 \mid \forall i \colon \mathsf{tr}_{<i,V} \notin S\right] \\
&\leq (k+1)/Q \leq 2k/Q.
\end{aligned}$$

□

6.2 Making the Prover Efficient - Proving Claim 6.3

We next show how to implement P'' efficiently assuming there is are no one-way functions.

Proof of Claim 6.3. We begin with defining a few distributions. Let $\mathcal{T}_{x,w}$ be the distribution $\langle P(x,w), V(x)\rangle$ and $\mathcal{T}_{x,w,i}$ be the distribution of $\langle P(x,w), V(x)\rangle_{<i,V}$. Let $\widehat{\mathcal{T}}_{x,w}$ and $\widehat{\mathcal{T}}_{x,w,i}$ be similarly defined with respect to the distribution of $\langle P(x,w), \widehat{V}(x)\rangle$.

Approximating transcript weights. We now define an algorithm Pred that approximates whether a transcript is more likely to come from a conversation with $\widehat{V}$ compared to the real transcript distribution. Consider the function $f_{x,w,i}(b, r_P, r_V) = (|T|, T, b)$ where $|T|$ is the bit length of T and:

- If $b = 0$ then $T \leftarrow \mathcal{T}_{x,w,i}$.
- If $b = 1$ then $T \leftarrow \widehat{\mathcal{T}}_{x,w,i}$.

Observe that $f_{i,x,w}$ can be computed in time polynomial in the combined running time of P and V. For a partial transcript tr, let

$$p_{x,w,i}(\mathsf{tr}) = \mathop{\mathbf{Pr}}_{\substack{b,r_P,r_V,\\(|T|,T,b)=f_{x,w,i}(b,r_P,r_V)}} [b = 0 \mid T = \mathsf{tr}].$$

We are interested in this probability due to the following derivation, showing its relationship to $\rho_w(\mathsf{tr})$:

$$\begin{aligned} p_{x,w,i}(\mathsf{tr}) &= \mathbf{Pr}[b = 0 \mid T = \mathsf{tr}] \\ &= \frac{\mathbf{Pr}[T = \mathsf{tr}, b = 0]}{\mathbf{Pr}[T = \mathsf{tr}, b = 1] + \mathbf{Pr}[T = \mathsf{tr}, b = 0]} \\ &= \frac{1/2 \cdot \mathbf{Pr}[\mathcal{T}_{x,w,i} = \mathsf{tr}]}{1/2 \cdot \mathbf{Pr}\left[\widehat{\mathcal{T}}_{x,w,i} = \mathsf{tr}\right] + 1/2 \cdot \mathbf{Pr}[\mathcal{T}_{x,w,i} = \mathsf{tr}]} \\ &= \frac{\mathbf{Pr}[\mathcal{T}_{x,w,i} = \mathsf{tr}]}{\mathbf{Pr}\left[\widehat{\mathcal{T}}_{x,w,i} = \mathsf{tr}\right] + \mathbf{Pr}[\mathcal{T}_{x,w,i} = \mathsf{tr}]}, \end{aligned}$$

which implies that either $p_{x,w,i}(\mathsf{tr}) = 0$ (corresponding to $\mathbf{Pr}[\mathcal{T}_{x,w,i} = \mathsf{tr}] = 0$) or

$$1/p_{x,w,i}(\mathsf{tr}) = \frac{\mathbf{Pr}\left[\widehat{\mathcal{T}}_{x,w,i} = \mathsf{tr}\right]}{\mathbf{Pr}[\mathcal{T}_{x,w,i} = \mathsf{tr}]} + 1 = \rho_w(\mathsf{tr}) + 1. \tag{15}$$

where $\rho_w(\mathsf{tr})$ is defined as in the description of P''.

We are therefore interested in efficiently approximating $p_{x,w,i}(\mathsf{tr})$. By Theorem 2.4, there exists a function family $\mathbf{F}$ such that if $\mathbf{F}$ is not an $\mathsf{ioAI\text{-}OWF}$ then, since $\epsilon \in 1/\mathrm{poly}$ there is an algorithm Pred that runs in time polynomial in the running time of f, such that for every (x, w, i):

$$\mathop{\mathbf{Pr}}_{\substack{b,r_P,r_V,\\(|T|,T,b)=f_{x,w,i}(b,r_P,r_V)}} [\mathsf{Pred}_{x,w,i}(|T|, T) \in (1 \pm \epsilon) \cdot p_{x,w,i}(T)] \geq 1 - \epsilon.$$

We shall henceforth omit the input $|T|$ from Pred (formally, this just means defining $\widehat{\mathsf{Pred}}_{x,w,i}(T)$ as the algorithm that receives T and runs $\mathsf{Pred}_{x,w,i}(|T|, T)$). It follows that

$$\mathop{\mathbf{Pr}}_{\substack{b=0,r_P,r_V\\(|T|,T,b)=f_{x,w,i}(b,r_P,r_V)}} [\mathsf{Pred}_{x,w,i}(T) \in (1 \pm \epsilon) \cdot p_{x,w,i}(T)] \geq 1 - 2\epsilon.$$

Or equivalently,

$$\mathop{\mathbf{Pr}}_{T \leftarrow \mathcal{T}_{x,w,i}} [\mathsf{Pred}_{x,w,i}(T) \in (1 \pm \epsilon) \cdot p_{x,w,i}(T)] \geq 1 - 2\epsilon. \tag{16}$$

Similarly,

$$\mathop{\mathbf{Pr}}_{T \leftarrow \widehat{\mathcal{T}}_{x,w,i}} [\mathsf{Pred}_{x,w,i}(T) \in (1 \pm \epsilon) \cdot p_{x,w,i}(T)] \geq 1 - 2\epsilon. \tag{17}$$

Describing the prover. We can now define the prover P' at round i:

1. On input x, w, a_i, tr and state st_{i-1} (if $i = 1$ then st_{i-1} is empty).
2. Let $\mathsf{tr}' = (\mathsf{tr}||a_i)$.
3. Run $p' \leftarrow \mathsf{Pred}_{x,w,i}(\mathsf{tr}'; r')$ and set

$$\widetilde{\rho}_w(\mathsf{tr}') = \begin{cases} 1/p' - 1 & p' > 0 \\ Q^2 + 1 & o.w. \end{cases}$$

4. Sample a number $\tau \leftarrow [0,1]$. If $\tau < \widetilde{\rho}_w(\mathsf{tr}')/Q^2$, set biased bit $\widetilde{\mathsf{Ind}}_i = 1$, abort and output $\perp$ (in which case the protocol ends). Otherwise, set $\widetilde{\mathsf{Ind}}_i = 0$, and continue to the next step.
5. Sample $(m_i, \mathsf{st}_i) \leftarrow P(x, w, a_i, \mathsf{tr}', \mathsf{st}_{i-1})$.
6. Output message m_i and state st_i.

Observe that P' runs in time $\mathrm{poly}(t)$ where t is the combined running time of P and $\widehat{V}$. Furthermore, if P is stateless, then so is P'.

Closeness of transcripts. We next show that

$$\mathrm{SD}(\langle P'(x,w), \widehat{V}(x)\rangle, \langle P''(x,w), \widehat{V}(x)\rangle) \leq 10k\epsilon.$$

The second item of the claim (that $\mathrm{SD}(\langle P'(x,w), V(x)\rangle, \langle P''(x,w), V(x)\rangle) \leq 12k\epsilon$) can be proven identically, replacing interaction with $\widehat{V}$ with V.

The proof is by a coupling argument. To bound the statistical distance, we consider a new prover $\overline{P}$, that in each round i, computes the value of both Ind_i and $\widetilde{\mathsf{Ind}}_i$, as in the description of P' and P'' respectively, but does not abort and continues to interact with $\widehat{V}$ as P does. If in some round $\mathsf{Ind}_i = 1$ (resp. $\widetilde{\mathsf{Ind}}_i = 1$), $\overline{P}$ sets the value of Ind_j (resp. $\widetilde{\mathsf{Ind}}_j$) to be 1 for every $j > i$. We observe that given a transcript of $\langle \overline{P}(x,w), \widehat{V}(x)\rangle$ together with the values of $\mathsf{Ind}_1, \ldots, \mathsf{Ind}_k$, we can generate the transcript of $\langle P''(x,w), \widehat{V}(x)\rangle$ (distributed according to the right distribution) by looking for the first location where $\mathsf{Ind}_i = 1$, sending $\perp$ as the prover message at that round, and truncating the protocol, indicating that the prover has aborted. Similarly, given a transcript of $\langle \overline{P}(x,w), \widehat{V}(x)\rangle$ together with the values of $\widetilde{\mathsf{Ind}}_1, \ldots, \widetilde{\mathsf{Ind}}_k$, we can generate the transcript of $\langle P'(x,w), \widehat{V}(x)\rangle$. Thus, by data-processing, it holds that

$$\begin{aligned}
&\mathrm{SD}(\langle P'(x,w), \widehat{V}(x)\rangle, \langle P''(x,w), \widehat{V}(x)\rangle) \\
&\leq \mathrm{SD}((\langle \overline{P}(x,w), \widehat{V}(x)\rangle, (I_1, \ldots, I_k)), (\langle \overline{P}(x,w), \widehat{V}(x)\rangle, (\widetilde{I}_1, \ldots, \widetilde{I}_k))) \\
&\leq \mathbf{Pr}\left[\exists i : I_i \neq \widetilde{I}_i\right] \leq \sum_i \mathbf{Pr}\left[I_i \neq \widetilde{I}_i \mid \text{ no prior indices are different}\right],
\end{aligned}$$

where I_i and $\widetilde{I}_i$ are the random variables taking the values of Ind_i and $\widetilde{\mathsf{Ind}}_i$ respectively. That is, it is enough to bound the probability that for some $i \in [k]$, $I_i \neq \widetilde{I}_i$.

For each round $i \in [k]$, let R_i and $\widetilde{R}_i$ be the jointly distributed random variable taking the value of $\rho_w(\mathsf{tr})$ and $\widetilde{\rho}_w(\mathsf{tr})$, respectively, in a random execution of $\langle \overline{P}(x,w), \widehat{V}(x)\rangle$, and let Z be the random variable taking the value of τ. Define $\gamma = 5\epsilon Q^2$. We have that

$$\begin{aligned}
&\mathbf{Pr}\left[I_i \neq \widetilde{I}_i \mid \text{ no prior indices are different}\right] \\
&\quad = \mathbf{Pr}\left[(R_i/Q^2 < Z < \widetilde{R}_i/Q^2) \vee (\widetilde{R}_i/Q^2 < Z < R_i/Q^2)\right] \\
&\quad \leq \mathbf{Pr}\left[\Big((R_i/Q^2 < Z < \widetilde{R}_i/Q^2) \vee (\widetilde{R}_i/Q^2 < Z < R_i/Q^2)\Big) \wedge |R_i - \widetilde{R}_i| < \gamma\right] \\
&\qquad + \mathbf{Pr}\left[|R_i - \widetilde{R}_i| \geq \gamma \wedge \min(R_i, \widetilde{R}_i) \leq Q^2\right] \\
&\quad \leq \mathbf{Pr}\left[|R_i - \widetilde{R}_i| < \gamma \ \wedge \ Z \text{ lies between } R_i/Q^2 \text{ and } \widetilde{R}_i/Q^2\right] \\
&\qquad + \mathbf{Pr}\left[|R_i - \widetilde{R}_i| \geq \gamma \wedge \min(R_i, \widetilde{R}_i) \leq Q^2\right] \\
&\quad \leq \frac{\gamma}{Q^2} + \mathbf{Pr}\left[|R_i - \widetilde{R}_i| \geq \gamma \wedge \min(R_i, \widetilde{R}_i) \leq Q^2\right] \\
&\quad = \frac{\gamma}{Q^2} + \mathbf{Pr}\left[|R_i - \widetilde{R}_i| \geq \gamma \wedge \min(R_i, \widetilde{R}_i) \leq Q^2\right].
\end{aligned}$$

Where the final inequality holds because Z is uniform over $[0,1]$ and the gap it needs to hit is of length at most $\frac{\gamma}{Q^2}$. We next want to show that

$$\mathbf{Pr}\left[|R_i - \widetilde{R}_i| \geq \gamma \wedge \min(R_i, \widetilde{R}_i) \leq Q^2\right] \leq 2\epsilon, \tag{18}$$

which implies that, $\mathbf{Pr}\left[I_i \neq \widetilde{I}_i\right] \leq \frac{\gamma}{Q^2} + 2\epsilon = 7\epsilon < 10\epsilon$.[18] To show Eq. (18), we observe that by the accuracy of Pred, (i.e. Eq. (17)), we have that

$$\mathop{\mathbf{Pr}}_{T \leftarrow \widehat{T}_{x,w,i}} [\mathsf{Pred}_{x,w,i}(T) \in (1 \pm \epsilon) \cdot p_{x,w,i}(T)] \geq 1 - 2\epsilon.$$

By Eq. (15), under the event in which $\mathsf{Pred}_{x,w,i}(T) \in (1 \pm \epsilon) \cdot p_{x,w,i}(T)$, we have that

$$1/\mathsf{Pred}_{x,w,i}(T) \in \frac{1}{(1 \pm \epsilon) \cdot p_{x,w,i}(T)} \subseteq \frac{(1 \pm 2\epsilon)}{p_{x,w,i}(T)} = (1 \pm 2\epsilon)(\rho_w(T) + 1).$$

It follows that with probability at least $1 - 2\epsilon$:

$$(1 - 2\epsilon)R_i - 2\epsilon < \widetilde{R}_i < (1 + 2\epsilon)R_i + 2\epsilon.$$

Suppose this holds. We show under this condition it cannot be that $|R_i - \widetilde{R}_i| \geq \gamma = 5\epsilon Q^2$ and that $\min(R_i, \widetilde{R}_i) \leq Q^2$. We look at two cases:

- If $R_i > 3Q^2/2$: then

  $$\widetilde{R}_i > (1 - 2\epsilon)R_i - 2\epsilon > (1 - 2\epsilon)3Q^2/2 - 2\epsilon > Q^2,$$

  where the final inequality assumes $\epsilon < 1/10$, which we can safely do since when this does not hold the main lemma statement becomes trivial. This invalidates the requirement that $\min(R_i, \widetilde{R}_i) \leq Q^2$.

- If $R_i \leq 3Q^2/2$: observe that by the bound on $\widetilde{R}$, we have $|R_i - \widetilde{R}_i| < 2\epsilon R_i + 2\epsilon$. Thus,

  $$|R_i - \widetilde{R}_i| < 3\epsilon Q^2 + 2\epsilon \leq 5\epsilon Q^2 = \gamma$$

  This invalidates the requirement that $|R_i - \widetilde{R}_i| \geq \gamma$.

□

7 CWI Arguments of Knowledge and AI-OWFs

In this section we derive our main theorems: one for general constant-round CWI protocols of knowledge, and the second for many-round CWI protocols of knowledge with a weakly non-adaptive extractor.

Theorem 7.1. *Let $\mathcal{R}$ be an NP relation with $\mathcal{L}(\mathcal{R}) \notin \mathsf{P}/\mathrm{poly}$ and assume there is a δ-CWI interactive argument for the OR relation $\mathcal{R} \vee \mathcal{R}$ with round-complexity k, completeness error α, and (η, μ)-knowledge soundness with a q-query extractor. If k is constant and there exists $\epsilon \in 1/\mathrm{poly}$ with:*

- $\alpha + \eta < 1 - \epsilon$, *and*
- $2\mu + 2kq^{k+1}\delta < 1 - \epsilon$.

Then infinitely-often auxiliary-input one-way functions exist.

Theorem 7.2. *Let $\mathcal{R}$ be an NP relation with $\mathcal{L}(\mathcal{R}) \notin \mathsf{ioP}/\mathrm{poly}$ and assume there is a δ-honest-CWI interactive argument for the OR relation $\mathcal{R} \vee \mathcal{R}$ with round complexity k, completeness error α, and (η, μ) computational knowledge soundness with a q-query extractor. If the knowledge extractor is weakly non-adaptive and there exist $\epsilon \in 1/\mathrm{poly}$ and $Q \in \mathrm{poly}$ with:*

- $\alpha + \eta + 12k/Q < 1 - \epsilon$, *and*
- $2\mu + 80qkQ^2\sqrt{\delta} < 1 - \epsilon$.

Then auxiliary-input one-way functions exist.

[18] We are taking a lot of slack with parameters here. This is only used to simplify the main statement: we will later need $\varepsilon < 1/10$, and bounding by 10ϵ means we can assume this always since otherwise the statement is trivial.

By plugging in negligible parameters and choosing ϵ and Q appropriately, we have the following corollaries. Recall that if not stated otherwise, CWI arguments of knowledge have negligible completeness, knowledge soundness and WI errors:

Corollary 7.3. *Let $\mathcal{R}$ be an* NP *relation with $\mathcal{L}(\mathcal{R}) \notin \mathsf{P}/\mathrm{poly}$. Suppose that there is a constant-round* CWI *argument of knowledge for the OR relation $\mathcal{R} \vee \mathcal{R}$. Then infinitely-often auxiliary-input one-way functions exist.*

Corollary 7.4. *Let $\mathcal{R}$ be an* NP *relation with $\mathcal{L}(\mathcal{R}) \notin \mathsf{ioP}/\mathrm{poly}$. Suppose that there is a* CWI *argument of knowledge with a weakly non-adaptive extractor for the OR relation $\mathcal{R} \vee \mathcal{R}$. Then auxiliary-input one-way functions exist.*

7.1 Constant Rounds - Proving Theorem 7.1

Proof of Theorem 7.1. Let $\epsilon' \in 1/\mathrm{poly}$ be a parameter to be specified later. Assume (towards contradiction to the assumption that $\mathcal{L}(\mathcal{R}) \notin \mathsf{P}/\mathrm{poly}$) that there are no ioAI-OWF. We go through a number of steps:

1. **CWI to SWI:** Since there are no ioAI-OWF through Lemma 2.15, we observe that the argument (P, V) is $(\delta + \epsilon')$-SWI.
2. **Stateless prover:** Utilizing Corollary 5.5 (since there are no ioAI-OWF), we derive an $(\delta + 2\epsilon')$-SWI argument system (P', V) for $\mathcal{R} \vee \mathcal{R}$ with completeness error $\alpha + \epsilon'$, and (η, μ) computational knowledge soundness with q extractor queries where the prover P' is stateless.
3. **Fooling the extractor:** Let E be the extractor for (P', V). By Lemma 4.3, we have that for every $(x_0, w_0), (x_1, w_1) \in \mathcal{R}$:
$$\mathrm{SD}\left(E^{P'(x_0 \vee x_1, w_0)}(x_0 \vee x_1), E^{P'(x_0 \vee x_1, w_1)}(x_0 \vee x_1)\right) \leq 2kq^{k+1}(\delta + 2\epsilon').$$
4. **Deciding the language:** Finally, we use Lemma 3.1 to show that $\mathcal{L}(\mathcal{R}) \in \mathsf{P}/\mathrm{poly}$, which is a contradiction. This requires that:
   (a) $1 - (\alpha + \epsilon') > \eta$ and,
   (b) there exists $\epsilon'' \in 1/\mathrm{poly}$ such that $2\mu + 2kq^{k+1}(\delta + 2\epsilon') < 1 - \epsilon''$.

   Setting $\epsilon' = \frac{\epsilon}{8kq^{k+1}}$ and $\epsilon'' = \epsilon/2$, these are both satisfied by the assumptions in the theorem statement.

□

7.2 Weakly Non-Adaptive Extractor - Proving Theorem 7.2

Proof of Theorem 7.2. Let $\epsilon' \in 1/\mathrm{poly}$ be a parameter to be specified later. We first assume $\mathcal{L}(\mathcal{R}) \notin \mathsf{P}/\mathrm{poly}$ and derive ioAI-OWF, and then in Remark 7.5 discuss how to derive AI-OWF from the assumption that $\mathcal{L}(\mathcal{R}) \notin \mathsf{ioP}/\mathrm{poly}$.

Assume $\mathcal{L}(\mathcal{R}) \notin \mathsf{P}/\mathrm{poly}$. We go through a number of steps:

1. **Honest-CWI to honest-SWI:** By Lemma 2.15 there is a function family $\mathbf{F}^{(1)}$ such that if $\mathbf{F}^{(1)}$ is not an ioAI-OWF, then the argument (P, V) is $(\delta + \epsilon')$-honest-SWI.
2. **Stateless prover:** Utilizing Corollary 5.2, there is a function family $\mathbf{F}^{(2)}$ such that if $\mathbf{F}^{(2)}$ is not an ioAI-OWF, then there is an $(\delta + 2\epsilon')$-honest-SWI argument system (P', V) for $\mathcal{R} \vee \mathcal{R}$ with completeness error $\alpha + \epsilon'$, and (η, μ) computational knowledge soundness with q extractor queries where the prover P' is stateless.
3. **Fooling the extractor:** Let E be the extractor for (P', V). By Lemma 4.5, we have that there exists a verifier $\widehat{V}$ such that for every P'' and every $(x_0, w_0), (x_1, w_1) \in \mathcal{R}$:
$$\begin{aligned}&\mathrm{SD}\left(E^{P''(x_0 \vee x_1, w_0)}(x_0 \vee x_1), E^{P''(x_0 \vee x_1, w_1)}(x_0 \vee x_1)\right)\\ &\leq 2q \cdot \mathrm{SD}(\langle P''(x_0 \vee x_1, w_0), \widehat{V}(x_0 \vee x_1)\rangle, \langle P''(x_0 \vee x_1, w_1), \widehat{V}(x_0 \vee x_1)\rangle).\end{aligned}$$

4. **SWI against $\widehat{V}$:** According to Lemma 6.1 there a function family $\mathbf{F}^{(3)}$ such that if $\mathbf{F}^{(3)}$ is not an ioAI-OWF, then we can derive a stateless prover P'' to protect against the verifier $\widehat{V}$ defined in the previous item with the properties:

   - For every $(x_0, w_0), (x_1, w_1) \in \mathcal{R}$, letting $x = x_0 \vee x_1$:

     $$\mathrm{SD}(\langle P''(x, w_0), \widehat{V}(x)\rangle, \langle P''(x, w_1), \widehat{V}(x)\rangle) \leq 20k(\epsilon' + Q^2\sqrt{\delta + 2\epsilon'}) < 40kQ^2\left(\sqrt{\epsilon'} + \sqrt{\delta}\right).$$

   - For every $(x, w) \in \mathcal{R} \vee \mathcal{R}$: $\mathrm{SD}(\langle P'(x, w), V(x)\rangle, \langle P''(x, w), V(x)\rangle) \leq 10k(\epsilon' + 1/Q)$.

   As a result, the protocol (P'', V) has completeness error at most

   $$\alpha + \epsilon' + 10k(\epsilon' + 1/Q) < \alpha + 12k(\epsilon' + 1/Q).$$

   Moreover, it holds that

   $$\mathrm{SD}\left(E^{P''(x_0 \vee x_1, w_0)}(x_0 \vee x_1), E^{P''(x_0 \vee x_1, w_1)}(x_0 \vee x_1)\right) \leq 80qkQ^2\left(\sqrt{\epsilon'} + \sqrt{\delta}\right)$$

5. **Deciding the language:** Finally, we use Lemma 3.1 to show that $\mathcal{L}(\mathcal{R}) \in \mathsf{P}/\mathrm{poly}$, contradicting our assumption. This requires that:

   (a) $1 - (\alpha + 12k(\epsilon' + 1/Q)) > \eta$ and,

   (b) there exists $\epsilon'' \in 1/\mathrm{poly}$ such that $2\mu + 80qkQ^2\left(\sqrt{\epsilon'} + \sqrt{\delta}\right) < 1 - \epsilon''$.

   Setting $\epsilon' = (\frac{\epsilon}{80qkQ^2})^2$ and $\epsilon'' = \epsilon/2$ which are both inverse-polynomial, the requirements are both satisfied by the assumptions in the theorem statement.

As a result, we derive that at least one of the function families $\mathbf{F}^{(1)}, \mathbf{F}^{(2)}, \mathbf{F}^{(3)}$ is an ioAI-OWF.

Remark 7.5 (AI-OWF vs. ioAI-OWF)**.** *We describe how we can derive the existence of* AI-OWF *if we start from the assumption that* $\mathcal{L}(\mathcal{R}) \notin \mathsf{ioP}/\mathrm{poly}$.

Recall that we derived function families $\mathbf{F}^{(1)}, \mathbf{F}^{(2)}, \mathbf{F}^{(3)}$ *by using Lemmas 2.15 and 6.1 and Corollary 5.2 and these definitions do not depend on each other being invertible. We observe that all of these proofs, apply by input-length in the sense that each statement applies on every auxiliary-input on which function family can be efficiently inverted.*

Denoting $\mathbf{F}^{(i)} = \{f^{(i)}_{a_i}\}_{a_i \in \{0,1\}^{\lambda_i(n)}}$*, construct the function family*

$$\mathbf{F} = \left\{f_{a_1,a_2,a_3}(x_1, x_2, x_3) = \left(f^{(1)}_{a_1}(x_1), f^{(2)}_{a_2}(x_2), f^{(3)}_{a_3}(x_3)\right)\right\}_{a_1 \in \{0,1\}^{\lambda_1(n)}, a_2 \in \{0,1\}^{\lambda_2(n)}, a_3 \in \{0,1\}^{\lambda_3(n)}}.$$

On every length n *on which we can invert* $\mathbf{F}$*, we can invert all* $\mathbf{F}^{(i)}$ simultaneously*, and thus make the same derivations as in Items 1, 2 and 4. This suffices in Lemma 3.1 for constructing a polynomial-size circuit deciding* $\mathcal{L}(\mathcal{R})$ *on inputs of size* n*. Thus, if* $\mathbf{F}$ *is not an* AI-OWF*, we get a polynomial-size family of circuits deciding* $\mathcal{L}(\mathcal{R})$ *for infinitely many* $n \in \mathbb{N}$.

□

8 One-Way Permutations and SWI

The goal of this section is to separate SWI and one-way functions. For this we use the Sam oracle introduced by [HHRS07]. Before stating our results we give some definitions.

8.1 The Oracle Sam

We start with the description of Sam. Sam^{π}_{d} takes as input an interactive oracle-aided Turing machine M together with a partial transcript tr of an interaction with M for at most d rounds and samples random coins for M^{π} that are consistent with the transcript, as long as the transcript continues a previous conversation with Sam^{π}_{d}. We use the following formulation from [PV10].

Definition 8.1 (Sam_d^π). *Let $\pi = \{\pi_n \colon \{0,1\}^n \to \{0,1\}^n\}_{n\in\mathbb{N}}$ be a family of permutations, and M an oracle-aided, interactive TM in a d-round protocol. For a partial transcript $\mathsf{tr} = (a_1, b_1, \ldots, a_i, b_i)$, let $R_{\mathsf{tr}}(M^\pi)$ be the set of all random tapes τ such that $M^\pi(a_1, b_1, \ldots, a_{j-1}, b_{j-1}; \tau) = a_j$ for all $j < i$.*

Sam_d^π takes as input $(M^\pi, \mathsf{tr}; r)$ where tr is a partial transcript, and $r \in \{0,1\}^*$ is a random string, and returns (τ', tr'), where $\tau' \in R_{\mathsf{tr}}(M^\pi)$ where for a random r, $\tau' \leftarrow R_{\mathsf{tr}}(M^\pi)$, and $\mathsf{tr}' = \mathsf{tr}||M^\pi(\mathsf{tr}; \tau')$, with the following restrictions:

1. *If $\mathsf{tr} = (a_1, b_1, \ldots, a_i, b_i)$ for $i > 1$, Sam_d^π only answers if $\mathsf{tr}' = (a_1, b_1, \ldots, a_i)$ is the output of a previous query made to Sam_d^π.*
2. *If $\mathsf{tr} = (a_1, b_1, \ldots, a_i, b_i)$, Sam_d^π only answers if $i \le d(n)$.*

Otherwise, Sam_d^π outputs $\perp$.

We remark that while the above algorithm is stateful, it can be made stateless using signatures, see [HHRS07] for details. [HHRS07] showed that for $d(n) = o(n/\log n)$, Sam_d^π is not powerful enough to break one-way permutations. In particular,

Theorem 8.2 ([HHRS07]). *For any efficient, oracle-aided algorithm A, there exists a negligible function μ such that*

$$\mathbf{Pr}\left[A^{\pi,\mathrm{Sam}^\pi_{o(n/\log n)}}(y) = \pi_n^{-1}(y)\right] \le \mu(n),$$

where the probability is over the randomness of A and Sam, the choice of a random permutation family $\pi = \{\pi_n \colon \{0,1\}^n \to \{0,1\}^n\}_{n\in\mathbb{N}}$ and the choice of $y \leftarrow \{0,1\}^n$.

Following [PV10], we define the class $\mathsf{RP}\text{-}\mathsf{CF}_d/\mathrm{poly} = \mathsf{RP}^{\mathrm{Sam}_d^\pi}/\mathrm{poly}$ to be the class of languages L with non-uniform efficient oracle aided TM M such that

- $\forall x \in L$, $\mathbf{Pr}_\pi\left[M^{\pi,\mathrm{Sam}_d^\pi}(x) = 1\right] \ge 1/3$
- $\forall x \notin L$, $\mathbf{Pr}_\pi\left[M^{\pi,\mathrm{Sam}_d^\pi}(x) = 1\right] = 0$

We note that our definition of $\mathsf{RP}\text{-}\mathsf{CF}_d/\mathrm{poly}$ is different from the class CF_d defined in [PV10], in few aspects. First we allow the algorithm to be non uniform. Second, our algorithm has one-sided error, instead of two-sided error.[19] See [PV10] for a discussion on the power of the class CF_d.

8.2 Oracle-Aided Interactive Arguments and SWI

We will next define $\mathcal{O}$-relativized interactive arguments. Below we will consider $\mathcal{O}$ which is the set of all permutations families $\pi = \{\pi_n \colon \{0,1\}^n \to \{0,1\}^n\}_{n\in\mathbb{N}}$.

Definition 8.3 ($\mathcal{O}$-Relativized Interactive Argument). *An $\mathcal{O}$-relativized interactive argument (IA) for a relation $\mathcal{R} \subseteq \{0,1\}^* \times \{0,1\}^*$ is a pair of oracle-aided PPT interactive machines (P, V) such that:*

- ***Completeness:*** *For every $(x, w) \in \mathcal{R}$,*

$$\mathop{\mathbf{Pr}}_{O\leftarrow\mathcal{O}}\left[\mathsf{Out}(P^O(x,w), V^O(x)) = 1\right] \ge 1 - \alpha(|x|).$$

- ***S-relativized Computational Knowledge Soundness:*** *There exists a polynomial-time oracle machine E such that for every x and every non-uniform efficient prover $\widehat{P}$:*

$$\mathop{\mathbf{Pr}}_{O\leftarrow\mathcal{O}}\left[\mathsf{Out}(\widehat{P}^{O,S}, V^O(x)) = 1\right] > \eta(|x|) \quad \Longrightarrow \quad \mathop{\mathbf{Pr}}_{O\leftarrow\mathcal{O}}\left[(x, E^{O,\widehat{P}^{O,S^O}}(x)) \in \mathcal{R}\right] \ge 1 - \mu(|x|).$$

Whenever α, η and μ are not explicitly specified, they are assumed to be negligible functions in $|x|$.[20]

[19] We also chose the arbitrary probability threshold for Yes-Instances to be 1/3 instead of the more common choice of 1/2. While this constant in our proof can be improved, we preferred a simpler proof.

[20] As in previous sections, our result holds with respect to a wider range of parameters.

We will also consider WI against *unbounded malicious verifiers*, in which we let $\widehat{V}$ in Definition 2.13 to be any (possibly inefficient) algorithm.

Definition 8.4. *An oracle-aided interactive argument (P, V) for a relation $\mathcal{R}$ is $\mathcal{O}$-*relativized unbounded malicious verifier statistical witness indistinguishability *with distinguishing advantage δ (unbounded malicious verifier δ-*SWI*) if for every $O \in \mathcal{O}$, x and every w_0, w_1 with $(x, w_0), (x, w_1) \in \mathcal{R}$, and every verifier $\widehat{V}$:*

$$\mathrm{SD}(\langle P^O(x, w_0), \widehat{V}(x)\rangle, \langle P^O(x, w_1), \widehat{V}(x)\rangle) < \delta(|x|).$$

The protocol is malicious verifier δ-SWI *if the above holds only with respect to efficient, non-uniform verifier $\widehat{V}$. Whenever δ is not explicitly specified, it is assumed to be negligible functions in $|x|$.*

8.3 Separating SWI from One-Way Permutations

We will prove the following theorem.

Theorem 8.5. *Let Π be the set of all permutations families, and assume that for some function $d = o(n/\log n)$ it holds that* NP $\not\subseteq$ RP-CF$_d$/poly. *Then there is no Π-relativized, malicious verifier,* SWI *for* NP *with d rounds and Sam_d-relativized computational-knowledge-soundness such that one of the following holds:*

- *d is a constant and the protocol is unbounded malicious verifier* SWI*, or*
- *the knowledge extractor is weakly non-adaptive.*

The proof of Theorem 8.5 follows the same lines of Corollary 4.2. We will need the following version of Lemma 3.1.

Lemma 8.6. *Let $\mathcal{R}$ be an* NP *relation and (P', V) be an k-round Π-relativized interactive argument for $\mathcal{R} \vee \mathcal{R}$ with completeness error α and (η, μ) Sam_d-relativized computational knowledge soundness with extractor E, and let $P^\pi := P^{\pi, \mathrm{Sam}_d^\pi}$ be an efficient oracle-aided interactive algorithm. If $1 - \alpha > \eta$ and there exist δ and $\varepsilon \in 1/\mathrm{poly}$ so that $2\mu + \delta \leq 1/12 - \varepsilon$ and for every $\pi \in \Pi, x_0, x_1, w_0, w_1$ with $(x_0, w_0) \in \mathcal{R}$ and $(x_1, w_1) \in \mathcal{R}$ it holds that*

$$SD\left(E^{\pi, P^\pi(x_0 \vee x_1, w_0)}(x_0 \vee x_1), E^{\pi, P^\pi(x_0 \vee x_1, w_1)}(x_0 \vee x_1)\right) \leq \delta,$$

then $\mathcal{L}(\mathcal{R}) \in$ RP-CF$_k$/poly.

Proof. We explain how to change the proof of Lemma 3.1. Recall that $\gamma = \frac{2\mu + \delta}{2}$. By a similar argument to the one in the proof of Lemma 3.1, we can derive that for every $x_0, x_1 \in \mathcal{L}_\mathcal{R}$:

$$\mathop{\mathbf{E}}_{\substack{x_0 \leftarrow S \\ r}} \left[\Pr_{\substack{x_1 \leftarrow S \\ \pi}} \left[(x_1, E_r^{\pi, P^\pi(x_{b^*} \vee x_{1-b^*}, w(x_0))}(x_{b^*} \vee x_{1-b^*})) \in \mathcal{R} \right] \right] \geq 1/2 - \gamma.$$

Let

$$A^\pi_{x_0, w(x_0), r, b^*}(x_1) := E_r^{\pi, P^\pi(x_{b^*} \vee x_{1-b^*}, w(x_0))}(x_{b^*} \vee x_{1-b^*}).$$

Then, for every set $S \subseteq \mathcal{L}_n$, there is a choice of x_0 and randomness r for the extractor, such that

$$\Pr_{\substack{x_1 \leftarrow S \\ \pi}} \left[(x_1, A^\pi_{x_0, w(x_0), r, b^*}(x_1)) \in \mathcal{R} \right] \geq 1/2 - \gamma.$$

Since

$$\Pr_{\substack{x_1 \leftarrow S \\ \pi}} \left[(x_1, A^\pi_{x_0, w(x_0), r, b^*}(x_1)) \in \mathcal{R} \right] \leq 1 \cdot \Pr_{x_1 \leftarrow S} \left[\Pr_\pi \left[(x_1, A^\pi_{x_0, w(x_0), r, b^*}(x_1)) \in \mathcal{R} \right] \geq 1/3 \right] + 1/3,$$

we get that

$$\Pr_{x_1 \leftarrow S} \left[\Pr_\pi \left[(x_1, A^\pi_{x_0, w(x_0), r, b^*}(x_1)) \in \mathcal{R} \right] \geq 1/3 \right] \geq 1/6 - \gamma \geq 1/\mathrm{poly}.$$

Namely, for every set $S \subseteq \mathcal{L}_n$, there is a choice of x_0 and randomness r for the extractor, such that A^π finds a witness for a $(1/6-\gamma)$-fraction of the set S, with probability at least $1/3$ over the choice of the oracle π. We proceed as in the proof of Lemma 3.1, where we next find $q = O(n/(1/6-\gamma)) \in \text{poly}$ tuples $(x^1, w^1, r^1, b^1), \ldots, (x^q, w^q, r^q, b^q)$, such that for every $x \in \mathcal{L}_n$, it holds that for some $i \in [q]$, $A^\pi_{x^i, w^i, r^i, b^i}$ finds a witness for x, with probability at least $1/3$ over the choice of π.

The proof now continues as in the proof of Lemma 3.1. □

We will also need the following version of Lemma 4.3 and Lemma 4.5, which consider an inefficient prover, and oracle-aided algorithms.

Lemma 8.7. *Let (P^π, V^π) be a k-round interactive argument with an (possibly* inefficient*) stateless prover for an* NP *relation $\mathcal{R}$ with completeness error α, (η, μ)-computational knowledge soundness against with q-query extractor E^π, and unbounded malicious verifier δ-*SWI*. Then for every x, w_0, w_1 with $(x, w_0) \in \mathcal{R}$ and $(x, w_1) \in \mathcal{R}$ it holds that*

$$\text{SD}\left(E^{\pi, P^\pi(x,w_0)}(x), E^{\pi, P^\pi(x,w_1)}(x)\right) \leq 2kq^{k+1}\delta.$$

Lemma 8.8. *Let E be a weakly non-adaptive extractor that makes q queries. There exists an efficient, non-uniform, verifier $\widehat{V}$ such that for every stateless prover P' and every x, w_0,w_1 with $(x, w_0), (x, w_1) \in \mathcal{R}$:*

$$\text{SD}(E^{\pi, P'^\pi(x,w_0)}(x), E^{\pi, P'^\pi(x,w_1)}(x)) \leq 2q \cdot \text{SD}(\langle P'^\pi(x, w_0), \widehat{V}(x)\rangle, \langle P'^\pi(x, w_1), \widehat{V}(x)\rangle).$$

The proof of Lemmas 8.7 and 8.8 follows by the same proof of Lemmas 4.3 and 4.5, by noticing that the proof is relativized, and that we only use the efficiency of the stateless prover in the proof of Lemma 8.7 to claim that the verifiers constructed in the proof is efficient.

We are now ready to prove Theorem 8.5.

Proof of Theorem 8.5. We start with the first item. The second item follow similarly using Lemma 4.5. Assume towards a contradiction that there exists such a black-box construction (P, V) with d rounds, and let E be its knowledge extractor.

Fix a permutation family π. Let P'^π be the inefficient stateless prover, that before sending each message sample its random tape uniformly at random from the set of tapes that are consistent with the transcript so far. We first observe that using Sam^π_d, we can construct an efficient prover $\widehat{P}^{\pi, \text{Sam}^\pi_d}$ that behaves like P'^π. That is,

$$E^{\pi, P'^\pi(x,w)}(x) \equiv E^{\pi, \widehat{P}^{\pi, \text{Sam}^\pi_d}(x,w)}(x).^{21} \tag{19}$$

Indeed, let M^π be the interactive TM that implements the rule of $P^\pi(x, w)$ in the protocol. Given partial transcript tr, $\widehat{P}^{\pi, \text{Sam}^\pi_d}$ simply query $\text{Sam}^\pi_d(M^\pi, \mathsf{tr})$. By definition the answer is distributed exactly as the answer of P', and the same holds under rewinding.

Next, we apply Lemma 8.7 on (P'^π, V^π). We get that

$$\text{SD}\left(E^{P'^\pi(x,w_0)}(x), E^{P'^\pi(x,w_1)}(x)\right) \leq 2kq^{k+1}\delta,$$

and thus by Eq. (19),

$$\text{SD}\left(E^{\widehat{P}^{\pi, \text{Sam}^\pi_d}(x,w_0)}(x), E^{\widehat{P}^{\pi, \text{Sam}^\pi_d}(x,w_1)}(x)\right) \leq 2kq^{k+1}\delta.$$

Finally, using Lemma 8.6 and the fact that $\widehat{P}^{\pi, \text{Sam}^\pi_d}$ is efficient, we get that $\text{NP} \subseteq \text{RP-CF}_d/\text{poly}$. □

[21] Note that we do not claim that $\widehat{P}$ is by itself stateless.

Acknowledgments

The authors are thankful to Willy Quach for very fruitful discussions. This work was initiated and much of it was conducted while Gal Arnon, Noam Mazor and Jad Silbak were research fellows at the Simons Institute for the Theory of Computing.

Gal Arnon is supported by the European Research Council (ERC) under the EU's Horizon 2020 research and innovation programme (Grant agreement No. 101019547), and Stellar Foundation Grant. Rafael Pass is supported in part by AFOSR Award FA9550-23-1-0387, AFOSR Award FA9550-23-1-0312, AFOSR Award FA9550-24-1-0267, ISF Award 2338/23 and ERC Advanced Grant KolmoCrypt - 101142322.

Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Government, the AFOSR, the European Union or the European Research Council Executive Agency.

References

[AR21] Gal Arnon and Guy N. Rothblum. "On Prover-Efficient Public-Coin Emulation of Interactive Proofs". In: *2nd Conference on Information-Theoretic Cryptography, ITC 2021, July 23-26, 2021, Virtual Conference*. Ed. by Stefano Tessaro. Vol. 199. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021, 3:1–3:15.

[BG92] Mihir Bellare and Oded Goldreich. "On Defining Proofs of Knowledge". In: *Advances in Cryptology - CRYPTO '92, 12th Annual International Cryptology Conference, Santa Barbara, California, USA, August 16-20, 1992, Proceedings*. Ed. by Ernest F. Brickell. Lecture Notes in Computer Science. Springer, 1992, pp. 390–420.

[BH26] Idan Baril and Iftach Haitner. *On the Complexity of Interactive Arguments*. Cryptology ePrint Archive, Paper 2026/272. 2026. URL: https://eprint.iacr.org/2026/272.

[BHT18] Itay Berman, Iftach Haitner, and Aris Tentes. "Coin flipping of any constant bias implies one-way functions". In: *Journal of the ACM (JACM)* 65.3 (2018), pp. 1–95.

[BIKM99] Amos Beimel, Yuval Ishai, Eyal Kushilevitz, and Tal Malkin. "One-way functions are essential for single-server private information retrieval". In: *Proceedings of the thirty-first annual ACM symposium on Theory of computing*. 1999, pp. 89–98.

[BKPRV24] Nir Bitansky, Chethan Kamath, Omer Paneth, Ron D Rothblum, and Prashant Nalini Vasudevan. "Batch proofs are statistically hiding". In: *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*. 2024, pp. 435–443.

[BL04] Boaz Barak and Yehuda Lindell. "Strict Polynomial-Time in Simulation and Extraction". In: *SIAM J. Comput.* 33.4 (2004), pp. 738–818. DOI: 10.1137/S0097539703427975. URL: https://doi.org/10.1137/S0097539703427975.

[BM82] Manuel Blum and Silvio Micali. "How to Generate Cryptographically Strong Sequences of Pseudo Random Bits". In: 1982, pp. 112–117.

[BOV07] Boaz Barak, Shien Jin Ong, and Salil P. Vadhan. "Derandomization in Cryptography". In: *SIAM J. Comput.* 37.2 (2007), pp. 380–400.

[BP15] Nir Bitansky and Omer Paneth. "ZAPs and Non-Interactive Witness Indistinguishability from Indistinguishability Obfuscation". In: *Theory of Cryptography - 12th Theory of Cryptography Conference, TCC 2015, Warsaw, Poland, March 23-25, 2015, Proceedings, Part II*. Ed. by Yevgeniy Dodis and Jesper Buus Nielsen. Vol. 9015. Lecture Notes in Computer Science. Springer, 2015, pp. 401–427.

[Bar01] Boaz Barak. "How to Go Beyond the Black-Box Simulation Barrier". In: *42nd Annual Symposium on Foundations of Computer Science, FOCS 2001, Las Vegas, Nevada, USA, October 14-17, 2001*. IEEE Computer Society, 2001, pp. 106–115.

[Blu83] Manuel Blum. "Coin flipping by telephone a protocol for solving impossible problems". In: *ACM SIGACT News* 15.1 (1983), pp. 23–27.

[CHK25] Suvradip Chakraborty, James Hulett, and Dakshita Khurana. "On Weak NIZKs, One-Way Functions and Amplification". In: *Annual International Cryptology Conference*. Springer. 2025, pp. 580–610.

[CHKT26] Suvradip Chakraborty, James Hulett, Dakshita Khurana, and Kabir Tomer. "Non-Trivial Zero-Knowledge Implies One-Way Functions". In: *arXiv preprint arXiv:2602.17651* (2026).

[CLV26] Rohit Chatterjee, Yunqi Li, and Prashant Nalini Vasudevan. "Weak Zero-Knowledge and One-Way Functions". In: *arXiv preprint arXiv:2602.16156* (2026).

[DH76] Whitfield Diffie and Martin E. Hellman. "New Directions in Cryptography". In: (1976), pp. 644–654.

[DN07] Cynthia Dwork and Moni Naor. "Zaps and Their Applications". In: *SIAM J. Comput.* 36.6 (2007), pp. 1513–1543.

[FLS99] Uriel Feige, Dror Lapidot, and Adi Shamir. "Multiple NonInteractive Zero Knowledge Proofs Under General Assumptions". In: *SIAM J. Comput.* 29.1 (1999), pp. 1–28.

[FS86] Amos Fiat and Adi Shamir. "How to prove yourself: Practical solutions to identification and signature problems". In: *Conference on the theory and application of cryptographic techniques*. Springer. 1986, pp. 186–194.

[FS90] Uriel Feige and Adi Shamir. "Witness Indistinguishable and Witness Hiding Protocols". In: *Proceedings of the 22nd Annual ACM Symposium on Theory of Computing, May 13-17, 1990, Baltimore, Maryland, USA*. Ed. by Harriet Ortiz. ACM, 1990, pp. 416–426.

[GGM86] Oded Goldreich, Shafi Goldwasser, and Silvio Micali. "How to Construct Random Functions". In: (1986), pp. 792–807.

[GM84] Shafi Goldwasser and Silvio Micali. "Probabilistic Encryption". In: (1984), pp. 270–299.

[GMR89] Shafi Goldwasser, Silvio Micali, and Charles Rackoff. "The Knowledge Complexity of Interactive Proof Systems". In: *SIAM J. Comput.* 18.1 (1989), pp. 186–208.

[GO94] Oded Goldreich and Yair Oren. "Definitions and Properties of Zero-Knowledge Proof Systems". In: *J. Cryptol.* 7.1 (1994), pp. 1–32.

[Gol90] Oded Goldreich. "A note on computational indistinguishability". In: *Information Processing Letters* 34.6 (1990), pp. 277–281.

[HHRS07] Iftach Haitner, Jonathan J Hoch, Omer Reingold, and Gil Segev. "Finding collisions in interactive protocols-a tight lower bound on the round complexity of statistically-hiding commitments". In: *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*. IEEE. 2007, pp. 669–679.

[HILL99] Johan Hastad, Russell Impagliazzo, Leonid A. Levin, and Michael Luby. "A pseudorandom generator from any one-way function". In: (1999), pp. 1364–1396.

[HN24] Shuichi Hirahara and Mikito Nanashima. "One-Way Functions and Zero Knowledge". In: *Proceedings of the 56th Annual ACM Symposium on Theory of Computing, STOC 2024, Vancouver, BC, Canada, June 24-28, 2024*. Ed. by Bojan Mohar, Igor Shinkar, and Ryan O'Donnell. ACM, 2024, pp. 1731–1738.

[HO14] Iftach Haitner and Eran Omri. "Coin flipping with constant bias implies one-way functions". In: *SIAM Journal on Computing* 43.2 (2014), pp. 389–409.

[HR07a] Iftach Haitner and Omer Reingold. "A new interactive hashing theorem". In: *Twenty-Second Annual IEEE Conference on Computational Complexity (CCC'07)*. IEEE. 2007, pp. 319–332.

[HR07b] Iftach Haitner and Omer Reingold. "Statistically-hiding commitment from any one-way function". In: *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*. 2007, pp. 1–10.

[IL89] Russell Impagliazzo and Michael Luby. "One-way functions are essential for complexity based cryptography". In: *30th Annual Symposium on Foundations of Computer Science*. IEEE Computer Society. 1989, pp. 230–235.

[KKS18] Yael Tauman Kalai, Dakshita Khurana, and Amit Sahai. "Statistical Witness Indistinguishability (and more) in Two Messages". In: *Advances in Cryptology - EUROCRYPT 2018 - 37th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Tel Aviv, Israel, April 29 - May 3, 2018 Proceedings, Part III*. Ed. by Jesper Buus Nielsen and Vincent Rijmen. Vol. 10822. Lecture Notes in Computer Science. Springer, 2018, pp. 34–65.

[KMNPRY22] Ilan Komargodski, Tal Moran, Moni Naor, Rafael Pass, Alon Rosen, and Eylon Yogev. "One-Way Functions and (Im)perfect Obfuscation". In: *SIAM J. Comput.* 51.6 (2022), pp. 1769–1795.

[KS06] Takeshi Koshiba and Yoshiharu Seri. "Round-efficient one-way permutation based perfectly concealing bit commitment scheme". In: *ECCC TR06-093* (2006).

[LMP24] Yanyi Liu, Noam Mazor, and Rafael Pass. "A note on zero-knowledge for NP and one-way functions". In: *Cryptology ePrint Archive* (2024).

[MPS25] Noam Mazor, Rafael Pass, and Tomer Solomon. "A Meta-Complexity Theoretic Approach to Indistinguishability Obfuscation and Witness Pseudo-Canonicalization". In: *Theory of Cryptography Conference*. Springer. 2025.

[MV24] Changrui Mu and Prashant Nalini Vasudevan. "Instance-hiding interactive proofs". In: *Theory of Cryptography Conference*. Springer. 2024, pp. 3–34.

[NOVY98] Moni Naor, Rafail Ostrovsky, Ramarathnam Venkatesan, and Moti Yung. "Perfect zero-knowledge arguments for NP using any one-way permutation". In: *Journal of Cryptology* 11.2 (1998), pp. 87–108.

[NY89] Moni Naor and Moti Yung. "Universal One-Way Hash Functions and their Cryptographic Applications". In: 1989, pp. 33–43.

[Nao91] Moni Naor. "Bit commitment using pseudorandomness". In: *Journal of cryptology* 4.2 (1991), pp. 151–158.

[OW93] Rafail Ostrovsky and Avi Wigderson. "One-Way Functions are Essential for Non-Trivial Zero-Knowledge". In: *Second Israel Symposium on Theory of Computing Systems, ISTCS 1993, Natanya, Israel, June 7-9, 1993, Proceedings*. IEEE Computer Society, 1993, pp. 3–17.

[Ost91] Rafail Ostrovsky. "One-Way Functions, Hard on Average Problems, and Statistical Zero-Knowledge Proofs." In: *SCT*. 1991, pp. 133–138.

[PV10] Rafael Pass and Muthuramakrishnan Venkitasubramaniam. "Private coins versus public coins in zero-knowledge proof systems". In: *Theory of Cryptography Conference*. Springer. 2010, pp. 588–605.

[PV20] Rafael Pass and Muthuramakrishnan Venkitasubramaniam. "Is it Easier to Prove Theorems that are Guaranteed to be True?" In: *61st IEEE Annual Symposium on Foundations of Computer Science, FOCS 2020, Durham, NC, USA, November 16-19, 2020*. Ed. by Sandy Irani. IEEE, 2020, pp. 1255–1267.

[Rom90] John Rompel. "One-Way Functions are Necessary and Sufficient for Secure Signatures". In: 1990, pp. 387–394.

[Vad06] Salil P Vadhan. "An unconditional study of computational zero knowledge". In: *SIAM Journal on Computing* 36.4 (2006), pp. 1160–1214.

ECCC ISSN 1433-8092

https://eccc.weizmann.ac.il