Unprovable Security of Perfect NIZK and Non-interactive Non-malleable Commitments

Rafael Pass*
Cornell University
rafael@cs.cornell.edu

March 19, 2017

Abstract

We present barriers to provable security of two important cryptographic primitives, *perfect non-interactive zero knowledge (NIZK)* and *non-interactive non-malleable commitments*:

- Black-box reductions cannot be used to demonstrate *adaptive* soundness (i.e., that soundness holds even if the statement to be proven is chosen as a function of the common reference string) of any statistical NIZK for **NP** based on any "standard" intractability assumptions.
- Black-box reductions cannot be used to demonstrate non-malleability of non-interactive, or even 2-message, commitment schemes based on any "standard" intractability assumptions.

We emphasize that the above separations apply even if the construction of the considered primitives makes a *non-black-box* use of the underlying assumption.

As an independent contribution, we suggest a taxonomy of game-based intractability assumptions.

*__Errata:__ *This version is essentially identical to an August 5, 2015 version (unformatted final submission) which (in its formatted form) appears in Computational Complexity, except that I have updated the definition of a security-preserving black-box reduction to restrict the number of queries made by the reduction to be polynomial. As pointed out by Dakshita Khurana and Amit Sahai, one of results (Theorem 5.9) on subexponential-time reductions implictly made this assumption. The only changes are on page 5,15 and 16 and are explictly marked.* A preliminary verion of this paper appeared in TCC'13. Pass is supported in part by a Alfred P. Sloan Fellowship, Microsoft New Faculty Fellowship, NSF Award CNS-1217821, NSF CAREER Award CCF-0746990, NSF Award CCF-1214844, AFOSR YIP Award FA9550-10-1-0093, and DARPA and AFRL under contract FA8750-11-2- 0211. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the US Government.

Contents

1 Introduction

Modern Cryptography relies on the principle that cryptographic schemes are proven secure based on mathematically precise assumptions; these can be *general*—such as the existence of one-way functions—or *specific*—such as the hardness of factoring products of large primes. The security proof is a *reduction* that transforms any attacker A of the scheme into a machine that breaks the underlying assumption (e.g., inverts an alleged one-way function). This endeavor has been extremely successful, and during the past three decades many cryptographic tasks have been put under rigorous treatment and numerous constructions realizing these tasks have been proposed under a number of well-studied complexity-theoretic hardness assumptions.

In this paper, we study two cryptographic primitives—*perfect non-interactive zero-knowledge with adaptive statements* and *non-interactive non-malleable commitments*—for which security proofs based on well-studied intractability assumptions have remained elusive.

Perfect NIZK with Adaptive Inputs. A non-interactive zero-knowledge (NIZK) protocol [BFM88] is a protocol between two parties, a Prover, and a Verifier, through which the Prover can non-interactively (i.e., by sending a single message π) convince the Verifier of the validity of a statement x, only if x is true (this is called the *soundness* property), while at the same time revealing nothing beyond the fact that x is true (this is called the *zero-knowledge* property). To make such constructs possible both parties are additionally assumed to have access to a "Common Reference String" (CRS) that has been ideally sampled according to some distribution. The original definition of [BFM88] only considered *non-adaptive* notions of soundness and zero-knowledge: Roughly speaking, the (non-adaptive) soundness condition requires that for every false statement $x \notin L$, with high probability over the choice of the CRS, any proof π output by a malicious prover will be rejected by the verifier. The (non-adaptive) zero-knowledge property, on the other hand, requires that for every true statement $x \in L$, the joint distribution consisting of the reference string, and an honestly generated proof, can be reconstructed by a simulator. In both of these properties, the statement x is required to be *fixed* before the reference string is known. Feige, Lapidot and Shamir [FLS90] introduced stronger *adaptive* notions of both soundness and zero-knowledge; roughly speaking, here soundness and zero-knowledge should hold even if the statement x is adversarially chosen *as a function of* the reference string.

As with traditional zero-knowledge protocols, NIZKs come in several flavors: *computational NIZK*, *statistical NIZK*, and *perfect NIZK*. In the computational notion, the simulator's output is only required to be computationally indistinguishable from an honestly generated view, whereas in the statistical (resp. perfect) variants, it is required to be statistically close (resp. identical) to an honestly generated view. Computational NIZK satisfying both adaptive zero-knowledge and adaptive soundness were constructed early on based on standard cryptographic intractability assumptions [FLS90, BY96], but constructions of statistical and perfect NIZK were elusive.

Only recently, a breakthrough result by Groth, Ostrovsky and Sahai (GOS) [GOS06] provided a construction of a perfect NIZK for **NP** based on the hardness of a number-theoretic assumption over bilinear groups. Their protocol satisfies the adaptive notion of zero-knowledge; however, it only satisfies the non-adaptive notion of soundness (that is, soundness is no longer guaranteed to hold if the attacker chooses a statement $x \notin L$ as a function of the common reference string). We focus on whether there exists a perfect NIZK for **NP** satisfying both adaptive soundness and adaptive zero-knowledge.

A step towards answering this question appears in the work of Abe and Fehr [AF07], which presented a perfect NIZK for **NP** with both adaptive soundness and adaptive zero-knowledge, using

an "knowledge-extractaction" assumption (simular to the "knowledge-of-exponent" assumption of [Dam91]), as opposed to a "standard" intractability assumption. Abe and Fehr also demonstrate that certain (arguably natural) types of proof techniques—which they refer to as "direct" black-box reductions—cannot be used to prove adaptive soundness of perfect NIZKs for **NP**. Their notion of a "direct" proof, however, is quite restrictive (very roughly speaking, it requires the security reduction to "directly embed" some hard instance into the CRS in a "structure preserving way").[1]

Non-interactive Non-malleable Commitments. Often described as the "digital" analogue of sealed envelopes, commitment schemes enable a *sender* to commit itself to a value while keeping it secret from the *receiver*. This property is called *hiding*. Furthermore, the commitment is *binding* in the sense that when in a later stage the commitment is opened, then it is guaranteed that the "opening" can yield only a single value determined in the committing stage. For many applications, however, the most basic security guarantees of commitments are not sufficient. For instance, the basic definition of commitments does not rule out an attack where an adversary, upon seeing a commitment to a specific value v, is able to commit to a related value (say, $v-1$), even though it does not know the actual value of v. This kind of attack might have devastating consequences if the underlying application relies on the *independence* of committed values (e.g., consider a case in which the commitment scheme is used for securely implementing a bidding mechanism). In order to address the above concerns, Dolev, Dwork and Naor introduced the concept of *non-malleable commitments* [DDN00]. Loosely speaking, a commitment scheme is said to be non-malleable if it is infeasible for an adversary to "maul" a commitment to a value v into a commitment to a related value $\tilde{v}$.

More precisely, we consider a *man-in-the-middle* (MIM) attacker that participates in two concurrent executions of a commitment scheme Π; in the "left" execution it interacts with an honest committer; in the "right" execution it interacts with an honest receiver. Additionally, we assume that the players have n-bit identities (where n is polynomially related to the security parameter), and that the commitment protocol depends only on the identity of the committer; we sometimes refer to this as the identity of the interaction.[2] Intuitively, Π being non-malleable means that if the identity of the right interaction is different than the identity of the left interaction (i.e., A does not use the same identity as the left committer), the value A commits to on the right does not depend on the value it receives a commitment to on the left; this is formalized by requiring that for any two values v_1 and v_2, the value A commits to after receiving left commitments to v_1 or v_2 are indistinguishable.[3]

The first non-malleable commitment protocol was constructed by Dolev, Dwork and Naor [DDN00] in 1991. The security of their protocol relies on the minimal assumption of one-way functions and requires $\Omega(\log n)$ rounds of interaction, where $n \in \mathbb{N}$ is the length of party identities. The round-complexity of non-malleable commitments has since been extensively studied (see e.g., [Bar02,

[1] Among other things, the structure preserving property requires that if the "hard instance" being directly embedded in the CRS is true, the CRS is valid, and if the hard instance is false, then the CRS is "invalid". This property can never hold when considering NIZK in the Uniform Reference String model (since every CRS is valid), and as such their result holds vacuously when considering NIZK in the Uniform Reference String model.

[2] Non-malleable commitments are sometimes also considered in settings where players do not have identities. However, as shown in [PR05b], any non-malleable commitment that handles sufficiently long identities can be turned into a non-malleable commitment without identities (and any non-malleable commitment without identities can trivially be turned into one with identities). Since our goal is to prove lower bounds, we focus on the more general (relaxed) notion of non-malleability with respect to identities.

[3] Note that the value A commits to is not efficiently computable from the transcript of the right interaction; nevertheless, if the commitment is statistically binding, the value is determined with overwhelming probability. Our focus is on non-malleability for statistically-binding commitments (as is typically the case in the literature).

PR05b, PR05a, LPV08, LP09, PW10, Wee10]), leading up to *constant round* protocols based on one-way functions [LP11, Goy11].

However, the question of whether *non-interactive*, or even 2-round, non-malleable commitments exist, is wide open. (We note that in the Common Reference String model, constructions of non-interactive non-malleable commitments are known [CIO98]; we focus on constructions in the plain model, without any set-up.) Some initial progress towards this question can be found in [PPV08] where a construction of non-interactive non-malleable commitments based on a new hardness assumption is given. This assumption, however, has a strong non-malleability flavor; as such, it provides little insight into the question of whether non-malleability can be obtained from a "pure" hardness assumptions (such as e.g., the hardness of factoring).

1.1 Our Results

The main result of this paper is showing that Turing (i.e., black-box) reductions cannot be used to base the security of the above-mentioned primitives on a general class of intractability assumptions.

More precisely, following Naor [Nao03] (see also [DOP05, HH09, RV10, Pas11, GW11]), we model an *intractability assumption* as an arbitrary game between a (potentially) unbounded challenger C, and an attacker A. The attacker A is said to break the challenger C with respect to the threshold t if it can make C output 1 with probability non-negligibly higher than the threshold t. An intractability assumption is defined as a pair (C, t) where C is a challenger and t is a threshold. All traditional cryptographic hardness assumptions (e.g., the hardness of factoring, the hardness of the discrete logarithm problem, the Decisional Diffie-Hellman (DDH) problem, etc.) can be modeled as 2-round challengers C with the threshold t being either 0 (in case of the factoring or discrete logarithm problems) or 1/2 (in case of the Decisional Diffie-Hellman problem).[4] In all these examples, C is polynomial-time; Naor [Nao03] and Gentry and Wichs [GW11] refer to such assumptions as "falsifiable". For generality, we (following [Pas11]) refer to these as "efficient-challenger" assumptions. More generally, we refer to an assumption where the challenger can be implemented in time (resp. circuit size) $T(\cdot)$ as a "$T(\cdot)$-time (resp. size) challenger assumption".

Our first result rules out basing statistical (and thus also perfect) NIZK with adaptive soundness on efficient-challenger (a.k.a falsifiable) assumptions.

Theorem 1.1 (Informally stated)**.** *Assume the existence of one-way functions secure against polynomial-size circuits. Let* Π *be a statistical non-interactive adaptively zero-knowledge argument for an* **NP***-complete language and let* (C, t) *be an efficient-challenger assumption. Assume there exists a polynomial-time (resp., polynomial circuit-size) Turing reduction* R *such that* R^A *breaks the* C *w.r.t. the threshold* t *for every* A *that breaks adaptive soundness of* Π*. Then,* C *can be broken in polynomial-time (resp. by a polynomial-size circuit) with respect to the threshold* t*.*

Moving on to non-interactive non-malleable commitments, we show that if non-malleability of a two-message (and thus also non-interactive) commitment scheme Π can be based on an efficient-challenger (resp., $T(\cdot)$-size) challenger assumption (C, t) using a polynomial-time (resp., circuit-size $T(\cdot)$) security reduction, then C can be broken in polynomial-time (resp., by a poly$(T(\cdot))$-sized circuit).

[4] For instance, for the case of factoring, the challenger C picks two random k-bit primes p an q, and outputs $N = pq$; the attacker A sends back a number p' and C outputs 1 iff $p' \in \{p, q\}$. For the case for DDH, the challenger picks some cyclic group G of order q where $|q| \approx 2^k$ (the assumption is parametrized by the group selecting algorithm), a generator g for G, uniformly selects $a, b, c \in Z_q$, lets $z_0 = g^{ab}$ and $z_1 = g^c$, and outputs a description of G, g, g^a, g^b, z_i where i is a random bit; the attacker is supposed to send back a bit i' and C outputs 1 iff $i' = i$.

Theorem 1.2 (Informally stated)**.** *Let* Π *be a two-message commitment scheme, and let* (C, t) *be an efficient-challenger (resp.,* $T(\cdot)$*-size) assumption. Assume there exists a polynomial-time (resp., circuit-size* $T(\cdot)$*) Turing reduction* R[5] *by such that* R^A *breaks* C *w.r.t. the threshold* t *for every* A *that breaks non-malleability of* Π*. Then* C *can be broken in polynomial-time (resp., by a* $\text{poly}(T(\cdot))$*-sized circuit) with respect to the threshold* t*.*

We emphasize that for all the above-mentioned results, the *construction* of the protocols Π need not make use of the underlying assumption in a black-box way; the only restriction we impose is that the security *reduction* (establishing the security of Π) is a Turing (i.e., black-box) reduction.

Why these primitives? On a very high-level, non-interactive statistical NIZK and two-round non-malleable commitments share three properties that enable our unprovability results: 1) they are both "two-round" primitives (for the case of NIZK, we can view the CRS as a "first message"), 2) whether the primitives get broken cannot be verified efficiently, and 3) they both have a zero-knowledge flavor (explicitly in the case of NIZK, and implicitly for the case of commitments). These are exactly the properties that we need for our unprovability results, and consequently both unprovability results have significant overlaps in terms of the techniques employed. Let us stress that our unprovability results do not apply to *all* primitives satisfying these properties; for instance, computational NIZK also satisfies them. Looking ahead, what is actually needed by our proof is that the primitives satisfy a zero-knowledge property even *conditioned* on the first message of the protocol (for "typical" first messages); statistical NIZK implies this property, but computational NIZK does not.

Dealing with Security Reductions with Non-uniform Advice. In this work we focus on ruling out security reductions that only use the attacker in a black-box way. Black-box reductions are restrictive in two ways: (a) they need to work even if the attacker they operate on is computationally unbounded, and (b) they get no (non-uniform) advice that *depend* on the attacker they operate on. Some quite commonly techniques in cryptography make use of reductions that do not satisfy property (b). For instance, a useful technique (originating in [GO94] and used in the context of non-malleable commitments in [LPV08]) is to use a hybrid argument that involves non-uniformly fixing some "good" prefix (which may not be efficiently computable) of some experiment and providing this prefix and perhaps some additional (not efficiently computable) information to the reduction.[6] We mention that a recent work by Chung, Lin, Mahmoody and Pass [CLMP13] provides techniques for extending certain types of separation results for the black-box setting to deal also with black-box reductions receiving such non-uniform advice. These techniques readily apply also to our results.

Let us stress that we still need to assume that the reduction works even if the attacker it operates on is computationally unbounded. As such, our results do not rule out *arbitrary* non-black-box reductions, such as those introduced by Barak [Bar01].[7]

A Taxonomy of Intractability Assumptions. As an independent contribution, we slightly generalize the notion of an intractability assumption from [Pas11] (see also [Nao03, DOP05, HH09,

[5] **Added on March 2017**: we here refer to $T(\cdot)$-size reductions that still only make polynomially many queries to their oracle.

[6] In fact, as we shall see later on, we also make use of this proof technique in this paper.

[7] Such arbitrary non-black-box techniques, however, are significantly less common and, as far as we know, no "practical" protocols have been analyzed using them. Furthemore, such techniques have only been successfully used to analyze cryptographic protocols with many communication rounds.

RV10, GW11]) and provide a natural taxonomy of intractability assumptions based on 1) the *security threshold*, 2) the number of *communication rounds* in the security game, 3) the *computational complexity* of the game challenger, 4) the *communication complexity* of the challenger. Our results, combined with [Pas11, GW11], demonstrate several natural primitives that may be (trivially) based on an assumption of a certain type (e.g., the soundness condition of a perfect NIZK can be viewed as a bounded-round assumption), but cannot be based on a different type of assumption (e.g., an assumption where the challenger is efficient). Our results focus on understanding limitations in terms of items 1, 2 and 3; we leave open an exploration of item 4 (i.e., the communication complexity of the challenger). More generally, we are optimistic that cryptographic tasks may be classified in this taxonomy, based on whether they can be achieved—even using a *non-black-box construction*—based on a class of assumptions in this taxonomy, but not on another.

A Note on Random Oracles. Let us point out that in the Random Oracle model [BR93], both of the above-mentioned primitives are easy to construct. For Perfect NIZK this was done in [BR93] (by relying on the "Fiat-Shamir heuristic" [FS87] which is sound in this model) and for non-interactive non-malleable commitments in [Pas03]. Indeed, many practical protocols rely on the assumption that a "good" hash function behaves like a non-interactive non-malleable commitment, and on non-interactive zero-knowledge arguments constructed by applying the "Fiat-Shamir heuristic" [FS87] to a three-message perfect zero-knowledge protocol. Our results show that such commonly used sub-protocols cannot be proven secure based on standard hardness assumptions. Note that these results are incomparable to those of e.g., [CGH04, GK03] on the "uninstantiability of random oracles". On the one hand, the results of [CGH04, GK03] are stronger in the sense that any instantiation of their scheme with a concrete function can actually be *broken*, whereas we just show that the instantiated scheme cannot be *proven secure* using a Turing reduction based on standard assumptions. On the other hand, the separations of [CGH04, GK03] consider "artificial protocols", whereas the protocols we consider are natural (and commonly used in practice).

1.2 Related Separation Results

There is a large literature on separation results between cryptographic primitives and/or assumptions. We distinguish between two types of results.

Separations for fully black-box constructions. The seminal work of Impagliazzo and Rudich [IR88] provides a framework for proving black-box separations between cryptographic primitives. We stress that this framework considers so-called "fully-black-box constructions" (see [RTV04] for a taxonomy of various black-box separations); that is, the framework considers both black-box *constructions* (i.e., the higher-level primitive only uses the underlying primitive as a black-box), and black-box *reductions*.

Separations for black-box reductions. In recent years, new types of black-box separations have emerged. These types of separation apply even to non-black-box constructions, but still only rule out black-box proofs of security: Pass [Pas06] and Pass, Tseng and Venkitasubramaniam [PTV11] (relying on the works of Brassard [Bra83] and Akavia et al [AGGM06], demonstrating limitations of "NP-hard Cryptography"[8]) demonstrate that under certain (new) complexity theoretic assumptions, various cryptographic task cannot be based on *one-way functions* using a

[8]See also the results of Feigenbaum and Fortnow [FF93] and the result of Bogdanov and Trevisan [BT03] that demonstrate limitations of NP-hard cryptography for *restricted* types of reductions.

black-box security reduction, even if the protocol uses the one-way function in a non-black-box way. Very recently, two independent works demonstrate similar types of separation bounds, but this time ruling our security reductions to a *general* set of intractability assumptions: Pass [Pas11] demonstrates impossibility of using black-box reductions to prove the security of several primitives (e.g., Schnorr's identification scheme, commitment scheme secure under weak notions of selective opening, Chaum Blind signatures, etc) based on any "bounded-round" intractability assumption (where the challenger uses an a-priori bounded number of rounds, but is otherwise computationally unbounded). Gentry and Wichs [GW11] demonstrate (assuming the existence of strong pseudorandom generators) impossibility of using black-box security reductions to prove soundness of "succinct non-interactive arguments" based on any "falsifiable" assumption (where the challenger is computationally bounded). Both of the above-mentioned work fall into the "meta-reduction" paradigm of Boneh and Venkatesan [BV98], which was earlier used to prove separations for *restricted* types of reductions (see e.g., [BMV08, HH09, HRS09, FS10]).[9] Our separation results are in the vein of these two works, and follows some of their techniques.

1.3 Proof Overview: Ruling out Perfect NIZK with Adaptive Inputs

Following is an overview of the proof of Theorem 1.1 restricted to the case of *perfect* (as opposed to statistical) NIZK. Assume there exists a perfect NIZK (P, V) for a *hard-on-the average* language $L \in \mathbf{NP}$; for simplicity, let us further assume that there exists efficient algorithms Sam_L and $Sam_{\bar{L}}$ such that a) with overwhelming probability, Sam_L samples instance $x \in L \cap \{0,1\}^n$ along with a witness $w \in R_L(x)$, b) $Sam_{\bar{L}}$ samples instances $x \in \{0,1\}^n \setminus L)$, and c) instances x sampled by Sam_L and $Sam_{\bar{L}}$ are indistinguishable by polynomial-size circuits. Such an L exists based on the existence of one-way functions (that are secure against polynomial-size circuits), which exist by hypothesis. In fact, any **NP** complete language has this property assuming the existence of one-way functions.

For simplicity, in this proof overview we further restrict ourselves to the case when the reference string is uniformly random (i.e., we consider only Perfect NIZK in the so-called Uniform Reference String (URS) Model). Assume, now, that there exists a Turing reduction R such that R^A breaks the assumption C (with respect to some thresholds t) whenever A breaks adaptive soundness of (P, V). Following the "meta-reduction" paradigm by Boneh and Venkatesan [BV98] (which is used in both [Pas11] and [GW11], and also [AF07]), we want to use R to directly break C.

More precisely (just as in [Pas11, GW11]) we exhibit a particular attacker A to the adaptive soundness of (P, V) and next show how to "emulate" this attacker for R without disturbing R's interaction with C (recall that C may be interactive). Whereas in [Pas11] the emulation was statistically close (and thus the separation could be applied also to unbounded challengers), in [GW11] the emulation was only *computationally indistinguishable*, but this still suffices for convincing C as long as C is computationally efficient. We here follow the approach of [GW11].

Let us turn to describing our attacker A, and next explain how to emulate it. Given a CRS ρ, attacker A first attempts to recover random coins r that may be used by the simulator S to output the CRS ρ; since the simulation is perfect, such a string r exists (but finding r might require super-polynomial time and so A is not necessarily polynomial-time). (Recall that since we are dealing with adaptive zero-knowledge, the zero-knowledge simulator needs to output a reference string ρ before knowing what statement it needs to simulate a proof of.) Next, A samples a false instance

[9] For instance, these separations results restrict to "algebraic" reductions, or reductions that run the attacker in a "straigth-line" fashion.

$x \notin L$ that is indistinguishable from a true instance (by hypothesis, this can be done efficiently).[10] Finally, it runs the simulator S on the random coins r to generate ρ, and next feeds it the instance x, and lets π denote the proof output by S (again this final step is efficient).

Let us argue that the proof π of x is accepted by $V(\rho)$. Towards this, consider a hybrid attacker A' that performs exactly the same steps as A, except that it samples a *true* instance $x \in L$. It follows from the ZK property (combined with the completeness property) that V accepts the proofs output by A'. Now, intuitively, it should follow from the hard-on-the-average property of L that V also accepts the proofs output by A. But there is a problem: Recall that A is *not (necessarily) efficient*. However, since it is only the first step of A that is inefficient, we can fix the random string r non-uniformly and still use the remaining steps of A and the efficient verifier V to contradict the hard-on-average property of L, as long as we assume that L is hard-on-average for non-uniform polynomial-time. Note that we here rely on the fact that A is allowed to choose the statement x *after* having seen the reference string ρ (i.e., we rely on A breaking *adaptive* soundness)—this is what allows us to non-uniformly choose r as a function of ρ, *before* sampling $x \in L$.

Now, given this breaker A, let us see an efficient attacker $\tilde{A}$ that is computationally indistinguishable from A. On inputs a CRS ρ, $\tilde{A}(\rho)$ simply samples a random true statement x together with a witness w, and next runs the honest prover strategy $P(\rho, x, w)$ to produce a proof π (this strategy is similar to the one used in [GW11]). It follows by the perfect ZK property that the outputs of C when communicating with $R^{\tilde{A}}$ and $R^{A'}$ are indistinguishable. Let us here point out that we cruicially rely on the *perfect* ZK property (in the URS model): this is because R may query its oracle on *any* CRS of its choice—in particular, the CRS may not be random and thus "standard" (computational) ZK would not suffice. However, if the ZK simulation is perfect, it follows that the simualtion is perfect for *every* CRS (of the right length). We can finally apply a similar argument as above to argue that the outputs of C when communicating with R^A and $R^{A'}$ are indistinguishable, and thus $R^{\tilde{A}}$ breaks C with roughly the same probability as R^A does. But this means that C can be broken in probabilistic polynomial time (since both R and $\tilde{A}$ are probabilistic polynomial-time machines).

1.4 Proof Overview: Ruling out Non-interactive Non-malleable Commitments

Following is an overview of the proof of Theorem 1.2. Assume there exists a non-interactive commitment scheme Π (for simplicity of exposition we focus only on non-interactive, as opposed to two-message, commitments). Assume, further, that there exists a Turing reduction R such that R^A breaks the assumption C (with respect to some thresholds t) whenever A breaks non-malleability of Π. Recall that an attacker A that breaks non-malleability of Π participates in two interactions—one on the "left" acting as a receiver, and one on the "right" acting as a committer. To be successful A needs to choose a different identity for the left and right interactions, and must commit to a value $\tilde{v}$ that is related to the value v it receives a commitment to on the left. Consider a strong attacker A that chooses identity 0 on the left, and identity 1 on the right, and upon receiving a commitment c recovers (using brute force) the unique value v that c is a commitment to (if the value is not unique v is set to $\perp$), and next honestly commits to v on the right. Clearly A breaks the non-malleability of Π (since the identity on the right is different from the identity on the left) and thus R^A also breaks C w.r.t. t.

Let us now see how to efficiently "emulate" A. We simply consider a "trivial" adversary $\tilde{A}$

[10] The careful reader may notice that A actually does not choose the statement x *adaptively*. The fact that the reduction needs to work for attackers A that *may* choose the statement adaptively, and as a consequence must output the reference string ρ before A gets to pick the statement, suffices for us.

that picks identity 0 on the left and 1 on the right (just as A), but instead of trying to commit to v on the right, it simply commits to 0 on the right. Now, intuitively, if the reduction R and the challenger C are polynomial-time, then it should follow by the hiding property of Π that $R^{\tilde{A}}$ still breaks C (w.r.t. t). Note, however, that R may be asking its oracle to break non-malleability of *multiple* commitments (rather than a single one as we implicitly assumed above), and since A is not efficiently computable, we need to be a bit careful when doing the hybrid argument. Nevertheless, using a careful ordering of the hybrids, and relying on the fact that the hiding property of a commitment scheme holds w.r.t. *non-uniform* PPT algorithms, we can show that $R^{\tilde{A}}$ still breaks C (w.r.t. t).

The case of super polynomial-time reductions. Note that the above proof idea applies to a very weak notion of "one-sided" non-malleability, where the attacker always uses identity 0 on the left and identity 1 on the right; Liskov et al [LLM+01] call commitments satisfying this weak notion of non-malleability, *mutually independent*. Interestingly, [LLM+01] shows a construction of a mutually independent commitment based on the existence of subexponentially-hard one-way permutations. The idea (a.k.a. "complexity-leveraging" [CGGM00]) is simple: Let Com_0 be a commitment scheme that is hiding for subexponential time, and let Com_1 be a (polynomial-time) secure commitment scheme whose hiding property can be fully broken (i.e., the commited value can be recovered) in subexponential time. A committer with identity $b \in \{0,1\}$ shall use Com_b. Now, if a MIM upon receiving a commitment of v using Com_0 is able to output a commitment to a related value $\tilde{v}$ using Com_1, then we can violate the hiding of Com_0 by simply breaking Com_1 by brute-force. This security reduction, however, is super-polynomial (subexponential) time. A natural question is whether super-polynomial time/size reductions may be helpful for constructing "full-fledged" (as opposed to one-sided) non-interactive commitments.[11] In Theorem 5.9, we rule out such reductions (or rather to show that if there exists such a reduction, then the reduction itself must already break the assumption, with quadratic overhead). We proceed to provide a (very high-level) proof idea.

Consider a $T(k)$-sized reduction R, where $T(k)$ is super-polynomial, for basing non-malleability on an efficient-challenger assumption C, and consider the algorithms A and $\tilde{A}$ described above.[12] Note that if R has super-polynomial size, we have no guarantees that $R^{\tilde{A}}$ breaks C even when R^A does; indeed, since hiding of Π is only required to hold for polynomial-sized algorithms, $R^{\tilde{A}}$'s success probability may be very different from R^A's success probability. We now show that if this is the case, we can *leverage* R as a commitment distinguisher for commitments using identity 1: intuitively, in case $R^{\tilde{A}}$ does not break C, then R (combined with C) must be able to break the hiding of commitments using identity 1 (since the only difference between the executions of $R^{\tilde{A}}$ and R^A is the values R receives commitments to using identity 1).

So, roughly, if $R^{\tilde{A}}$ does not already break C, we can use R (in conjunction with C) to obtain a circuit D that distinguishes, say, commitments to 0^k and 1^k *using identity 1*.[13] We may then use D to construct a *different* man-in-the-middle attacker A' that uses "switched" identities: it chooses identity 1 on the *left* and 0 on the right (as opposed to 0 on the left and 1 on the right,

[11] Indeed, [PW10] rely on intuitions similar to those from mutually independent commitments to construct a "full-fledged" non-malleable commitment, but this construction requires multiple communication rounds.

[12] The assumption that C is an efficient challenger is only made here to simplify exposition; our actual proof also works when C is $T(k)$-sized.

[13] As in the previous proof, to obtain a machine that breaks the hiding of the commitment, we need to rely a polynomial-length non-uniform advice to deal with the above-mentioned inefficiency issue in the hybrid argument; this is why we work with circuits here.

as A and $\tilde{A}$ did) and upon receiving a commitment on the left, it uses D to get a "guess" for the value of the commitment (recall that by assumption D distinguishes commitments to 0^k and 1^k using identity 1), and then commits to the guess on the right (using identity 0). (Note that the reason we require A' to use "switched" identities, is that we want to leverage the fact that R could distinguish commitments given by A and $\tilde{A}$, and those commitments are given using identity 1.)

We can finally use R combined with A' to directly break C. So, summarizing, either $R^{\tilde{A}}$ works, or else, we use R in order to construct an MIM A' that breaks non-malleability, and then use $R^{A'}$ to break C—in essence, we use R "on itself" to break C.

1.5 Overview of the Paper

We provide some preliminaries and standard definitions in Section 2. Definitions of intractability assumptions and black-box reductions can be found in Section 3; this section also contains our taxonomy of intractability assumptions. We formally state and prove our results about NIZK in Section 4, and we finally formally state and prove our results about non-malleable commitments in Section 5.

2 Preliminaries

2.1 Notation

Integer, Strings and Vectors. We denote by $\mathbb{N}$ the set of natural numbers. Unless otherwise specified, when given as an input to an algorithm, a natural number is presented in its binary expansion (with no *leading* 0s). For $n \in \mathbb{N}$, we denote by 1^n the unary expansion of n (i.e., the concatenation of n 1's).

Probabilistic notation. We employ the following probabilistic notation from [GMR88]. We focus on probability distributions $X : S \to R^+$ over finite sets S.

Probabilistic assignments. If D is a probability distribution then "$x \leftarrow D$" denotes the elementary procedure consisting of choosing an element x at random according to D and returning x. If F is a finite set, then the notation "$x \leftarrow F$" denotes the act of choosing x uniformly from F.

Probabilistic experiments. Let p be a predicate and $D_1, D_2, \ldots$ probability distributions, then the notation $\Pr[x_1 \leftarrow D_1;\ x_2 \leftarrow D_2;\ \ldots\ :\ p(x_1, x_2, \ldots)]$ denotes the probability that $p(x_1, x_2, \ldots)$ holds after the ordered execution of the probabilistic assignments $x_1 \leftarrow D_1;\ x_2 \leftarrow D_2;\ \ldots$

New probability distributions. If $D_1,\ D_2,\ \ldots$ are probability distributions, the notation $\{x \leftarrow D_1; y \leftarrow D_2; \cdots : (x, y, \cdots)\}$ denotes the new probability distribution over $\{(x, y, \cdots)\}$ generated by the ordered execution of the probabilistic assignments $x \leftarrow D_1$, $y \leftarrow D_2, \cdots$.

Probability ensembles. A *probability ensemble* is an infinite sequence of random variables $X = \{X_n\}_{n\in\mathbb{N}}$. We will consider ensembles of the form $X = \{X_n\}_{n\in\mathbb{N}}$ where X_n ranges over strings of length $p(k)$, for some fixed, positive polynomial p.

In order to simplify notation, we sometimes abuse notation and employ the following "shortcut": Given a probability distribution X, we let X denote the random variable obtained by selecting $x \leftarrow X$ and outputting x.

Algorithms. We employ the following notation for algorithms.

Probabilistic algorithms. By a probabilistic algorithm we mean a Turing machine that receives an auxiliary random tape as input. If M is a probabilistic algorithm, then for any input x, the notation "$M_r(x)$" denotes the output of the M on input x when receiving r as random tape. We let the notation "$M(x)$" denote the probability distribution over the outputs of M on input x where each bit of the random tape r is selected at random and independently, and then outputting $M_r(x)$.

Interactive Algorithms. We assume familiarity with the basic notions of an *Interactive Turing Machine* [GMR89] (ITM for brevity) and a *protocol*. (Briefly, a protocol is pair of ITMs computing in turns. In each turn, called a round, only one ITM is active. A round ends with the active machine either halting—in which case the protocol halts—or by sending a message m to the other machine, which becomes active with m as a special input. By an interactive algorithm we mean a (probabilistic) interactive Turing Machine.

Given a pair of interactive algorithms (A, B), we let $\langle A(a), B(b)\rangle(x)$ denote the probability distribution over the outputs of $B(b)$ after interacting with $A(a)$ on the common input x.

Oracle algorithms. An oracle algorithm is a machine that gets oracle access to another machine. Given a probabilistic oracle algorithm M and a probabilistic algorithm A, we let $M^A(x)$ denote the probability distribution over the outputs of the oracle algorithm M on input x, when given oracle access to A. We emphasize that if the algorithm A is probabilistic, M does not get to control the randomness of A, and each time A is invoked fresh randomness is used by A. The fact that we do not allow black-box reductions to access the randomness of the oracle they communicate with may seem restrictive. As we shall see later on, however, our result apply even if we consider reductions that work only for *deterministic* attackers (using a technique from Goldreich and Krawczyk [GK96]); see Section 4.3.1 for more details.

Negligible functions. The term "negligible" is used for denoting functions that are asymptotically smaller than the inverse of any positive polynomial. More precisely, a function $\mu(\cdot)$ from non-negative integers to reals is called *negligible* if for every constant $c > 0$ and all sufficiently large k, it holds that $\mu(k) < k^{-c}$.

2.2 Indistinguishability

The following definition of (computational) indistinguishability originates in the seminal paper of Goldwasser and Micali [GM84].

Let X be a set of strings. A *probability ensemble indexed by* X is a sequence of random variables indexed by X. Namely, any element of $A = \{A_x\}_{x\in X}$ is a random variable indexed by X.

Definition 2.1 (Indistinguishability)**.** *Let $X \subseteq \{0,1\}^*$. Two ensembles $\{A_{n,x}\}_{n\in\mathbb{N},x\in X}$ and $\{B_{n,x}\}_{n\in\mathbb{N},x\in X}$ are said to be* computationally indistinguishable*, if for every probabilistic machine D (the "distinguisher") whose running time is polynomial in the length of its first input, there exists a negligible function $\mu(\cdot)$ so that for every $n \in \mathbb{N}, x \in X$:*

$$|\Pr[D(1^n, x, A_{n,x}) = 1] - \Pr[D(1^n, x, B_{n,x}) = 1]| < \mu(n)$$

$\{A_{n,x}\}_{n\in\mathbb{N},x\in X}$ and $\{B_{n,x}\}_{n\in\mathbb{N},x\in X}$ are said to be statistically indistinguishable *over X if the above condition holds for all (possibly computationally unbounded) machines D.*

We stress that (as is usually the case) the distinguisher D gets as input the index of the ensemble.

2.3 Witness Relations

We recall the standard definition of a witness relation for an $\mathcal{NP}$ language.

Definition 2.2 (Witness relation)**.** *A* witness relation *for a language $L \in \mathcal{NP}$ is a binary relation R_L that is polynomially bounded, polynomial time recognizable and characterizes L by $L = \{x : \exists w \, s.t. \, (x, w) \in R_L\}$.*

We say that w is a *witness* for the membership $x \in L$ if $(x, w) \in R_L$. We will also let $R_L(x)$ denote the set of witnesses for the membership $x \in L$; i.e., $R_L(x) = \{w : (x, w) \in L\}$.

3 Intractability Assumptions and Black-box Reductions

Our definition of an intractability assumption closely follows [Pas11]. Following Naor [Nao03] (see also [DOP05, HH09, RV10]), we model an intractability assumption as an interaction (or game) between a probabilistic machine C—called the challenger—and an attacker A. Both parties get as input 1^k where k is the security parameter. Any such challenger C, together with a threshold function $t(\cdot)$ intuitively corresponds to the assumption:

> *For every polynomial-time adversary A, there exists a negligible function μ such that for all $k \in \mathbb{N}$, the probability that C outputs 1 after interacting with A is bounded by $t(k) + \mu(k)$.*

Hence, we say that *A breaks C w.r.t. threshold t with probability p on common input* 1^k if

$$\Pr\left[\langle A, C\rangle(1^k) = 1\right] \geq t(k) + p(k).$$

Computational efficiency of C. If the challenger C is polynomial-time in the length of the messages it receives, we say that the assumption is *efficient challenger*; such assumptions are referred to as *falsifiable* assumptions by Naor [Nao03] and Gentry and Wichs [GW11]. More generally, we refer to an assumption as having a $T(\cdot, \cdot)$-time (resp., circuit-size) challenger if C can be implemented in time (resp., circuit-size) $T(k, \ell)$ on input the security parameter 1^k, and when receiving messages of length ℓ. Hence, (C, t) is an efficient-challenger assumption if C is a $T(\cdot, \cdot)$-assumption where $T(k, \ell)$ is polynomial in both k and ℓ. For simplicity, here we consider either $\text{poly}(k, \ell)$-time (or circuit-size) challengers, or $T(k, \ell) = T(k)$-time (or circuit-size) challengers, where the running-time of the challenger is bounded only as a function of the security parameter.

For instance, the assumption that a particular function f is (strongly) one-way corresponds to the threshold $t(k) = 0$ and the 2-round challenger C that on input 1^k pick a random input x of length k, sends $f(x)$ to the attacker, and finally outputs 1 iff the attacker returns an inverse to $f(x)$. More esoteric assumptions such as the "one-more discrete logarithm assumption" [BNPS03, BP02], or "adaptive one-way functions" [PPV08], are not efficient-challenger assumptions: for these notions of one-way functions, the attacker gets restricted access to some (restricted) inversion oracle that cannot be implemented in polynomial-time. For instance, adaptive one-wayness of a function f requires that inverting $y = f(x)$ for a random x is hard even if the attacker has access to an oracle that inverts f on any point $y' \neq y$. Such assumptions, however, can be modeled as *exponential-time* challenger assumptions—the challenger can now implement the oracle for the attacker.

The security threshold t. Indistinguishability assumptions (such as, e.g., the Decisional Diffie-Hellman (DDH) problem, or the assumption that a particular function g is a pseudorandom generator) can also be modelled as 2-round challengers but now we have the threshold $t(k) = 1/2$. For instance, the assumption that a length-doubling function g is a pseudorandom generator corresponds to the 2-round challenger C that on input 1^k proceeds as follows: 1) pick a random input x of length k, and let $y_0 = g(x)$, 2) pick a random input y_1 of length $2k$, 3) send y_b to the attacker where b is a random bit, and 4) finally output 1 iff the attacker returns a bit b' such that $b' = b$. Note that any threshold t assumption can always be turned into a threshold $t + \delta$ assumption by having the challenger simply "accept" with probability δ, and otherwise proceed as before. The other direction is less clear.[14]

Round/Communication Efficiency of C. Note the above-mentioned adaptive one-way function assumption requires the challenger C not only to be inefficient, but also to have an a-priori *unbounded number of communication rounds* (as C must answer any number of, potentially adaptively selected, inversion queries). We may also consider restricted types of intractability assumptions which bound the round complexity of C. For instance, in [Pas11] we considered challengers C that are computationally unbounded, but for which there exists a polynomial upper bound (in the terms of the security parameter k) on the number of communications rounds by C; we refer to these assumptions as *bounded-round* intractability assumptions. Another interesting class of assumptions is obtained by further restricting the *communication complexity* of C; for instance, we may require that there is a polynomial bound (again in terms of the security parameter k) on the total communication of C; we refer to these assumptions as *bounded-communication intractability assumption.*

A Taxonomy of Intractability Assumptions. The above way of modeling assumptions provides a natural taxonomy of intractability assumptions based on 1) the *security threshold* t, 2) the number of *communication rounds* used by C, 3) the *computational complexity* of C, and 4) the *communication complexity* of C.

Let us also note that "knowledge-extraction" assumptions (similar to the "knowledge-of-exponent" assumption of [Dam91]) do not fit within our taxonomy of intractability assumptions. In our opinion, such assumptions are perhaps better thought of as *tractability* assumptions. Roughly speaking, such assumptions stipulate *feasibility* of efficiently performing some particular task (namely "extraction" of some inputs from every machine that wins in some game).

Classifying Cryptographic Tasks. As mentioned above, the assumption that a particular function is a one-way function can be formalized as a 2-round efficient-challenger assumption; so can the DDH assumption. But also security properties of more elaborate *cryptographic tasks* can be formalized as intractability assumptions of the above kind:[15]

[14] As noted in [HH09], in some cases (and in particular the DDH assumption), we can use "Chernoff-type parallel repetition theorems" [IJK07, HPWP10] to redue a threshold 1/2 assumption to a threshold 0 assumption; using such parallel repetition, however, increases the communication complexity. In particular, even if the original threshold 1/2 assumption has a-priori polynomially-bounded communication complexity, the new assumption has not. (Roughly, this is because we need to let the attacker in the new threshold 0 assumption to dictate a lower bound number of parallel repetitions to ensure that "completness" holds).

[15] Specifically, this refers to all cryptographic task with "game-based" definition of security (e.g., one-way functions, signatures etc), as opposed to simulation-based definitions of security (e.g. zero-knowledge).

- For instance, the notion of "IND-CPA security" of an encryption scheme [GM84] can be formalized as a 4-round efficient-challenger, bounded-communication, assumption using the standard CPA security game. (Recall that the classic notion of Chosen Plain-text Attack (CPA) security considers an attacker that first receives a randomly sampled public key, next picks two messages m_1 and m_2, then receives back an randomly generated encryption c of a randomly selected choice of these messages m_b and wins if it manages to guess the bit b.)

- On the other hand, the security game of a signature scheme [GMR88] requires using an unbounded-round (and thus also communication) challenger, but still has an efficient challenger. (Recall that the standard notion of security for signatures schemes considers an attacker that receives the verification key for a signature scheme, and then gets oracle access to a "signing" oracle that signs any message of the attacker's choice; the attacker finally wins the game if it manages to come up with a valid signature on a "new" message m on which it has not queries its signature oracle. To model this as an assumption (C, t), we simply have the challenger C generate random verification and signing keys, and then use the signing key to implement the signing oracle for the attacker.[16])

- We also note that non-malleability of a two-round commitment and adaptive soundness of non-interactive proofs can be formalized as bounded-round, bounded-communication assumptions, but they require an inefficient challenger (to check whether the attack was successful). For instance, the assumption that a non-interactive proof is adaptively sound for **NP** statements of length $q(k) = k$ can be stated as an exponential-time challenger assumption: The challenger first sends a CRS to the attacker; upon receiving back a statement x and proof π, the challenger ouputs 1 iff π is an accepting proof of x (which can be efficiency checked) and $x \in \{0,1\}^k$ is a false statement (which can be checked in time $2^{\text{poly}(k)}$).

We refer to the intractability assumption associated with the (game-based) security definition of an instantiation of a cryptographic task as the *trivial* intractability assumption on which it can be based (e.g., the security of a particular signature scheme can be based on the intractability assumption that the scheme is secure.) Note, however, that not all cryptographic tasks have even a trivial intractability assumption on which they can be based (e.g., it is not clear whether the zero-knowledge property of a protocol can be formalized as a game-based security property).

Many major results in the cryptographic literature demonstate "jumps" in our taxonomy: we base the security of some cryptographic tasks on an intractability assumption of a more restrictive nature. We believe that these "jumps" demonstrate the fundamental nature of these results. For instance,

- When we construct a pseudorandom generator from a particular one-way permutation [BM84, GL89] or even a one-way function [HILL99], we base a primitive (the pseudorandom generator) whose trivial associated intractability has 2-rounds, bounded-communication, efficient challenger, and *threshold* of 1/2, on an intractability assumption that is 2-round, bounded-communication, efficient-challenger, but threshold 0.

- When we base the security of a signature scheme on one-way functions [NY89, Rom90], or when we based pseudorandom functions on one-way functions [GGM86, HILL99], we base primitives whose associated trivial intractability assumption are *unbounded-round*, on a 2-round, bounded-communication, efficient-challenger assumption.

[16]The notion of "IND-CCA" security for encryption schemes can be modeled in a similar fashion using an unbounded-round efficient challenger C.

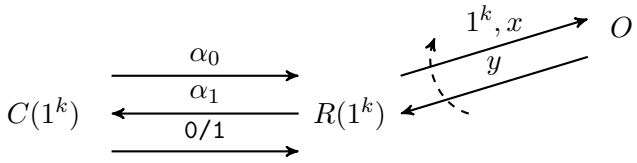

Figure 1: A black-box reduction R.

In contrast, some intractability assumptions may be unbridgable at least as far as black-box reductions are concerned. Indeed, as mentioned above, the results of [Pas11, GW11] yield some results in this direction, separating unbounded-round and bounded-round assumptions [Pas11] and unbounded-challenger and efficient-challenger assumptions [GW11]. The results in this paper further elucidate this landscape.

Black-box Reductions. We consider probabilistic polynomial time Turing reductions—i.e., *black-box reductions*. A black-box reduction refers to a probabilistic polynomial-time oracle algorithm. Roughly speaking, a black-box reduction for basing the security of a primitive P on the hardness of an assumption (C, t), is a probabilistic polynomial-time oracle machine R such that whenever the oracle O "breaks" P with respect to the security parameter k, then R^O "breaks" (C, t) with respect to a *polynomially-related* security parameter k' such that k' can be efficiently computed given k; see Figure 1.

We restrict ourselves to the case when $k' = k$. This is without loss of generality because we can always redefine the challenger C so that it on input k acts as if its input actually was k' (since k' can be efficiently computed given k). To formalize this notion, we thus restrict the oracle machines R such that on input 1^k, they always query their oracle on inputs $(1^k, \cdot)$. *Additionally, we restrict our attention to black-box reductions that only make polynomially many queries to its oracle. When discussing polynomial-time reductions, this restriction vacously holds, but, in general, a super-polynomial time reduction may make a super-polynomial number of queries to its oracle. As far as we are aware (at least at the time when this paper was originally written) typical super-polynomial time reductions all adhere to this restriction. Additionally, in our eyes, reductions making only a polynomial number of queries are the most relevant one from a practical point of view.*[17]

Definition 3.1. *We say that R is a* security-parameter preserving *black-box reduction if R is an oracle machine and there exists some polynomial $p(\cdot)$ such that $R(1^k)$ only queries its oracle with inputs of the form $(1^k, x)$, where $x \in \{0,1\}^*$, and additionally R only makes at most $p(k)$ queries to its oracle.*[18]

A more liberal notion of a black-box reduction allows the reduction R to (on input 1^k) query its oracle on multiple security parameters (that are all polynomially related to k). In our eyes, such a liberal notion is less justified from a practical point of view (and as far as we are aware, cryptographic reductions typically do not rely on such liberal reductions); nevertheless, all our

[17] **Added in March 2017:** The emphasized text was added in March 2017. We are very grateful to Dakshita Khurana and Amit Sahai for pointing out that this restriction was needed for our result on super-polynomial reductions (i.e., Theorem 5.9). We point out that our proof actually already explictly considered the impact of the number of queries (See Claim 4.10 where the explicit bound is given in terms of the number of queries) but the earlier Theorem 5.9 was incorrect as stated, as our notion of a reduction did not make this cruicial restriction.

[18] **Added in March 2017:** The restriction to the number of queries was added. Perhaps a better name for these types of security-preserving reductions, suggested by Amit Sahai, is an "instance-to-instance" reduction.

proofs for the case of polynomial-time reductions directly apply also for such a notion of black-box reductions.[19]

Interpreting Reductions to Assumptions. Note that our notion of an assumption (C, t) is not fully-specified in the sense that it does not talk about the complexity of the attacker A that attempts to break the assumption. This enables us to talk about, for instance, "polynomial-time hardness of (C, t)" (by restricting to polynomial-time attackers) or "subexponential hardness of (C, t)" (by restricting to subexponential-time attackers).

To simplify our treatment, we do not directly discuss the complexity classes against which an assumption is assumed to be hard, but rather quantify our results in terms of the complexity of the reduction used in the proof of security. The complexity of a reduction together with the complexity class of the attacker for which we want a primitive P to be secure, determine the complexity class with respect to which the underlying assumption needs to be secure: if we have a $T(\cdot)$-time security-preserving reduction R for basing the security of a primitive P on the hardness of assumption (C, t), then to conclude that P is secure w.r.t. $S(\cdot)$-time attackers, we need to assume that (C, t) cannot be broken w.r.t. $T(\cdot) + \text{poly}(S(\cdot))$ time attackers.[20] Note that we here rely on the security-parameter preserving property of R: If R were not security preserving, the combined running-time of the reduction and presumed attacker could be $S(T(k))$.

4 Security of Perfect Adaptive NIZK

We recall the definition of non-interactive proofs in the Common Reference String (CRS) model. For generality (and since we are proving a lower bound) we allow the CRS to be generated by an arbitrary polynomial-time samplable distribution (as opposed to requiring it to be uniformly distributed over strings of a specific length). In the adaptively-sound notion of a non-interactive proof/argument,[21] we require that soundness holds even if the attacker may adaptively pick the statement after having seen the CRS. We consider only proofs/arguments for languages in **NP** where the *prover is efficient* when given an **NP**-witness.

Definition 4.1 (Non-Interactive Proofs/Arguments)**.** *A triple of algorithms, $(\mathcal{D}, P, V)$, is called a non-interactive proof system (with non-adaptive soundness) for a language L if the algorithm $\mathcal{D}$ (the "CRS generator") is probabilistic polynomial-time, the algorithm V (the "verifier") is a deterministic polynomial-time, and P (the "prover") is probabilistic polynomial-time such that the following two conditions hold:*

- Completeness: *There exists a negligible function μ such for every $x \in L$, every $w \in R_L(x)$ and every $k \in \mathbb{N}$,*

$$\Pr\left[\ \rho \leftarrow \mathcal{D}(1^k, 1^{|x|});\ \pi \leftarrow P(1^k, x, w, \rho)\ :\ V(1^k, x, \rho, \pi) = 1\ \right] \geq 1 - \mu(k)$$

[19] In fact, as remarked in [CMP09], in the context of black-box separations, restricting to reductions that only query its oracle on a single security parameter is actually without loss of generality if we consider primitives with "standard" cryptographic definitions where to break security an attacker only needs to be successful for *infinitely many* input lengths.

[20] **Added in March 2017:** With our previous notion of a security preserving reduction (without the polynomial query restriction,) the expression was $T(\cdot) \cdot S(\cdot)$.

[21] Recall that an interactive *argument* is an interactive proof where the soundness condition is only required to hold w.r.t. *computationally-bounded* provers.

- Soundness: *For every algorithm B and every positive polynomial q, there exists a negligible function μ such that for every $k \in \mathbb{N}$ and every $x \notin L$ such that $|x| \leq q(k)$*

$$\Pr\left[\ \rho \leftarrow \mathcal{D}(1^k, 1^{|x|});\ \pi' \leftarrow B(1^k, x, \rho)\ :\ V(1^k, x, \rho, \pi') = 1\ \right] \leq \mu(k)$$

If additionally the following condition holds, then we call $(\mathcal{D}, P, V)$ an adaptively-sound non-interactive proof system*:*

- Adaptive Soundness: *For every algorithm B and every polynomial q, there exists a negligible function μ such that for every $k \in \mathbb{N}, n \in [q(k)]$*

$$\Pr\left[\ \rho \leftarrow \mathcal{D}(1^k, 1^n);\ (x, \pi') \leftarrow B(1^k, 1^n, \rho)\ :\ V(1^k, x, \rho, \pi') = 1 \wedge |x| = n \wedge x \notin L\ \right] \leq \mu(k)$$

Finally, if the soundness (resp adaptive soundness) condition only holds w.r.t polynomial-time adversaries B, we call $(\mathcal{D}, P, V)$ a non-interactive argument *(resp., an* adaptively-sound *non-interactive argument).*

Let us turn to defining zero-knowledge. Also here there is a non-adaptive and an adaptive version. In the *non-adaptive* definition of zero-knowledge from [BFM88], there is a single simulator that after seeing the statement to be proven, generates both the CRS and the proof at the same time. In the *adaptive* definition from [FLS90], there are two simulators—the first of which must output a CRS ρ before seeing any statements to the proven, and the second generates a proof for a given statement x (w.r.t. to the selected CRS ρ) . The stronger adaptive definition guarantees zero-knowledge even when the statement to be proved is chosen as a function of the CRS. We focus only on adaptive zero-knowledge.

Definition 4.2 (Non-Interactive Zero-Knowledge)**.** *Let $(\mathcal{D}, P, V)$ be an non-interactive proof system for the language L. We say that $(\mathcal{D}, P, V)$ is (adaptively) zero-knowledge if there exists two probabilistic polynomial-time simulators S_1 and S_2 such that for every polynomial q, every non-uniform polynomial-time statement-witness chosing algorithm $c(\cdot, \cdot, \cdot)$ that on input $(1^k, 1^n, \rho)$ outputs a n-bit statement x and witness w such that $(x, w) \in R_L$, the following two ensembles are computationally indistinguishable*

$$\left\{\rho \leftarrow \mathcal{D}(1^k, 1^n);\ x, w \leftarrow c(1^k, 1^n, \rho); \pi \leftarrow P(1^k, x, w, \rho)\ :\ (\rho, x, \pi)\right\}_{k \in \mathbb{N}, n \in [q(k)], z \in \{0,1\}^*}$$

$$\left\{(\rho, \mathsf{aux}) \leftarrow S_1(1^k, 1^n);\ x, w \leftarrow c(1^k, 1^n, \rho); \pi' \leftarrow S_2(1^k, x, \mathsf{aux})\ :(\rho, x, \pi')\right\}_{k \in \mathbb{N}, n \in [q(k)], z \in \{0,1\}^*}$$

We furthermore say that $(\mathcal{D}, P, V)$ is perfect (resp., statistical) zero-knowledge if the above ensembles are identically distributed (resp., statistically close).

Note that in the ensembles considered in the definition of zero-knowledge, the input z (in the index of the ensemble) is not given to any of the algorithms in the experiments that define the ensembles; this input is there only to be provided as an input to the distinguisher (to ensure that indistinguishability holds w.r.t. non-uniform PPT distinguishers)—recall that in the definition of computational indistinguishability (see Definition 2.1), the distinguisher get the ensemble index as input. (Let us also remark that although z may be of an arbitrary length, since distinguisher's running time is bounded by poly(k), it can only read a polynomial-lenght prefix of z, and thus effectively, we are only considering polynomial-legnth advice strings.)

We use the (common) acronym "NIZK" to denote a non-interactive zero-knowledge proof or argument. Feige, Lapidot and Shamir and Bellare and Yung [FLS90, BY96] (building on [BFM88]) show that the existence of enhanced trapdoor permutations [GR13] implies that all of **NP** has an adaptively-sound NIZK, but the zero-knowledge property is only computational. As mentioned, Groth, Ostrovsky and Sahai [GOS06] show (under some number-theoretic assumptions) that all of **NP** has a *perfect* NIZK with *non-adaptive* soundness. More recently, Abe and Fehr [AF07] present a perfect NIZK for **NP** also with adaptive soundness but based the soundness property on a "knowledge-extractaction" assumption (simular to the "knowledge-of-exponent" assumption of [Dam91]) rather than an intractability assumption.

We aim to prove limitations of basing notions of adaptive soundness for perfect or statistical NIZK for **NP** on standard intractability assumptions. Let us first explicitly define what it means to break adaptive soundness of a NIZK.

Definition 4.3 (Breaking Adaptive Soundness)**.** *We say that a probabilistic algorithm A* breaks adaptive soundness of $(\mathcal{D}, P, V)$ w.r.t the language L on input lengths $q(\cdot)$ with probability $\epsilon(\cdot)$ *if for every $k \in N$,*

$$\Pr\left[\ \rho \leftarrow \mathcal{D}(1^k, 1^{q(k)});\ (x, \pi') \leftarrow A(1^k, \rho)\ :\ V(1^k, x, \rho, \pi') \wedge |x| = q(k) \wedge x \notin L\ \right] \geq \epsilon(k)$$

Indeed, if ϵ is not negligible, then the adaptive soundness condition of Definition 4.1 is violated. Let us turn to defining what it means to base adaptive soundness on an intractability assumption C.

Definition 4.4 (Basing Adaptive Soundness on the Hardness of C)**.** *We say that R is a* black-box reduction for basing adaptive soundness of $(\mathcal{D}, P, V)$ w.r.t. L and input lengths $q(\cdot)$ on the hardness of C w.r.t threshold $t(\cdot)$ *if R is a security-parameter preserving black-box reduction and there exists a polynomial $p(\cdot, \cdot)$ such that for every probabilistic machine A that breaks adaptive soundness of $(\mathcal{D}, P, V)$ w.r.t L and inputs lengths $q(\cdot)$ with probability $\epsilon(\cdot)$, for every $k \in \mathbb{N}$, the combined machine R^A breaks C w.r.t threshold t with probability $p(\epsilon(k), 1/k)$ on input 1^k.*

We are now ready to formally state Theorem 1.1 from the introduction.

Theorem 4.5. *Assume the existence of one-way functions that are secure against polynomial-size circuits. Let $(\mathcal{D}, P, V)$ be a statistical non-interactive adaptively zero-knowledge argument for an* **NP***-complete language L, let $q(k) \in \Theta(k^c)$ be a "nice" function[22] (where $c > 0$) and let (C, t) be any efficient-challenger assumption. If there exists a polynomial-time (resp., polynomial circuit-size) black-box reduction R for basing adaptive soundness of $(\mathcal{D}, P, V)$ w.r.t L and input lengths $q(\cdot)$ on the hardness of C w.r.t threshold t, then there exists a probabilistic polynomial-time (resp., polynomial circuit-size) machine B and a polynomial $p'(\cdot)$ such that for infinitely many $k \in \mathbb{N}$, machine B breaks C w.r.t threshold t with probability $\frac{1}{p'(k)}$ on input 1^k.*

Let us remark that as shown in [Ost91, OW93, PS05], any (even computational) NIZK for a hard-on-the-average language (for non-uniform polynomial-time algorithms), implies the existence of non-uniformly hard one-way functions. So the assumption of one-way functions could actually be relaxed to assume that there exists a hard-on-the average language in **NP**.

Let us also remark that since Theorem 4.5 applies to all statistical non-interactive zero-knowledge *arguments* for **NP**-complete languages, it also applies to statistical non-interactive zero-knowledge *proofs*. But this conclusion is less interesting as it was already known that only languages in **BPP**/1

[22] By nice we here mean that $q(k)$ can be computed in poly(k) time.

(which, in particular, is contained in $\mathbf{P}/poly$) can have statistical NIZK proofs with adaptive zero-knowledge and just non-adaptive soundness [PS05].[23]

4.1 Proof of Theorem 4.5 for the Case of Perfect NIZK

We start by considering a simplified case when the zero-knowledge property is *perfect*; that is, we consider perfect NIZK.

Proof outline. Let $g : \{0,1\}^* \to \{0,1\}^*$ be a length-doubling PRG (which can be constructed based on one-way functions [HILL99]). Consider the language $L = \{\ g(s)\ \mid\ s \in \{0,1\}^*\ \}$. Assume there exists a perfect NIZK $(\mathcal{D}, P, V)$ for L, and assume there exists a black-box reduction R for basing adaptive soundness of $(\mathcal{D}, P, V)$ w.r.t L and input length $q(\cdot)$ (where $q(k) \in \Theta(k^c)$ is a "nice" function and $c > 0$) on the hardness of C w.r.t threshold t; the existence of these objects follows from the hypothesis of Theorem 4.5 as $L \in \mathbf{NP}$. That is, there exists a polynomial $p(\cdot,\cdot)$ such that for every probabilistic machine A that breaks adaptive soundness of $(\mathcal{D}, P, V)$ w.r.t L and inputs lengths $q(\cdot)$ with probability $\epsilon(\cdot)$, for every $k \in \mathbb{N}$, the combined machine R^A breaks C w.r.t. threshold t with probability $p(\epsilon(k), 1/k)$ on input 1^k; that is,

$$\Pr\left[\langle R^A, C\rangle(1^k) = 1\right] \geq t(k) + p(\epsilon(k), 1/k). \tag{1}$$

Our overall proof strategy proceeds in two steps:

- We first devise a particular *inefficient* attacker A that breaks adaptive soundness of $(\mathcal{D}, P, V)$ with *overwhelming probability*. By Equation 1, it then follows that there exists a positive polynomial $\hat{p}(\cdot)$ such that R^A breaks C with probability $t(k) + \frac{1}{\hat{p}(k)}$ for infinitely many k.
- We then show how to simulate A by an efficient "emulator" $\tilde{A}$[24] such that the emulator is indistinguishable from the original. This implies that $R^{\tilde{A}}$ breaks C with probability $t(k) + \frac{1}{\hat{p}(k)} - \mu(k)$ for infinitely many k, where $\mu(k)$ is a negligible function, which proves the theorem since $R^{\tilde{A}}$ can be implemented in polynomial time.

In our actual proof, the above two steps happen somewhat in parallel. We simultaneously construct the attacker A and the emulator $\tilde{A}$, and use the emulator to argue that the attacker is actually successful in breaking soundness. Towards this goal, we start by first constructing a "basic" attacker A_0, and the emulator $\tilde{A}$, and show that $\tilde{A}$ indistinguishably emulates A_0 given any "well-formed" CRS (in the range of the CRS generating algorithm); $\tilde{A}$, however, does not indistinguishably emulate A_0 given "invalid" CRS. Our actual attacker A is then constructed by running A_0 when given as input a well-formed CRS, and running $\tilde{A}$ otherwise.

For simplicity of notation, we assume that $q(k) = 2k$; it is easy to see that the same proof works as long as $q(k) \in \Theta(k^c)$ and q is efficiently computable.

Constructing the Basic Attacker A_0. We present a *randomized* attacker A_0 that on each query (i.e., invokation by the reduction) uses fresh random coins; recall that by the assumption that R is a security-parameter preserving reduction each such query is of the form $1^k, \rho$, where $\rho \in \{0,1\}^*$. (Recall that our notion of a black-box reduction requires the reduction to work even if

[23] [PS05] also showed that only languages in $\mathbf{AM} \cap co\mathbf{AM}$ can have statistical NIZKs satisfying even just non-adaptive zero-knowledge and non-adaptive soundness.

[24] We use the term emulator to describe $\tilde{A}$ in order not to overload the word simulator.

the attacker is probabilistic. As we point out in Section 4.3.1, at the cost of a minor complication, the proof can be adapted to work also if only considering reductions that work as long as the attacker is deterministic.)

Let S_1, S_2 be the zero-knowledge simulators associated with $(\mathcal{D}, P, V)$; recall that S_1 outputs a simulated CRS ρ (and some state information $\texttt{aux}$), and S_2 on input a statement x (and the passed along state $\texttt{aux}$ from S_1) produces a simulated proof of x w.r.t. the CRS ρ output by S_1.

On input 1^k and a string ρ (to be interpreted as a CRS), attacker A_0 proceeds as follows (letting $n = q(k) = 2k$):

- **Inverting S_1:** A_0 *uniformly* picks a random tape r such that $S_1(1^k, 1^n)$ outputs $\rho, \texttt{aux}$ when given the random tape r, where $\texttt{aux}$ is some arbitrary string. If no such that r exists, it returns $\perp$.[25] *Note that this step is not necessarily efficient.*
- **Pick FALSE statement x:** Next, A_0 uniformly picks a string $x \in \{0,1\}^n$. Note that, except with probability at most 2^{-k}, it holds that $x \notin L$ (there are 2^{2k} strings, and at most 2^k can be in the range of the PRG g).
- **Generate SIMULATED proof π:** Finally, A_0 runs the simulator $S_2(1^k, 1^n, x, \texttt{aux})$ to produce the proof π, and returns (x, π).

As noted above, with high probability the statement x picked by A_0 is false. But it remains to argue that the proof π of x output by A_0 is accepting (for the reference string ρ). (Very roughly speaking, the intuition for why π ought to be accepting is that the statement x is indistinguishable from a true statement (in the range of the PRG), and for such statements the simulator ought to produce accepting proofs.) As mentioned above, towards formalizing this intuition, we first present an *efficient* emulator, $\tilde{A}$, for A_0.

Constructing the Emulator $\tilde{A}$. We devise an efficient "emulator", $\tilde{A}$, that on input 1^k and a reference string ρ, proceeds as follows:

- **Pick TRUE statement x:** $\tilde{A}$ uniformly picks a string $s \in \{0,1\}^k$, and lets $x = g(s)$. Note that by definition $x \in L$.
- **Generate HONEST proof π:** $\tilde{A}$ runs the honest prover algorithm $P(1^k, \rho, x, s)$ to produce the proof π, and returns (x, π).

The following lemma—which cruicially relies on the *perfect* NIZK property of $(\mathcal{D}, P, V)$—shows that $\tilde{A}(1^k, \rho)$ is a good emulator for $A_0(1^k, \rho)$ for *every ρ in the range of $\mathcal{D}(1^k, 1^n)$*.

Lemma 4.6. *The following two ensembles are computationally indistinguishable*[26] *ensembles is there to*

$$\left\{A_0(1^k, \rho)\right\}_{k \in \mathbb{N}, \rho \in \mathcal{D}(1^k, 1^n), z \in \{0,1\}^*}$$

$$\left\{\tilde{A}(1^k, \rho)\right\}_{k \in \mathbb{N}, \rho \in \mathcal{D}(1^k, 1^n), z \in \{0,1\}^*}$$

Proof. To prove the lemma, we consider a hybrid attacker that proceeds just as A_0 but instead picks a *true* statement. More precisely, $Sim(1^k, \rho)$ proceeds as follows:

[25] Since we assume that the zero-knowledge simulation is perfect, this can only happen in case ρ is not in the range of $\mathcal{D}(1^k, 1^n)$.

[26] The variable z in the index to the ensembles is there to ensure that indistinguishability holds also w.r.t. non-uniform polynomial-time distinguishers.

- **Inverting S_1:** Pick a random tape r such that $S_1(1^k, 1^{2k})$ outputs $\rho, \texttt{aux}$ given the random tape r, where $\texttt{aux}$ is some arbitrary string. (If no such that r exists, it returns $\perp$.)
- **Picking TRUE statement x:** Pick a string $s \in \{0,1\}^k$, and let $x = g(s)$.
- **Generate SIMULATED proof π:** Run the simulator $S_2(1^k, 1^{2k}, x, \texttt{aux})$ to produce the proof π, and return (x, π).

Consider any CRS ρ in the range of $\mathcal{D}(1^k, 1^n)$. Note that the only difference between $Sim(1^k, \rho)$ and $\tilde{A}(1^k, \rho)$ is whether the proof π is being simulated (as in Sim), or honestly generated (as in $\tilde{A}$). It thus follows directly from the *perfect* adaptive zero-knowledge property of $(\mathcal{D}, P, V)$ that the outputs of $Sim(1^k, \rho)$ and $\tilde{A}(1^k, \rho)$ are identically distributed. (Note that we crucially rely on the perfect ZK property since we are comparing executions *conditioned* on a fixed CRS.)

Next, note that for every CRS ρ in the range of $\mathcal{D}(1^k, 1^n)$, the only difference between $A_0(1^k, \rho)$ and $Sim(1^k, \rho)$ is choice of the statement x; random in the case of A_0 and pseudorandom in case of Sim. Intuively it should follow from the security of the pseudorandom generator g that the outputs of $A_0(1^k, \rho)$ and $Sim(1^k, \rho)$ are indistinguishable. But there is a problem: A_0 and Sim are not efficiently computable (recall that they both "invert S_1"), so *efficiently* contradicting the pseudorandomness property of g becomes problematic. This, issue, however, can be dealt with by noting that the only inefficient parts of A_0 (and Sim) happen *before* A_0 chooses the statement x. We can thus non-uniformly fix these inefficient computations, and rely on the fact that the pseudorandomness property of g holds even against non-uniform polynomial-time algorithms. □

We can now show that A_0 breaks adaptive soundness of $(\mathcal{D}, P, V)$ with overwhelming probability.

Claim 4.7. *There exists a negligible function μ such that for every $k \in \mathbb{N}$,*

$$\Pr\left[\ \rho \leftarrow \mathcal{D}(1^k, 1^n);\ (x, \pi') \leftarrow A_0(1^k, \rho)\ :\ V(1^k, x, \rho, \pi') = 1 \wedge x \notin L\ \right] \geq 1 - \mu(k)$$

Proof. Let us first note that by the completeness property of $(\mathcal{D}, P, V)$, we have that there exists a negligible function μ' such that for every $k \in \mathbb{N}$,

$$\Pr\left[\ \rho \leftarrow \mathcal{D}(1^k, 1^n);\ (x, \pi') \leftarrow \tilde{A}(1^k, \rho)\ :\ V(1^k, x, \rho, \pi') = 1\ \right] \geq 1 - \mu'(k)$$

It then follows by Lemma 4.6 that there exists a negligible function μ'' such that

$$\Pr\left[\ \rho \leftarrow \mathcal{D}(1^k, 1^n);\ (x, \pi') \leftarrow A_0(1^k, \rho)\ :\ V(1^k, x, \rho, \pi') = 1\ \right] \geq 1 - \mu''(k)$$

The claim follows from the (earlier made) observation that A_0 picks false statements with overwhelming probability. □

We have thus shown that A_0 is a "good" attacker. The problem, however, is that our emulator $\tilde{A}$ only emulates A_0 when given a "well-formed" CRS as input. We now show how to modify A_0 so that it is (always) indistinguishably emulated by $\tilde{A}$, yet it still is a good attacker.

Constructing the actual attacker A. Our actual attacker A will run A_0 when given a "well-formed" CRS, and otherwise it will "give-up" and simply run $\tilde{A}$. More precisely, let $A(1^k, \rho)$ be an algorithm that proceeds as follows:

- If ρ is in the range of $\mathcal{D}(1^k, 1^n)$, let $(x, \pi) \leftarrow A_0(1^k, \rho)$, and otherwise let $(x, \pi) \leftarrow \tilde{A}(1^k, \rho)$;
- Output (x, π).

The following corollary, which is direct consequence of Lemma 4.6 and the construction of A shows that $\tilde{A}(1^k, \rho)$ is a good emulator for $A(1^k, \rho)$ for *every* ρ (not necessarily only those in the range of $\mathcal{D}(1^k, 1^n)$).

Corollary 4.8. *The following two ensembles are computationally indistinguishable*

$$\left\{A(1^k, \rho)\right\}_{k\in\mathbb{N},\rho\in\{0,1\}^*,z\in\{0,1\}^*}$$

$$\left\{\tilde{A}(1^k, \rho)\right\}_{k\in\mathbb{N},\rho\in\{0,1\}^*,z\in\{0,1\}^*}$$

As corollary of Claim 4.7 and the observation that A runs A_0 whenever it receives a well-formed CRS as input (which is always the case in "soundness experiment"), we have that A still breaks the adaptive soundness of $(\mathcal{D}, P, V)$:

Corollary 4.9. *There exists a negligible function μ such that for every $k \in \mathbb{N}$,*

$$\Pr\left[\ \rho \leftarrow \mathcal{D}(1^k, 1^n);\ (x, \pi') \leftarrow A(1^k, \rho)\ :\ V(1^k, x, \rho, \pi') = 1 \wedge x \notin L\ \right] \geq 1 - \mu(k)$$

As a consequence of Corollary 4.9 and the fact that R is a good reduction (i.e., Equation 1), there exists a positive polynomial $\hat{p}(\cdot)$ such that R^A breaks C with probability $t(k) + \frac{1}{\hat{p}(k)}$ for infinitely many k; that is,

$$\Pr\left[\langle R^A, C\rangle(1^k) = 1\right] \geq t(k) + \frac{1}{\hat{p}(k)} \tag{2}$$

We are not yet done, however, since A is not efficient.

Proving that $\tilde{A}$ is a good emulator for A. To conclude the proof of the theorem, we finally show that the efficient $\tilde{A}$ is a "good" emulator for A, even if A is repeatedly invoked by R (in an interaction with C), and even if R queries A on an *arbitrary* CRS ρ. R may query A on "untypical" and even invalid CRS, and it is to deal with this fact that we required indistinguishability of A and $\tilde{A}$ to hold for *every* CRS in Corollary 4.8.

Claim 4.10. *For every efficient C and R, there exists a negligible function μ such that for every $k \in \mathbb{N}$,*

$$\left|\Pr\left[\langle R^{\tilde{A}}, C\rangle(1^k) = 1\right] - \Pr\left[\langle R^A, C\rangle(1^k) = 1\right]\right| \leq \mu(k).$$

Proof. The key idea behind the proof is that all queries by R are answered *independently* (by both A and $\tilde{A}$), and thus we can perform a hybrid argument which essentially reduces us to the "single-shot" scenario considered in Corollary 4.8. Let us point out that to have this independence property, it is cruicial that A (like A_0 and $\tilde{A}$) generate a "fresh" random statement (independent of all earlier statements) on each query it receives. (If, for instance, A had been stateful and had picked a random statement x *once* (before seing any CRS), and then provided proofs of the *same* x in every query, then this independence property would not hold. This clarifies why our proof does not extend to rule out also reductions proving *non-adaptive* soundness of $(\mathcal{D}, P, V)$.)

We proceed to a formal proof, which proceeds using a rather standard hybrid argument over the oracle queries made by R; the only (minor) subtlety is that we need to order the hybrids so that "later" oracle queries are answered by the efficient $\tilde{A}$. We can then use non-uniformity to deal with the fact that A is inefficient.

In more detail, assume for contradiction that the claim is false. That is, there exists a polynomial p' such that for infinitely many $k \in \mathbb{N}$,

$$\left|\Pr\left[\langle R^{\tilde{A}}, C\rangle(1^k) = 1\right] - \Pr\left[\langle R^{A}, C\rangle(1^k) = 1\right]\right| \geq \frac{1}{p'(k)}.$$

Let $m(k)$ be an upper-bound on the number of oracle queries by R on input 1^k, and fix a canonical k for which the above happens. Consider a sequence of hybrid experiments $H_0, \ldots, H_{m(k)}$, where H_i is defined as the output of $C(1^k)$ after communicating with $R(1^k)$ where the first i oracle queries of R are answered by A, and the remaining ones are answered by the efficient $\tilde{A}$. Note that both A and $\tilde{A}$ are *stateless* and thus these hybrid experiments are well defined. Furthemore, note that

- $H_0 = \langle R^{\tilde{A}}, C\rangle(1^k)$, and
- $H_{m(k)} = \langle R^{A}, C\rangle(1^k)$

It follows that there exists some j such that

$$|\Pr[H_{j+1} = 1] - \Pr[H_j = 1]| \geq \frac{1}{m(k)p'(k)}.$$

Note that up until the point when R receives its $(j+1)$st statement-proof pair (x_{j+1}, π_{j+1}) back from its oracle, the two experiments H_j, and H_{j+1} proceed identically the same. Thus, if they are distinguishable with probability $\frac{1}{m(k)p'(k)}$, there exists some prefix τ^k of the execution of H_j[27], up until and including the $(j+1)$th query, ρ_{j+1}, of R, such that conditioned on this prefix τ^k, H_j and H_{j+1} are distinguishable with probability $\frac{1}{m(k)p'(k)}$. Recall that the only difference between H_j and H_{j+1} is that in H_j the $(j+1)$th query is answered by A whereas in H_{j+1} it is answered by (the efficient) $\tilde{A}$. This contradicts Corollary 4.8, since in both experiements, all subsequent queries of R are answered by the efficient $\tilde{A}$. □

The proof of the theorem (w.r.t. Perfect NIZK) is concluded by combining Equation 2 with Claim 4.10 and observing that $R^{\tilde{A}}$ can be implemented in polynomial time.

4.2 Proof of Theorem 4.5 for the Case of Statistical NIZK

We now show how to extend the proof to deal with *statistical* NIZK. In our proof above, we relied on the perfect NIZK property to prove that $\tilde{A}$ emulates A_0 for *every* CRS in the range of $\mathcal{D}(1^k, 1^n)$. This is no longer true for statistical NIZK: there may be "deviating" CRS for which the zero-knowledge simulation has a large statistical deviation (and thus $\tilde{A}$ will not correctly emulate A_0 for such CRS). Furthermore, R may query its oracle on such deviating CRS.

We note, however, that "deviating" CRS must be "rare" (or else $(\mathcal{D}, P, V)$ could not be statistical zero-knowledge); we can thus afford to have A "give-up" (and instead run $\tilde{A}$) on such deviating CRS (and not just on CRS that are outside the range of $\mathcal{D}$). Since deviating CRS are rare, such a modified A will still succeed in breaking soundness. Furthemore, such an A can now be emulated by $\tilde{A}$.

More precisely, we say that a CRS ρ is (k, α)-*deviating* if the outputs of $\tilde{A}(1^k, \rho)$ and $Sim(1^k, \rho))$ are α-far in statistical distance, where Sim is defined as in the proof of Lemma 4.6. The following claim shows that deviating CRS are rare.

[27] Technically, the prefix includes the random tape of C and R and all the answers to the first j queries by R.

Claim 4.11. *There exists a negligible function μ' such that for every $k \in \mathbb{N}$,*

$$\Pr\left[\ \rho \leftarrow \mathcal{D}(1^k, 1^n)\ :\ \rho \text{ is } (k, \mu'(k))\text{-deviating}\ \right] \leq \mu'(k)$$

Proof. Assume for contradiction that there exists a positive polynomial $p(\cdot)$ such that for infinitely many k, with probability at least $\frac{1}{p(k)}$ over the choice of ρ, the statistical distance between the outputs of $\tilde{A}(1^k, \rho)$ and $Sim(1^k, \rho))$ is at least $\frac{1}{p(k)}$. Thus, the statistical distance between the following distributions is at least $\frac{1}{p(n)^2}$:

$$\left\{\rho \leftarrow \mathcal{D}(1^k, 1^n); s \leftarrow \{0,1\}^k; x = g(s); \pi \leftarrow P(1^k, x, s)\ :(\rho, x, \pi)\ \right\}$$

$$\left\{\rho \leftarrow \mathcal{D}(1^k, 1^n); x, \pi' \leftarrow Sim(1^k, \rho)\ :(\rho, x, \pi')\ \right\}$$

By the statistical ZK property of $(\mathcal{D}, P, V)$, there exists some negligible function μ'' such that the latter distribution is $\mu''(k)$-close to the following distribution:

$$\left\{\rho, \mathtt{aux} \leftarrow S_1(1^k, 1^n); x, \pi' \leftarrow Sim(1^k, \rho)\ :(\rho, x, \pi')\ \right\},$$

which in turn is identical to

$$\left\{\rho, \mathtt{aux} \leftarrow S_1(1^k, 1^n); s \leftarrow \{0,1\}^k; x = g(s); \pi' \leftarrow S_2(1^k, x, \mathtt{aux})\ :(\rho, x, \pi')\ \right\}$$

But this contradicts the statistical ZK property of $(\mathcal{D}, P, V)$. □

Let us also note that a slightly modified version of Lemma 4.6, where we restrict to CRS ρ that are *not* $(k, \mu'(k))$-deviating, holds using the same proof. More precisely, let $Good(1^k)$ denote the set of $\rho \in \mathcal{D}(1^k, 1^n)$ such that ρ is not $(k, \mu'(k))$-deviating.

Lemma 4.12. *The following two ensembles are computationally indistinguishable*

$$\left\{A_0(1^k, \rho)\right\}_{k \in \mathbb{N}, \rho \in Good(1^k), z \in \{0,1\}^*}$$

$$\left\{\tilde{A}(1^k, \rho)\right\}_{k \in \mathbb{N}, \rho \in Good(1^k), z \in \{0,1\}^*}$$

We now modify A to run $\tilde{A}$ not only when the CRS is not well-formed, but also when the CRS is $(k, \mu'(k))$-deviating. Given this new A, Corollary 4.8 follows as before (from the new Lemma 4.12). Recall that Corollary 4.9 (i.e., that A breaks the adaptive soundness property) previously followed directly from Claim 4.7 (i.e., that A_0 breaks the adaptive soundness property) and the observation that A runs *always* runs A_0 whenever it receives a well-formed CRS (which is the case in the "soundness experiment"). This is no longer true for the modifed A. However, by Claim 4.11 we have that A_0 is invoked with *overwhelming probability* (instead of always) in the "soundness experiment", and thus Corollary 4.9 still follows by a union bound.

The rest of the proof is unchanged. This concludes the proof of Theorem 4.5.

4.3 Ramifications

In this section we discuss some extensions of Theorem 4.5.

4.3.1 Extensions to Deterministic Attackers

In the proof of Theorem 4.5, we consider a *randomized* oracle A, and thus only rule out reductions that work for all randomized attackers. Following [Pas11], and using a technique from Goldreich and Krawczyk [GK96], we can extend the proof to also rule out reductions that only work for deterministic attackers.

First, if we consider a deterministic attacker R, then we may assume (w.l.o.g.) that R never asks the same query $(1^k, \rho)$ twice (since can we can internally emulate responses to all repeated queries). Next, given a randomized attacker A, we define a *deterministic* attacker A^f that on input $(1^k, \rho)$ sets the random tape of A to equal $f(1^k, \rho)$ and next executes $A(1^k, \rho)$, as defined above, using this pre-selected random tape. (Note that this attacker is "adaptive": it selects the statement to prove as a function of the CRS ρ.) Let RF_k be a randomly chosen $M(k)$-wise independent hash function $\{0,1\}^{\text{poly}(k)} \to \{0,1\}^{\text{poly}(k)}$ (for appropriate-length polynomials), where $M(k)$ is an upperbound on the number of oracle queries made by $R(1^k)$. Note that $A_0^{RF_k}(1^k)$ (resp. $A^{RF_k}(1^k)$) acts exactly as the attacker $A_0(1^k)$ (resp. $A(1^k)$) defined in our proof as long as R never asks the attacker the same query twice (which it does not according to the above convention). It follows that R's view when communnicating with A^{RF_k} and A are identical. We can thus replace A_0 with $A_0^{RF_k}$ in Lemma 4.6 and Claim 4.7, and A with A^{RF_k} in Corollary 4.8, Corollary 4.9 and Claim 4.10. (Let us point that for the case of Lemma 4.6 it is important that perfect ZK holds w.r.t. *non-uniform* polynomial-time input-witness chosing algorithms, as the statement x is chosen by applying a function $f \leftarrow RF_k$ to ρ and such a function f has polynomial-length description.) Finally, by an averaging argument (see [Pas11] for more details), with overwhelming probability over the choice of a $f \leftarrow RF_k$, machine A^f breaks adaptive soundness of $(\mathcal{D}, P, V)$ with overwhelming probability, and for each such "good" choice of f we have that $R^{A^f}(1^k)$ breaks C with advantage negligbly close to $\frac{1}{\hat{p}(k)}$, where $\hat{p}(\cdot)$ is a polynomial; thus $R^{A^{RF_k}}(1^k)$ also breaks C with advantage negligibly close to $\frac{1}{\hat{p}(k)}$, and thus the proof can be concluded just as before.

4.3.2 Extensions to Reductions with Non-uniform Advice

Our current lower bound only considers security reductions that do not get any non-uniform advice that depends on the attacker. As mentioned in the introduction, some quite commonly used proof techniques in cryptography (and in particular the techniques used in the proof of Theorem 4.5!) rely on the reduction receiving a non-unform advice string that depends on the attacker. However, the recent work of [CLMP13] provides techniques that extend our separation result to deal also with reductions receiving such non-uniform advice; we mention that to apply the techniques of [CLMP13], it is cruicial that we can rule out reductions that only need to work for deterministic attackers. We refer the reader to [CLMP13] for further details.

4.3.3 On Adaptive Culpable Soundness

Groth, Ostrovsky and Sahai [GOS06] presented a weaker definition of adaptive soundness, which they call *adaptive culpable soundness*. Roughly speaking, (1) they restrict themselves to languages in $\mathbf{NP} \cap co\mathbf{NP}$, and (2) require a successful attacker to not only prove a false statement, but also provide an witness of the fact that the statement is false (which is possible since the language is in $co\mathbf{NP}$).

Let us remark that simply restricting to languages in $\mathbf{NP} \cap co\mathbf{NP}$ does not suffice for overcoming our lower-bound: Assuming the existence of one-way permutations, our impossibility result rules

out statistical NIZK with adaptive soundness also for $\mathbf{NP} \cap co\mathbf{NP}$. By the Goldreich-Levin Theorem [GL89], the existence of one-way permutations implies the existence of a hard-on-the-average language in $\mathbf{NP} \cap co\mathbf{NP}$ (see [Gol01]), which suffices for concluding the proof of Theorem 4.5.

5 Security of Non-interactive Non-malleable Commitments

Commitment schemes are used to enable a party, known as the *sender (or committer)*, to commit itself to a value while keeping it secret from the *receiver* (this property is called hiding). Furthermore, the commitment is binding in the sense that at a later stage when the commitment is opened, it is guaranteed that the "opening" can yield only a single value determined in the committing phase. In this work, we consider commitment schemes that are statistically-binding; namely, the binding property is required to hold against unbounded adversaries. In such a case, the hiding property can only hold against computationally bounded (non-uniform) adversaries.

Following [PR03, DDN00], we consider a generalized form of *tag-based commitment schemes* where, in addition to the security parameter, the committer and the receiver also receive a "tag"—a.k.a. the identity—id as common input. (The reason we consider tag-based commitments is that non-malleability is typically defined for such commitment schemes. Typically, we think of the tag as the identity of the committer.[28])

Since our unprovability result only refer to *two-round* commitment schemes (in which the commit-phase consists of a single message from the receiver that is followed by a single message from the committer), we only provide a definition of such two-round schemes.

Definition 5.1. *A pair of probabilistic algorithms* (Com, Rec) *whose running-time is polynomial in the length of its first input, is a* tag-based (two-round) commitment scheme *if the following two properties hold:*

- Statistical Binding: *There exists a negligible function* μ *such that for every* $k \in \mathbb{N}$*, every* $\mathsf{id} \in \{0,1\}^*$*, the following holds*

$$\Pr\left[\ \rho \leftarrow Rec(1^k, \mathsf{id})\ :\ \mathsf{bindfail}(\rho)\ \right] \leq \mu(k)$$

  where $\mathsf{bindfail}(\rho)$ *is true iff there exists* $v_0, v_1 \in \{0,1\}^k, r_0, r_1 \in \{0,1\}^*$ *such that*

$$Com_{r_0}(1^k, \mathsf{id}, v_0, \rho) = Com_{r_1}(1^k, \mathsf{id}, v_1, \rho)$$

- (Computational) Hiding: *The following two ensembles are computationally indistinguishable*

$$\Big\{Com(1^k, \mathsf{id}, v_0, \rho)\Big\}_{k\in\mathbb{N}, \mathsf{id}\in\{0,1\}^*, v_0, v_1\in\{0,1\}^k, \rho, z\in\{0,1\}^*}$$

$$\Big\{Com(1^k, \mathsf{id}, v_1, \rho)\Big\}_{k\in\mathbb{N}, \mathsf{id}\in\{0,1\}^*, v_0, v_1\in\{0,1\}^k, \rho, z\in\{0,1\}^*}$$

Note that in the ensembles considered in the definition of computational hiding, the input z (in the index of the ensemble) is not given to any of the algorithms in the experiments; this input is

[28] Non-malleable commitments are sometimes also considered in settings where players do not have identities. However, as shown in [PR05b], any non-malleable commitment that handles sufficiently long identities can be turned into a non-malleable commitment without identities (and any non-malleable commitment without identities can trivially be turned into one with identities). Since our goal is to prove lower bounds, we focus on the more general (relaxed) notion of non-malleability with respect to identities.

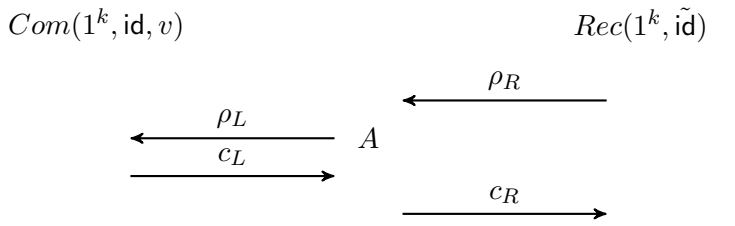

Figure 2: A man-in-the-middle execution.

there only to be provided as an input to the distinguisher (to ensure that indistinguishability holds w.r.t. non-uniform PPT distinguishers).

Let us turn to defining non-malleable commitments. We use a definition essentially due to [PR05b, PR05a], which follows the earlier definition from [DDN00]. Let $\Pi = (Com, Rec)$ be a tag-based commitment scheme, and let $k \in \mathbb{N}$ be a security parameter. Consider a man-in-the-middle adversary A that, on inputs 1^k, participates in one "left" and one "right" interaction, as depicted in Figure 2: In the left interaction, the man-in-the-middle adversary A interacts with a committer, receiving a commitment to the value $v \in \{0,1\}^k$ using an identity id of length $\ell(k)$; in the right interaction, A interacts with a receiver, attempting to commit a value $\tilde{v} \in \{0,1\}^k$, using an identity $\tilde{\mathsf{id}} \neq \mathsf{id}$ of length $\ell(k)$. (To simplify notation, we do not explicitly provide the identities id and $\tilde{\mathsf{id}}$ as an input to A; as we will consider a non-uniform PPT attacker A, it can always receive them as a non-uniform advice.)

If the right commitment is invalid, or undefined, its value is set to $\perp$. Let $\mathsf{MIM}^{\Pi}(1^k, A, v, \mathsf{id}, \tilde{\mathsf{id}})$ denote a random variable that describes the value $\tilde{v}$ in the above experiment.[29] More precisely, let $\mathsf{MIM}^{\Pi}(1^k, A, v, \mathsf{id}, \tilde{\mathsf{id}})$ denote the following probability distribution (the experiment generating it is also depicted in Figure 2):

$$\left\{ \rho_R \leftarrow Rec(1^k, \tilde{\mathsf{id}}); \rho_L \leftarrow A(1^k, \rho_R); c_L \leftarrow Com(1^k, \mathsf{id}, v, \rho_L); c_R \leftarrow A(1^k, \rho_R, \rho_L, c_L) \ : \ \mathsf{val}(\rho_R, c_R) \right\}$$

where $\mathsf{val}(\rho_R, c_R)$ is defined as $\tilde{v}$ if there exists some *unique* value $\tilde{v}$ such that $c_R \in Com(1^k, \tilde{\mathsf{id}}, \tilde{v}, \rho_R)$, and $\perp$ otherwise.

Definition 5.2. *A tag-based commitment scheme $\Pi = (Com, Rec)$ is said to be* non-malleable for identities of length *$\ell(\cdot)$ if for every non-uniform polynomial-time man-in-the-middle adversary A, the following ensembles are computationally indistinguishable.*

$$\left\{ \mathsf{MIM}^{\Pi}(1^k, A, v_0, \mathsf{id}, \tilde{\mathsf{id}}) \right\}_{k \in \mathbb{N}, v_0, v_1 \in \{0,1\}^k, \mathsf{id}, \tilde{\mathsf{id}} \in \{0,1\}^{\ell(k)}}$$
$$\left\{ \mathsf{MIM}^{\Pi}(1^k, A, v_1, \mathsf{id}, \tilde{\mathsf{id}}) \right\}_{k \in \mathbb{N}, v_0, v_1 \in \{0,1\}^k, \mathsf{id}, \tilde{\mathsf{id}} \in \{0,1\}^{\ell(k)}}$$

Our unprovability results will apply to very *weak* notions of non-malleability where we restrict to identities of length 1, and restrict to only two message, 0^k and 1^k. Let us start by explicitly defining what it means to break non-malleability *in this particular setting.*

[29] We note that more recent definitions of non-malleability [LPV08, LP09] additionally output the view of A in the above experiment. Since we are proving a lower-bound, we simply state the weaker definition.

Definition 5.3 (Breaking Non-malleability)**.** *Let* $\Pi = (Com, Rec)$ *be a tag-based commitment scheme. We say that a probabilistic algorithm* A breaks non-malleability of Π *with probability* $\epsilon(\cdot)$ *if for every* $k \in N$*, there exists some bit* $\mathsf{id} \in \{0,1\}$ *such that*

$$\left|\Pr\left[\mathsf{MIM}^{\Pi}(1^k, A, 1^k, \mathsf{id}, 1-\mathsf{id}) = 1^k\right] - \Pr\left[\mathsf{MIM}^{\Pi}(A, 0^k, \mathsf{id}, 1-\mathsf{id}) = 1^k\right]\right| > \epsilon(k)$$

We furthermore say that A breaks one-sided non-malleability of Π *with probability* $\epsilon(\cdot)$ *if the above holds w.r.t.* $\mathsf{id} = 0$*.*

We call a commitment weakly non-malleable (resp., one-sided non-malleable) if no attacks of the above kind exist. Let us turn to defining what it means to base weak non-malleability on an intractability assumption.

Definition 5.4 (Basing Weak Non-malleability on the Hardness of C)**.** *We say that* R *is a black-box reduction for basing weak non-malleability (resp., one-sided non-malleability) of* (Com, Rec) *on the hardness of* C *w.r.t threshold* $t(\cdot)$ *if* R *a security-parameter preserving black-box reduction and there exists a polynomial* $p(\cdot,\cdot)$ *such that for every probabilistic machine* A *that breaks non-malleability (resp., one-sided non-malleability) of* (Com, Rec) *with probability* $\epsilon(\cdot)$*, for every* $k \in \mathbb{N}$*,* R^A *breaks* C *w.r.t. threshold* t *with probability* $p(\epsilon(k), 1/k)$ *on input* 1^k*.*

5.1 One-sided Schemes

We are now ready to state our first unprovability result for non-malleable commitments, which focuses on one-sided non-malleability.

Theorem 5.5. *Let* (Com, Rec) *be a two-round tag-based commitment scheme, and let* (C, t) *be any efficient-challenger assumption. If there exists a probabilistic polynomial-time black-box reduction* R *for basing weak* one-sided *non-malleability of* (Com, Rec) *on the hardness of* C *w.r.t threshold* t*, then there exists a probabilistic polynomial-time machine* B *and a polynomial* $p'(\cdot)$ *such that for infinitely many* $k \in \mathbb{N}$*, machine* B *breaks* C *w.r.t. threshold* t *with probability* $\frac{1}{p'(k)}$ *on input* 1^k*.*

Before proceeding to the proof of Theorem 5.5, let us remark that Theorem 5.5 only applies to schemes (Com, Rec) that already satisfying the standard notions of binding and hiding of commitments (see Definition 5.1). While the hiding property of commitments follows from (standard) non-malleability, *one-sided* non-malleability only implies hiding for commitments using identity 0. For the proof of Theorem 5.5 we require hiding to hold also w.r.t. identity 1.[30] (Looking forward, in Theorem 5.9, which deals with *two-sided* non-malleability, it will suffice to just assume that (Com, Rec) satisfies binding.) Let us also note that binding of (Com, Rec) does not follow from non-malleability.[31]

Proof of Theorem 5.5. The proof follows the same overall structure as the proof of Theorem 4.5. Assume there exists a two-round tag-based commetment scheme (Com, Rec) and assume there exists a black-box reduction R for basing weak one-sided non-malleability of (Com, Rec) on the hardness of C w.r.t threshold t. That is, there exists a polynomial $p(\cdot,\cdot)$ such that for every probabilistic machine A that breaks weak one-sided non-malleability of (Com, Rec) with probability

[30] This is inherent. There are binding (but not hiding) schemes (Com, Rec) that satisfy one-sided non-malleability: Simply consider a commitment that is hiding w.r.t. identity 0 but fully reveals the value committed to w.r.t identity 1. Non-malleability directly follows from the hiding property w.r.t. identity 0.

[31] If (Com, Rec) is *never* binding (e.g., $Com(v) = 0$), then non-malleability holds trivially.

$\epsilon(\cdot)$, for every $k \in \mathbb{N}$, R^A breaks C w.r.t. threshold t with probability $p(\epsilon(k), 1/k)$ on input 1^k; that is,

$$\Pr\left[\langle R^A, C\rangle(1^k) = 1\right] \geq t(k) + p(\epsilon(k), 1/k). \tag{3}$$

As in the proof of Theorem 4.5, our proof strategy proceeds in two steps:

- We first devise a particular *inefficient* attacker A that breaks weak one-sided non-malleability of (Com, Rec) with *overwhelming probability*. By Equation 3, it then follows that there exists a positive polynomial $\hat{p}(\cdot)$ such that R^A breaks C with probability $t(k) + \frac{1}{\hat{p}(k)}$ for infinitely many k.
- We then show how to *indistinguishably* simulate A by an efficient emulator $\tilde{A}$ to conclude that $R^{\tilde{A}}$ breaks C with probability $t(k) + \frac{1}{\hat{p}(k)} - \mu(k)$ for infinitely many k, where $\mu(k)$ is a negligible function, and thus prove the theorem.

Constructing the Attacker A. We define the attacker A as follows. Let $\mathsf{id} = 0$.

- On input $1^k, \rho_R$, the attacker A samples $\rho_L \leftarrow Rec(1^k, \mathsf{id})$ and outputs ρ_L.
- On input $1^k, \rho_R, \rho_L, c_L$, the attacker A proceeds as follows.[32]
  - **Inverting Com:** A checks whether there exists a *unique* value v for which there exists some string $r \in \{0,1\}^*$ such that $c_L = Com_r(1^k, \mathsf{id}, v_0, \rho_L)$. If such a unique value does not exists, it sets $v = 0^k$. *Note that this step is not efficient.*
  - **Generate commitment:** A generates $c_R \leftarrow Com(1^k, 1 - \mathsf{id}, v, \rho_R)$ and outputs c_R.
- (On all other inputs, the attacker A its simply outputs $\perp$.)

We now show the following claim (which is an analog of Corollary 4.9 in the proof of Theorem 4.5).

Claim 5.6. *There exists a negligible function $\mu(\cdot)$ such that A breaks one-sided non-malleability of (Com, Rec) with probability $1 - \mu(\cdot)$.*

Proof. It follows directly from the binding property of (Com, Rec) (recall that by hypothesis (Com, Rec) is a tag-based two-round commitment scheme) that there exists a negligible function μ' such that except with probability $\mu'(k)$, the commitment message c_L sent in the left interaction in the experiment $\mathsf{MIM}^{\Pi}(1^k, A, v, \mathsf{id}, 1 - \mathsf{id})$ (where $v \in \{0,1\}^k$) determines the unique value v. Whenever this happens A will recover this unique value and honestly produce a commitment to it (using identity $1 - \mathsf{id}$) in the right interaction. By the binding property of (Com, Rec), except with probability $\mu'(k)$, the value committed to the right interaction will also be uniquely defined as v. It follows that A breaks one-sided non-malleability with probability $1 - 2\mu'(\cdot)$. □

As a consequence of Claim 5.6 and the fact that R is a good reduction (i.e., Equation 3), there exists a positive polynomial $\hat{p}(\cdot)$ such that R^A breaks C with probability $t(k) + \frac{1}{\hat{p}(k)}$ for infinitely many k; that is,

$$\Pr\left[\langle R^A, C\rangle(1^k) = 1\right] \geq t(k) + \frac{1}{\hat{p}(k)} \tag{4}$$

[32] Recall that our notion of an oracle machine requires the oracle to use *fresh* randomness on every invocation; this simplifies notation (and only makes our impossibility result stronger). A thus needs to receive ρ_L as input (even though it previously generated it). Note, however, that if the attacker had been deterministic (and as we shall see later on, we also rule out reductions that only work for deterministic attackers), clearly ρ_L does not need to be provided as an input to the attacker (as it can simply recompute it).

Note, however, that R^A is not necessarily implementable in polynomial time. We now turn to show how to indistinguishably simulating A by an *efficient* emulator $\tilde{A}$.

Constructing the Emulator $\tilde{A}$. The emulator $\tilde{A}$ proceeds exactly as A except that instead of performing the "Inverting Com" step, $\tilde{A}$ simply sets $v = 0^k$.

The following lemma (which can be viewed as analog of Corollary 4.8 in the proof of Theorem 4.5) shows that $\tilde{A}$ is a good emulator for A for *all* inputs x; in particular, this holds for inputs of the form $(1^k, \rho_R)$ (and as such $\tilde{A}$ correctly emulates the first part of the algorithm A), and inputs of form $(1^k, \rho_L, rho_R, c_L)$ (and as such $\tilde{A}$ correctly emulates the second part of the algorithm A).

Lemma 5.7. *The following two ensembles are computationally indistinguishable*

$$\left\{A(1^k, x)\right\}_{k\in\mathbb{N},x,z\in\{0,1\}^*}$$

$$\left\{\tilde{A}(1^k, x)\right\}_{k\in\mathbb{N},x,z\in\{0,1\}^*}$$

Proof. If x is not of the form (ρ_L, rho_R, c_L), the output of $A(1^k, x)$ is identically distributed to the output of $\tilde{A}(1^k, x)$. When x is of the form (ρ_L, ρ_R, c_L), indistinguishability follows directly from the hiding property of (Com, Rec) w.r.t. the tag $\tilde{\mathsf{id}} = 1$. (Note that it is important that computational hiding of commitments holds even w.r.t. auxiliary inputs, as the additional inputs ρ_L, c_L need to be interpreted as an auxilary input to a commitment distinguisher; recall that by hypothesis (Com, Rec) is a tag-based two-round commitment scheme satisfying such a hiding property (see Definition 5.1). □

Given Lemma 5.7, we can finally show that $\tilde{A}$ is a "good" emulator for A, even if A is repeatedly invoked by R (in an interaction with C), and even if R queries A on illegal transcripts. The following claim follows in exactly the same way as Claim 4.10 in the proof of Theorem 4.5 (by using Lemma 5.7).

Claim 5.8. *For every efficient C and R, there exists a negligible function μ such that for every $k \in \mathbb{N}$,*

$$\left|\Pr\left[\langle R^{\tilde{A}}, C\rangle(1^k) = 1\right] - \Pr\left[\langle R^A, C\rangle(1^k) = 1\right]\right| \leq \mu(k).$$

The proof of the Theorem 5.5 is concluded by combining Equation 4 with Claim 5.8. □

We remark that Theorem 5.5 is tight in the sense that we cannot hope to rule out also super-polynomial-time reductions for one-sided non-malleability: Liskov et al [LLM$^+$01] present a construction of such commitments assuming the existence of one-way permutations with subexponential security (relying on the "complexity-leveraging" idea of [CGGM00]) and using a subexponential security reduction.

5.2 General Schemes and Super-polynomial Reductions

Let us now turn to ruling out also super polynomial-time reductions for "two-sided" non-malleability (i.e., where identity of the left interaction can be either 0 or 1).

Theorem 5.9. *Let (Com, Rec) be a two-round tag-based commitment scheme, and let (C, t) be a $T(\cdot)$-size challenger intractability assumption for an arbitrary function $T(\cdot)$. If there exists a $T(\cdot)$-size randomized black-box reduction R for basing weak non-malleability of Π on the hardness of C w.r.t threshold t, then there exists a poly$(T(\cdot))$-sized attacker B and a polynomial $p'(\cdot)$ such that for infinitely many $k \in \mathbb{N}$, machine B breaks C w.r.t. threshold t with probability $\frac{1}{p'(k)}$ on input 1^k.*

Proof. Consider the attacker A and emulator $\tilde{A}$ from the proof of Theorem 5.5. We consider two cases:

- **Case 1:** $R^{\tilde{A}}$ breaks C w.r.t. threshold t with inverse polynomial probability for infinitely many k.
- **Case 2:** There exists a negligible function μ such that $R^{\tilde{A}}$ breaks C w.r.t. threshold t with probability at most $\mu(k)$ for every $k \in \mathbb{N}$.

If Case 1 happens, we are done. Let us now consider Case 2. As (implicitly) shown in the proof of Theorem 5.5 (in Claim 5.8 which in turn uses Lemma 5.7), in this case, R together with C and using a polynomial-size advice string may distinguish commitments to 0^k and 1^k using identity $\tilde{\mathsf{id}} = 1$ with inverse polynomial probability for infinitely many k. (In the proof of Theorem 5.5, this lead to a contradiction since R and C were polynomial-time. Here, however, both are *super-polynomial-time*, so there is no contradiction with hiding property of (Com, Rec).) More precisely, there exists a poly$(T(\cdot))$-sized *boolean* circuit D, a polynomial $p(\cdot)$ and an infinite sequence of strings $\rho^1, \rho^2, \ldots$ such that for infinitely many $k \in \mathbb{N}$, D distinguishes between $(1^k, Com(1^k, 1, 1^k, \rho^k))$ and $(1^k, Com(1^k, 1, 0^k, \rho^k))$ with probability $\frac{1}{p(k)}$; that is,

$$\left|\Pr\left[D(1^k, Com(1^k, 1, 1^k, \rho^k)) = 1\right] - \Pr\left[D(1^k, Com(1^k, 1, 0^k, \rho^k)) = 1\right]\right| \geq \frac{1}{p(k)} \tag{5}$$

Explicitly, the distinguisher $D(1^k, c)$ outputs

$$\left\langle R^{\tilde{A}}_{(\tau^k_R, c)}, C_{\tau^k_C} \right\rangle (1^k)$$

where $R_{(\tau^k_R, c)}$ denotes the machine R whose initial view (i.e., its random tape, messages from C and answers to its initial oracle queries) is fixed to (τ^k_R, c), whereas $C_{\tau^k_C}$ denotes the machine C whose initial view (i.e., random tape and messages from R) is fixed to τ^k_C, and τ^k_R, τ^k_C are the "prefix views" of respectively R and C obtained from the hybrid argument in the proof of Claim 5.8; ρ^k is the last message sent by R in the prefix view τ^k_R.

We now show how to use the sequence $\rho^1, \rho^2, \ldots$ and the distinguisher D (having τ^k_R, τ^k_C hard-coded) to construct a new attacker A' that uses *identity 1 on the left* and 0 on the right (recall that our earlier attacker A used *identity 0 on the left* and 1 on the right). A' proceeds as follows:

- $A'(1^k, \rho_R)$ outputs $\rho_L = \rho^k$;
- $A'(1^k, \rho_R, \rho_L, c_L)$ lets $b \leftarrow D(1^k, c_L)$ and outputs $c \leftarrow Com(1^k, 0, b^k, \rho_R)$. (It may seem a bit surprising that we here ignore the input ρ_L. The point is that we only require this attacker to work in an "honest" man-in-the-middle execution (as in the definition of MIM), and in such an execution $\rho_L = \rho^k$, and thus D will be a "good" distinguisher for the commitment received.)
- (On all other inputs, A' simply outputs $\perp$.)

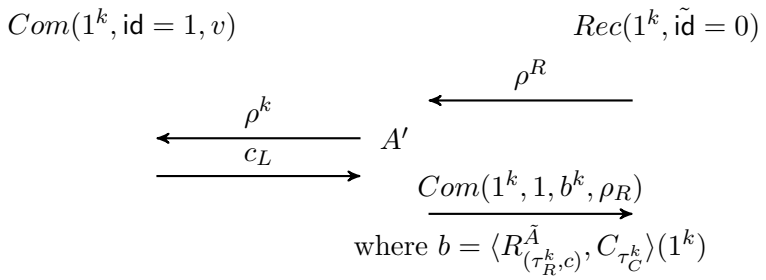

Figure 3: The second attacker A' receiving $\tau^k_R, \tau^k_C, \rho^k$ as non-uniform advice.

See Figure 3 for a pictoral representation of A' (with the description of D expanded out).

The following claim shows that A' breaks the non-malleability of (Com, Rec).

Claim 5.10. *There exists a polynomial $p'(\cdot)$ such that A' breaks non-malleability of (Com, Rec) with probability $\frac{1}{p'(k)}$ for infinitely many k.*

Proof. By the binding property of (Com, Rec) w.r.t. the identity 0, and the construction of A', there exists some negligible function $\mu(\cdot)$ such that the following inequalities hold:

$$\left|\Pr\left[\mathsf{MIM}^{\Pi}(1^k, A', 1^k, 1, 0) = 1^k\right] - \Pr\left[D(1^k, Com(1^k, 1, 1^k, \rho^k)) = 1\right]\right| \leq \mu(k)$$

$$\left|\Pr\left[\mathsf{MIM}^{\Pi}(1^k, A', 0^k, 1, 0) = 1^k\right] - \Pr\left[D(1^k, Com(1^k, 1, 0^k, \rho^k)) = 1\right]\right| \leq \mu(k)$$

Combing these inequalities with Equation 5, we have

$$\left|\Pr\left[\mathsf{MIM}^{\Pi}(1^k, A', 1^k, 1, 0) = 1^k\right] - \Pr\left[\mathsf{MIM}^{\Pi}(A', 0^k, 1, 0) = 1^k\right]\right| \geq \frac{1}{p(k)} - 2\mu(k)$$

which concludes the claim. □

As a consequence of Claim 5.10 and the fact that R is a good reduction (i.e., Equation 3), there exists a positive polynomial $\hat{p}(\cdot)$ such that $R^{A'}$ breaks C with probability $t(k) + \frac{1}{\hat{p}(k)}$ for infinitely many k. Since R is of circuit-size bounded by $T(\cdot)$ (and thus can make at most $T(\cdot)$ queries to its oracle) and A' is of circuit-size bounded by $\text{poly}(T(\cdot))$, it follows that $R^{A'}$ can be described by a circuit of size $\text{poly}(T(\cdot))$, which concludes the proof of the theorem. (We remark that we here rely on the fact that R is security-parameter preserving and thus, even if it is super-polynomial time, only queries its oracle on the same (or a polynomially related) security parameter; otherwise the running time of $R^{A'}$ could be $\text{poly}(T(T(\cdot))$.) □

A Remark on Deterministic Attackers and Reductions with Non-uniform Advice. Just as the proof of Theorem 4.5, the proofs of Theorem 5.5 and Theorem 5.9 readily extend to rule out reductions that only work with deterministic attackers, and, relying on the techniques from [CLMP13], also reductions with non-uniform advice.

6 Acknowledgements

I am grateful to Kai-min Chung and Mohammad Mahmoody for many helpful comments and definitional discussions. I am extremely grateful to Oded Goldreich for his incredibly detailed, insightful and helpful comments.

References

[AF07] Masayuki Abe and Serge Fehr. Perfect NIZK with adaptive soundness. In *TCC*, pages 118–136, 2007.

[AGGM06] Adi Akavia, Oded Goldreich, Shafi Goldwasser, and Dana Moshkovitz. On basing one-way functions on NP-hardness. In *STOC '06*, pages 701–710, 2006.

[Bar01] Boaz Barak. How to go beyond the black-box simulation barrier. In *FOCS '01*, volume 0, pages 106–115, 2001.

[Bar02] Boaz Barak. Constant-round coin-tossing with a man in the middle or realizing the shared random string model. In *FOCS '02: Proceedings of the 43rd Symposium on Foundations of Computer Science*, pages 345–355, Washington, DC, USA, 2002. IEEE Computer Society.

[BFM88] Manuel Blum, Paul Feldman, and Silvio Micali. Non-interactive zero-knowledge and its applications (extended abstract). In *STOC*, pages 103–112, 1988.

[BM84] Manuel Blum and Silvio Micali. How to generate cryptographically strong sequences of pseudo-random bits. *SIAM Journal on Computing*, 13(4):850–864, 1984.

[BMV08] Emmanuel Bresson, Jean Monnerat, and Damien Vergnaud. Separation results on the "one-more" computational problems. In *CT-RSA*, pages 71–87, 2008.

[BNPS03] Mihir Bellare, Chanathip Namprempre, David Pointcheval, and Michael Semanko. The one-more-rsa-inversion problems and the security of chaum's blind signature scheme. *J. Cryptology*, 16(3):185–215, 2003.

[BP02] Mihir Bellare and Adriana Palacio. Gq and schnorr identification schemes: Proofs of security against impersonation under active and concurrent attacks. In *CRYPTO*, pages 162–177, 2002.

[BR93] Mihir Bellare and Phillip Rogaway. Random oracles are practical: A paradigm for designing efficient protocols. In *ACM Conference on Computer and Communications Security*, pages 62–73, 1993.

[Bra83] Gilles Brassard. Relativized cryptography. *IEEE Transactions on Information Theory*, 29(6):877–893, 1983.

[BT03] Andrej Bogdanov and Luca Trevisan. On worst-case to average-case reductions for np problems. In *FOCS*, pages 308–317, 2003.

[BV98] Dan Boneh and Ramarathnam Venkatesan. Breaking rsa may not be equivalent to factoring. In *EUROCRYPT*, pages 59–71, 1998.

[BY96] Mihir Bellare and Moti Yung. Certifying permutations: Noninteractive zero-knowledge based on any trapdoor permutation. *J. Cryptology*, 9(3):149–166, 1996.

[CGGM00] Ran Canetti, Oded Goldreich, Shafi Goldwasser, and Silvio Micali. Resettable zero-knowledge (extended abstract). In *STOC '00*, pages 235–244, 2000.

[CGH04] Ran Canetti, Oded Goldreich, and Shai Halevi. The random oracle methodology, revisited. *J. ACM*, 51(4):557–594, 2004.

[CIO98] Giovanni Di Crescenzo, Yuval Ishai, and Rafail Ostrovsky. Non-interactive and non-malleable commitment. In *STOC*, pages 141–150, 1998.

[CLMP13] Kai-min Chung, Huijia Lin, Mohammad Mahmoody, and Rafael Pass. On the power of non-uniform proof of security. In *ITCS'13*, 2013.

[CMP09] Kai-min Chung, Mohammad Mahmoody, and Rafael Pass. A note on black-box reductions. Manuscript, 2009.

[Dam91] Ivan Damgård. Towards practical public key systems secure against chosen ciphertext attacks. In *CRYPTO*, pages 445–456, 1991.

[DDN00] Danny Dolev, Cynthia Dwork, and Moni Naor. Nonmalleable cryptography. *SIAM Journal on Computing*, 30(2):391–437, 2000.

[DOP05] Yevgeniy Dodis, Roberto Oliveira, and Krzysztof Pietrzak. On the generic insecurity of the full domain hash. In *CRYPTO*, pages 449–466, 2005.

[FF93] Joan Feigenbaum and Lance Fortnow. Random-self-reducibility of complete sets. *SIAM Journal on Computing*, 22(5):994–1005, 1993.

[FLS90] Uriel Feige, Dror Lapidot, and Adi Shamir. Multiple non-interactive zero knowledge proofs based on a single random string. In *FOCS '90*, pages 308–317, 1990.

[FS87] Amos Fiat and Adi Shamir. How to prove yourself: practical solutions to identification and signature problems. In *Proceedings on Advances in cryptology—CRYPTO '86*, pages 186–194, London, UK, 1987. Springer-Verlag.

[FS10] Marc Fischlin and Dominique Schröder. On the impossibility of three-move blind signature schemes. In *EUROCRYPT*, pages 197–215, 2010.

[GGM86] Oded Goldreich, Shafi Goldwasser, and Silvio Micali. How to construct random functions. *J. ACM*, 33(4):792–807, 1986.

[GK96] Oded Goldreich and Hugo Krawczyk. On the composition of zero-knowledge proof systems. *SIAM Journal on Computing*, 25(1):169–192, 1996.

[GK03] Shafi Goldwasser and Yael Tauman Kalai. On the (in)security of the fiat-shamir paradigm. In *FOCS '03*, pages 102–111, 2003.

[GL89] Oded Goldreich and Leonid A. Levin. A hard-core predicate for all one-way functions. In *STOC*, pages 25–32, 1989.

[GM84] Shafi Goldwasser and Silvio Micali. Probabilistic encryption. *J. Comput. Syst. Sci.*, 28(2):270–299, 1984.

[GMR88] Shafi Goldwasser, Silvio Micali, and Ronald L. Rivest. A digital signature scheme secure against adaptive chosen-message attacks. *SIAM J. Comput.*, 17(2):281–308, 1988.

[GMR89] Shafi Goldwasser, Silvio Micali, and Charles Rackoff. The knowledge complexity of interactive proof systems. *SIAM Journal on Computing*, 18(1):186–208, 1989.

[GO94] Oded Goldreich and Yair Oren. Definitions and properties of zero-knowledge proof systems. *Journal of Cryptology*, 7:1–32, 1994.

[Gol01] Oded Goldreich. *Foundations of Cryptography — Basic Tools*. Cambridge University Press, 2001.

[GOS06] Jens Groth, Rafail Ostrovsky, and Amit Sahai. Perfect non-interactive zero knowledge for np. In *EUROCRYPT*, pages 339–358, 2006.

[Goy11] Vipul Goyal. Constant round non-malleable protocols using one way functions. In *STOC*, pages 695–704, 2011.

[GR13] Oded Goldreich and Ron D Rothblum. Enhancements of trapdoor permutations. *Journal of cryptology*, 26(3):484–512, 2013.

[GW11] Craig Gentry and Daniel Wichs. Separating succinct non-interactive arguments from all falsifiable assumptions. In *STOC*, pages 99–108, 2011.

[HH09] Iftach Haitner and Thomas Holenstein. On the (im)possibility of key dependent encryption. In *TCC*, pages 202–219, 2009.

[HILL99] Johan Håstad, Russell Impagliazzo, Leonid Levin, and Michael Luby. A pseudorandom generator from any one-way function. *SIAM Journal on Computing*, 28:12–24, 1999.

[HPWP10] Johan Håstad, Rafael Pass, Douglas Wikström, and Krzysztof Pietrzak. An efficient parallel repetition theorem. In *TCC'10*, pages 1–18, 2010.

[HRS09] Iftach Haitner, Alon Rosen, and Ronen Shaltiel. On the (im)possibility of arthur-merlin witness hiding protocols. In *TCC '09*, pages 220–237, 2009.

[IJK07] Russell Impagliazzo, Ragesh Jaiswal, and Valentine Kabanets. Chernoff-type direct product theorems. In *CRYPTO '07*, pages 500–516, 2007.

[IR88] Russell Impagliazzo and Steven Rudich. Limits on the provable consequences of one-way permutations. In *CRYPTO '88*, pages 8–26, 1988.

[LLM+01] Moses Liskov, Anna Lysyanskaya, Silvio Micali, Leonid Reyzin, and Adam Smith. Mutually independent commitments. In *ASIACRYPT*, pages 385–401, 2001.

[LP09] Huijia Lin and Rafael Pass. Non-malleability amplification. In *STOC '09*, pages 189–198, 2009.

[LP11] Huijia Lin and Rafael Pass. Constant-round non-malleable commitments from any one-way function. In *STOC*, pages 705–714, 2011.

[LPV08] Huijia Lin, Rafael Pass, and Muthuramakrishnan Venkitasubramaniam. Concurrent non-malleable commitments from any one-way function. In *TCC '08*, pages 571–588, 2008.

[Nao03] Moni Naor. On cryptographic assumptions and challenges. In *CRYPTO*, pages 96–109, 2003.

[NY89] Moni Naor and Moti Yung. Universal one-way hash functions and their cryptographic applications. In *STOC*, pages 33–43, 1989.

[Ost91] Rafail Ostrovsky. One-way functions, hard on average problems, and statistical zero-knowledge proofs. In *Structure in Complexity Theory Conference*, pages 133–138, 1991.

[OW93] Rafail Ostrovsky and Avi Wigderson. One-way functions are essential for non-trivial zero-knowledge. In *Theory and Computing Systems, 1993*, pages 3–17, 1993.

[Pas03] Rafael Pass. On deniability in the common reference string and random oracle model. In *CRYPTO*, pages 316–337, 2003.

[Pas06] Rafael Pass. Parallel repetition of zero-knowledge proofs and the possibility of basing cryptography on np-hardness. In *IEEE Conference on Computational Complexity*, pages 96–110, 2006.

[Pas11] Rafael Pass. Limits of provable security from standard assumptions. In *STOC*, pages 109–118, 2011.

[PPV08] Omkant Pandey, Rafael Pass, and Vinod Vaikuntanathan. Adaptive one-way functions and applications. In *CRYPTO 2008: Proceedings of the 28th Annual conference on Cryptology*, pages 57–74, Berlin, Heidelberg, 2008. Springer-Verlag.

[PR03] Rafael Pass and Alon Rosen. Bounded-concurrent secure two-party computation in a constant number of rounds. In *FOCS '03*, pages 404–, 2003.

[PR05a] Rafael Pass and Alon Rosen. Concurrent non-malleable commitments. In *FOCS '05*, pages 563–572, 2005.

[PR05b] Rafael Pass and Alon Rosen. New and improved constructions of non-malleable cryptographic protocols. In *STOC '05*, pages 533–542, 2005.

[PS05] Rafael Pass and Abhi Shelat. Unconditional characterizations of non-interactive zero-knowledge. In *CRYPTO*, pages 118–134, 2005.

[PTV11] Rafael Pass, Wei-Lung Dustin Tseng, and Muthuramakrishnan Venkitasubramaniam. Towards non-black-box lower bounds in cryptography. In *TCC*, pages 579–596, 2011.

[PW10] Rafael Pass and Hoeteck Wee. Constant-round non-malleable commitment from strong one-way functions. In *Eurocrypt '10*, 2010.

[Rom90] John Rompel. One-way functions are necessary and sufficient for secure signatures. In *STOC*, pages 387–394, 1990.

[RTV04] Omer Reingold, Luca Trevisan, and Salil P. Vadhan. Notions of reducibility between cryptographic primitives. In *TCC*, pages 1–20, 2004.

[RV10] Guy N. Rothblum and Salil P. Vadhan. Are pcps inherent in efficient arguments? *Computational Complexity*, 19(2):265–304, 2010.

[Wee10] Hoeteck Wee. Black-box, round-efficient secure computation via non-malleability amplification. In *FOCS*, pages 531–540, 2010.