From Weak to Strong Zero-Knowledge and Applications

Kai-Min Chung
Cornell University
chung@cs.cornell.edu

Edward Lui
Cornell University
luied@cs.cornell.edu

Rafael Pass
Cornell University
rafael@cs.cornell.edu

May 7, 2013

Abstract

The notion of *zero-knowledge* [GMR85] is formalized by requiring that for every malicious efficient verifier V^*, there exists an efficient simulator S that can reconstruct the view of V^* in a true interaction with the prover, in a way that is indistinguishable to *every* polynomial-time distinguisher. *Weak zero-knowledge* weakens this notions by switching the order of the quantifiers and only requires that for every distinguisher D, there exists a (potentially different) simulator S_D.

In this paper we consider various notions of zero-knowledge, and investigate whether their weak variants are equivalent to their strong variants. Although we show (under complexity assumption) that for the standard notion of zero-knowledge, its weak and strong counterparts are not equivalent, for meaningful variants of the standard notion, the weak and strong counterparts are indeed equivalent. Towards showing these equivalences, we introduce new non-black-box simulation techniques permitting us, for instance, to demonstrate that the classical 2-round graph non-isomorphism protocol of Goldreich-Micali-Wigderson [GMW91] satisfies a "distributional" variant of zero-knowledge.

Our equivalence theorem has other applications beyond the notion of zero-knowledge. For instance, it directly implies the *dense model theorem* of Reingold et al (STOC '08), and the leakage lemma of Gentry-Wichs (STOC '11), and provides a modular and arguably simpler proof of these results (while at the same time recasting these result in the language of zero-knowledge).

1 Introduction

The notion of *zero-knowledge*, and the *simulation-paradigm* used to define it, is of fundamental importance in modern cryptography—most definitions of protocol security rely on it. In a zero-knowledge protocol, a prover P can convince a verifier V of the validity of some mathematical statement $x \in L$, while revealing "zero (additional) knowledge" to V. This zero-knowledge property is formalized by requiring that for every potentially malicious efficient verifier V^*, there exists an efficient simulator S that, without talking to P, is able to "indistinguishably reconstruct" the view of V^* in a true interaction with P. The traditional way of defining what it means to "indistinguishably reconstruct" is to require that the output of S cannot be distinguished (with more than negligible probability) from the true view of V^* by *any* efficient distinguisher D; that is, we have a *universal* simulator that works for *all* distinguishers D.

A seemingly weaker way to define the zero-knowledge property is to require that for every distinguisher D, there exists a "distinguisher-dependent" simulator S_D such that the output of S_D cannot be distinguished from the true view of V^* by the *particular* distinguisher D; following [DNRS03], we refer to this weaker notion of zero-knowledge as *weak zero-knowledge*.

The main question addressed in this paper is whether this switch in the order of the quantifiers yields an equivalent notion. More specifically, we consider various notions of zero-knowledge, and investigate whether their weak (distinguisher-dependent simulator) variants are equivalent to their strong (universal simulator) variants. Towards addressing this question, we introduce new non-black-box simulation techniques permitting us, for instance, to demonstrate that the classical 2-round graph non-isomorphism protocol of Goldreich-Micali-Wigderson [GMW91] satisfies a "distributional" variant of zero-knowledge. Our results also reveal deep connections between the notion of zero-knowledge and the *dense model theorem* of Reingold et al [RTTV08] (which in turn is related to questions such as the XOR Lemma [Yao82] and Szemeredi's regularity lemma [FK99]; see [TTV09] for more details).

1.1 From Weak to Strong Zero-Knowledge

Our first result shows that (under plausible complexity-theoretic assumptions) for the standard definition of zero-knowledge, weak zero-knowledge is a strictly weaker requirement than (strong) zero-knowledge.

Theorem 1 (Informally stated). *Assume the existence of "timed commitments" and "timed one-way permutations". Then, there exists an interactive proof for a language $L \in \mathsf{NP}$ that is weak zero-knowledge but not (strong) zero-knowledge.*

Motivated by this separation, we turn to consider relaxed notions of zero-knowledge. We first consider a concrete security variant of the notion of zero-knowledge. Roughly speaking, we call a protocol (t, ϵ)-zero-knowledge if the zero-knowledge property holds with respect to all $t(n)$-time bounded distinguishers (as opposed to all polynomial-time distinguishers), and we require that the distinguishability gap is bounded by $\epsilon(n)$ (as opposed to being negligible), where n is the length of the statement x being proved. Weak (t, ϵ)-zero-knowledge is define analogously (by again switching the order of the quantifiers).

Note that if (P, V) is (t, ϵ)-zero-knowledge (resp. weak (t, ϵ)-zero-knowledge) for some super-polynomial function t and some negligible function ϵ, then (P, V) is zero-knowledge (resp. weak zero-knowledge) in the classic sense. We here consider a slightly relaxed notion where we only require (P, V) to be (t, ϵ)-zero-knowledge for all polynomials t and all inverse polynomials ϵ. (Note that this is weaker than the standard definition of zero-knowledge since now the running-time of the simulator may depend on the bounds t and ϵ.) Perhaps surprisingly, we show that for this relaxed notion of zero-knowledge, the weak and strong versions lead to an equivalent definition.

Theorem 2 (Informally stated)**.** *If an interactive proof (P, V) is weak (t, ϵ)-zero knowledge for every polynomial t and every inverse polynomial ϵ, then (P, V) is also (t', ϵ')-zero knowledge for every polynomial t' and every inverse polynomial ϵ'.*

We highlight that the "universal" simulator S constructed in the proof of Theorem 2 makes use of the malicious verifier V^* in a non-black-box way. On a very high-level (and significantly oversimplifying), the idea behind Theorem 2 is to rely on Von Neumann's minimax theorem to obtain the universal simulator from the "distinguisher-dependent" simulators; the non-black-box nature of the universal simulator comes from the fact that defining the "utility function" we use with the minimax theorem requires knowing the auxiliary inputs received by V^*, and thus we make non-black-box use of V^*.

Implementing this approach becomes quite non-trivial since we require the existence of a *uniform* polynomial-time simulator for every uniform polynomial-time verifier—the minimax theorem only guarantees the existence of a *distribution* over polynomial-time machines that simulates the view of the verifier, but it is not clear if this distribution can be computed in uniform polynomial time. We overcome this issue by instead relying on a multiplicative weights algorithm to appropriately implement an approximate minimax strategy; see Section 1.4 for more details.

1.2 From Super-Weak to Strong Distributional Zero-Knowledge

Note that although in the definition of weak zero-knowledge the simulator may depend on the distinguisher, we still require that the probability that the distinguisher outputs 1 when given the output of the simulator is *close* to the probability that the distinguisher outputs 1 when given the true view of the malicious verifier V^*. An even weaker condition (considered in [HP10]) only requires that the simulator manages to make the distinguisher output 1 with *at least as high probability* (minus some "small" gap) as the probability that the distinguisher outputs 1 when given a true view of V^*. That is, we only consider "one-sided" indistinguishability. We refer to such a zero-knowledge property as *super-weak zero-knowledge*.

It is not hard to see that super-weak (t, ϵ)-zero-knowledge is not equivalent to weak (t, ϵ)-zero-knowledge (see Appendix D for the proof). Thus, we here consider an alternative "distributional" notion of zero-knowledge (a la [Gol93]) where indistinguishability of the simulation is only required for any distribution of statements (and auxiliary inputs). Additionally, we here model both the distinguisher and the simulator as non-uniform polynomial-time algorithms (as opposed to uniform ones). (The combination of these variants was previously considered by [DNRS03].[1]) We refer to such a notion of zero-knowledge as *distributional zero-knowledge*, and analogously define *distributional (t, ϵ)-zero-knowledge* as well as *weak (resp. super-weak) distributional (t, ϵ)-zero-knowledge.* Roughly speaking, distributional zero-knowledge captures the intuition that proofs of "random" statements do not provide the verifier with any new knowledge (beyond the statement proved). Perhaps surprisingly, we show that super-weak distributional (t, ϵ)-zero-knowledge is equivalent to (strong) distributional (t, ϵ)-zero-knowledge if we consider all polynomials t and all inverse polynomials ϵ.

Theorem 3 (Informally stated)**.** *If an interactive proof (P, V) is super-weak distributional (t, ϵ)-zero-knowledge for every polynomial t and every inverse polynomial ϵ, then (P, V) is also distributional (t', ϵ')-zero knowledge for every polynomial t' and every inverse polynomial ϵ'.*

[1] More specifically, the notion of "ultra-weak zero-knowledge" of [DNRS03] considers both of these relaxations, but relaxes the notion even further.

In contrast to Theorem 2, the proof of Theorem 3 follows from a rather direct use of the minimax theorem; see Section 1.4 for more details. We also show that any protocol where the prover is "laconic" [GVW01]—that is, it sends only $O(\log n)$ bits in total, is super-weak (distributional) zero-knowledge; combining this result with Theorem 3 thus yields the following theorem.

Theorem 4 (Informally stated)**.** *Let (P, V) be an interactive proof with a laconic prover for a language L. Then (P, V) is distributional (t, ϵ)-zero-knowledge for every polynomial t and every inverse polynomial ϵ.*

Given Theorem 3, the proof of Theorem 4 is very straight-forward: to show that laconic proofs are super-weak zero-knowledge, have the simulator simply enumerate all possible prover messages and keep the one that the distinguisher "likes the most" (i.e., makes the distinguisher output 1 with as high probability as possible); note that we here rely crucially on the fact that we only need to achieve "one-sided" indistinguishability.

Theorem 4 may seem contradictory. An interactive proof with a laconic prover (i.e., with small prover communication complexity) can reveal, say, the first $\log n$ bits of the witness w to the statement x proved, yet Theorem 4 states that such a protocol satisfies a notion of zero-knowledge. But if we leak something specific about the witness, how can we expect the protocol to be "zero-knowledge"? The key point here is that (as shown in Theorem 4), for *random* statements x, the information revealed about the witness can actually be efficiently generated. In other words, the *whole* process where the prover first picks the statement (at random), and then provides the proof, is zero-knowledge.

Despite the simplicity of the proof of Theorem 4, it has many (in our eyes) intriguing corollaries. The first one is that the classic two-round graph non-isomorphism protocol of [GMW91] (which is only known to be "honest-verifier" zero-knowledge) is distributional (t, ϵ)-zero-knowledge for every polynomial t and every inverse polynomial ϵ.[2] In fact, by the complete problem for SZK [SV03], we can show that every language in SZK has a 2-round interactive proof that is distributional (t, ϵ)-zero-knowledge for every polynomial t and every inverse polynomial ϵ.

Theorem 5 (Informally stated)**.** *For every language $L \in SZK$ and every polynomial p, there exists a 2-round interactive proof (P, V) for L with completeness $1 - \mathrm{negl}(\cdot)$ and soundness error $\frac{1}{p(\cdot)}$, and is distributional (t, ϵ)-zero-knowledge for every polynomial t and every inverse polynomial ϵ.*

We proceed to outline two other applications of Theorem 4.

Leakage Lemma of Gentry-Wichs. Roughly speaking, the "Leakage Lemma" of Gentry-Wichs [GW11] states that for every joint distribution $(X, \pi(X))$, where $|\pi(x)| = O(\log |x|)$ (π should be thought of as leakage on X), and for every distribution Y that is indistinguishable from X, there exists some leakage $\widetilde{\pi}$ such that the joint distributions $(X, \pi(X))$ and $(Y, \widetilde{\pi}(Y))$ are indistinguishable. As we now argue, this lemma (and in fact, a stronger version of it) is a direct consequence of Theorem 4.

In the language of zero-knowledge, let X be a distribution over statements, and consider a one-message interactive proof where $\pi(x)$ denotes the distribution over the prover's message when the statement is x. By Theorem 4, this protocol is distributional zero-knowledge, and thus there exists an *efficient* simulator S that can simulate the interaction (i.e, $(X, S(X))$ is indistinguishable from $(X, \pi(X))$). By the indistinguishability of Y and X (and the efficiency

[2] Recall that in the classic Graph Non-Isomorphism protocol the prover sends just a single bit and thus is very laconic.

of S), it directly follows that $(Y, S(Y))$ is indistinguishable from $(X, \pi(X))$. Thus we have found $\widetilde{\pi} = S$.

Let us note that our proof of the leakage lemma yields an even stronger statement—namely, we have found an efficient simulator $\widetilde{\pi}$; such a version of the leakage lemma was recently established by Jetchev and Pietrzak [JP12]. (As an independent contribution, our proof of Theorem 4 is actually significantly simpler than both the proof of [GW11] and [JP12].) Additionally, since our result on zero-knowledge applies also to *interactive* protocols, we directly also get an interactive version of the leakage lemma.

Dense Model Theorem. Roughly speaking, the Dense Model Theorem of [RTTV08, TTV09] states that if X is indistinguishable from the uniform distribution over n-bits, U_n, and R is δ-dense[3] in X, then there exists a "model-distribution" M that is (approximately) δ-dense in U_n such that M is indistinguishable from R. Again, we show that this lemma is a direct consequence of Theorem 4. (Furthermore, our proof of Theorem 4 is arguably simpler and more modular than earlier proofs of the dense model theorem.)

Let us first translate the statement of the dense model theorem into the language of zero-knowledge. Let X be a distribution over statements x, and consider some distribution R that is δ-dense in X, i.e., there exists a joint distribution $(X, B(X))$ with $\Pr[B(X) = 1] \geq \delta$ such that $R = X|(B(X) = 1)$. Define a one-bit proof where the prover sends the bit $B(x)$, where x is the statement. By Theorem 4, there exists a simulator S for this interactive proof; let $M = U_n|(S(U_n) = 1)$. By the indistinguishability of the simulation, $(X, S(X))$ is indistinguishable from $(X, B(X))$, and thus by indistinguishability of X and U_n, $(U_n, S(U_n))$ is indistinguishable from $(X, B(X))$. It follows that M is (approximately) δ-dense in U_n, and M is indistinguishable from R.

1.3 A Note on Our Non-Black-Box Simulation Technique

The universal simulators in Theorem 3, 4, and 5 are indirectly obtained via the minimax theorem used in the proof of Theorem 3, and again we make non-black-box usage of the verifier V^*. We remark that our non-black-box usage of V^* is necessary (assuming standard complexity-theoretic assumptions): We show that black-box simulation techniques cannot be used to demonstrate distributional (t, ϵ)-zero-knowledge for 2-round proof systems for languages that are hard-on-average.

Theorem 6 (Informally stated)**.** *Let L be any language that is hard-on-average for polynomial-size circuits, and let (P, V) be any 2-round interactive proof (with completeness $2/3$ and soundness error $1/3$) for L. Then, there exists a polynomial t such that for every $\epsilon(n) < 1/12$, (P, V) is not black-box distributional (t, ϵ)-zero-knowledge*

As as consequence we have that as long as SZK contains a language that is hard-on-average, our non-black-box techniques are necessary (otherwise, Theorems 5 and 6 would contradict each other). As far as we know, the above yields the first example where a non-black-box simulation technique can be used to analyze "natural" protocols (e.g., the classic graph non-isomorphism protocol) that were not "tailored" for non-black-box simulation, but for which black-box simulation is not possible. This stands in sharp contrast to the non-black-box technique of Barak [Bar01] and its follow-ups (see e.g., [PR03, Pas04, PR05, BS05, DGS09, BP12, CPS13, BP13]), where non-black-box simulation is enabled by a very specific protocol design. This gives hope that non-black-box techniques can be used to analyze simple/practical protocols.

[3] R is said to be δ-*dense* in X if for every r, $\Pr[R = r] \leq (1/\delta) \cdot \Pr[X = r]$; equivalently, R is δ-dense in X if there exists a joint distribution $(X, B(X))$ with $\Pr[B(X) = 1] \geq \delta$ such that $R = X|(B(X) = 1)$.

Let us finally remark that in our non-black-box technique, we only need to make non-black-box use of the malicious verifier V^*'s auxiliary input z and its running-time t, but otherwise we may treat V^*'s Turing machine as a black-box. Although the non-black-box simulation technique of Barak [Bar01] also makes non-black-box usage of V^*'s Turing machine, it is not hard to see that also this technique can be modified to only make non-black-box usage of z and t (but not its Turing machine)—since the description of V^*'s Turing machine is of constant length the non-black-box simulator can simply enumerate all possible Turing machines in the protocol of Barak.

1.4 Our Techniques

As mentioned, both Theorem 2 and 3 rely on the minimax theorem from game theory. Recall that the minimax theorem states that in any finite two-player zero-sum game, if for every distribution over the actions of Player 1, there exists some action for Player 2 that guarantees him an expected utility of v, then there exists some (universal) distribution of actions for Player 2 such that no matter what action Player 1 picks, Player 2 is still guaranteed an expected utility of v. For us, Player 1 will be choosing a distinguisher, and Player 2 will be choosing a simulator; roughly speaking, Player 2's utility will be "high" if the simulation is "good" for the distinguisher chosen by Player 1. Now, by the weak zero-knowledge property, we are guaranteed that for every distinguisher chosen by Player 1, there exists some simulator for Player 2 that guarantees him a high utility. Thus intuitively, by the minimax theorem, Player 2 should have a simulator that yields him high utility with respect to any distinguisher.

There are two problems with this approach. First, to apply the minimax theorem, we require the existence of a good "distinguisher-dependent" simulator for every *distribution* over distinguishers. Secondly the minimax theorem only guarantees the existence of a distribution over simulators that works well against every distinguisher. We resolve both of these issues in quite different ways for Theorem 3 and Theorem 2.

In the context of Theorem 3, since we model both the simulator and distinguisher as non-uniform machines, we can use standard techniques to "derandomize" any distribution over simulators/distinguishers into a single simulator/distinguisher that gets some extra non-uniform advice: we simply approximate the original distribution by sufficiently many samples from it, and these samples can be provided to a single machine as non-uniform advice. (Such "derandomization" techniques originated in the proof of the hard-core lemma [Imp95].)

In the context of Theorem 2, the situation is more difficult since we need both the distinguisher and the simulator to be uniform. In particular, we are only guaranteed the existence of a good distinguisher-dependent simulator for every *uniform* distinguisher and not necessarily for non-uniform ones. Here, we instead try to efficiently and uniformly find the "minimax" distribution over simulator strategies. If this can be done, then we do have a single uniform (and efficient) simulator algorithm. Towards this, we use a *multiplicative weights algorithm*, which can be used to approximately find the minimax strategies of two-player zero-sum games (e.g., see [FS99]). The multiplicative weights algorithm roughly works as follows. In the first round, Player 1 chooses the uniform distribution over the set of all $t(n)$-time Turing machines with description size $\leq \log n$ (note that any $t(n)$-time uniform distinguisher will be a member of this set for sufficiently large n), and then Player 2 chooses a "good simulator" that yields high payoff with respect to Player 1's distribution (note that since Player 1's distribution is uniformly and efficiently computable, we can view the process of sampling from it, and next running the sampled distinguisher, as a single uniform and efficient distinguisher, and thus we may rely on the weak zero-knowledge definition to conclude that a good simulator exists). In the next round, Player 1 updates its distribution using a multiplicative update rule that depends on Player 2's chosen simulator in the previous round; Player 2 again chooses a simulator that yields high payoff with respect to Player 1's new distribution, etc. By repeating this procedure for polynomially many rounds, Player 2 obtains a sequence of simulators such that the uniform distribution over the multiset of simulators yields high payoff no matter what distinguisher Player 1 chooses.

There are some issues that need to be resolved. In each round, we need to pick a simulator that works well against a (uniformly and efficiently computable) distribution over $t(n)$-time distinguishers. Although the running-time of the underlying distinguishers is bounded by $t(n)$, the time needed to sample from this distribution could be growing (exponentially) in each round, which in turn could potentially lead to an exponential growth in the running-time of the simulator. Thus after polynomially many rounds, it is no longer clear that the simulator or the distribution over distinguishers is polynomial-time.[4] To deal with this issue, we rely on the "good" distinguisher-dependent simulator for a single universal distinguisher that receives as auxiliary input the code of the actual distinguisher it is running; we can then at each step approximate the distribution over distinguishers and feed this approximation as auxiliary input to the universal distinguisher.

Another important issue to deal with is the fact that to evaluate the "goodness" of a simulation w.r.t. to some distinguisher (i.e., to compute the utility function), we need to be able to sample true views of the malicious verifier in an interaction with the honest prover—but if we could do this, then we would already be done! Roughly speaking, we overcome this issue by showing that the goodness of a simulation w.r.t. a particular distinguisher D can be approximated by using the distinguisher-dependent simulator S_D for D.

We remark that in both of the above proofs, the reason that we work with a (t, ϵ)-notion of zero-knowledge is that the running-time of the simulator we construct is polynomial in t and $1/\epsilon$.

1.5 Related Work

As mentioned above, the notion of weak zero-knowledge was first introduced by Dwork, Naor, Reingold and Stockmeyer [DNRS03]. Dwork et al also considered non-uniform versions and distributional versions of zero-knowledge; distributional versions of zero-knowledge were first considered by Goldreich [Gol93] in a uniform setting (called uniform zero-knowledge).

The minimax theorem from game-theory has been applied in various contexts in complexity theory (e.g., [Imp95, BSW03, RTTV08, TTV09]) and more recently also in cryptography (e.g., [RTTV08, DP08, CKLR11, GW11, JP12]). The proof of Theorem 4 is related to the approaches taken in these previous works, and most closely related to the approach taken in [TTV09].

However, as far as we know, none of the earlier results have applied the minimax theorem in the context of zero-knowledge. Nevertheless, as we mentioned above, our Theorem 4 implies some of these earlier results (and shows that they can be understood in the language of zero-knowledge).

1.6 Overview

In Section 2, we show that weak zero-knowledge is not equivalent to zero-knowledge (Theorem 1 above). In Section 3, we show that weak and strong (t, ϵ)-zero-knowledge are equivalent (Theorem 2 above). In Section 4, we show that super-weak and strong distributional (t, ϵ)-zero-knowledge are equivalent (Theorem 3 above), and interactive proofs with a laconic prover are distributional zero-knowledge (Theorem 4 above), and we also describe applications of this result. In Appendix D, we separate the notion of super-weak and weak (t, ϵ)-zero-knowledge.

[4] A similar issue appeared in a recent paper by us in the context of forecast testing [CLP13], where we used a related, but different, technique to overcome it.

2 Separation of Weak and Strong Zero-Knowledge

Given a prover P, a verifier V^*, and $x, z \in \{0,1\}^*$, let $Out_{V^*}[P(x) \leftrightarrow V^*(x,z)]$ denote the output of $V^*(x,z)$ after interacting with $P(x)$. We now state the definition of zero-knowledge for convenient reference.

Definition 7 (zero-knowledge). Let (P,V) be an interactive proof system for a language L. We say that (P,V) is *zero-knowledge* if for every PPT adversary V^*, there exists a PPT simulator S such that for every PPT distinguisher D, there exists a negligible function $\nu(\cdot)$ such that for every $n \in \mathbb{N}$, $x \in L \cap \{0,1\}^n$, and $z \in \{0,1\}^*$, we have

$$|\Pr[D(x,z,Out_{V^*}[P(x) \leftrightarrow V^*(x,z)]) = 1] - \Pr[D(x,z,S(x,z)) = 1]| \leq \nu(n).$$

Remark. If L is a language in NP with witness relation R_L, we usually require the prover P to be efficient, but on common input x, we also give any witness $y \in R_L(x)$ to the prover P. We refer to such a notion as *efficient prover* zero-knowledge. More formally, in the definition of zero-knowledge above, we would change "$x \in L \cap \{0,1\}^n$, and $z \in \{0,1\}^*$" to "$x \in L \cap \{0,1\}^n$, $y \in R_L(x)$, and $z \in \{0,1\}^*$", and we would change $P(x)$ to $P(x,y)$ and require P to be efficient. All subsequent definitions can be extended to an efficient prover setting in an obvious way.

One can relax the definition of zero-knowledge by switching the order of the quantifiers $\exists S$ and $\forall D$ so that the simulator S can depend on the distinguisher D. We call the relaxed definition *weak zero-knowledge* (following [DNRS03]).

Definition 8 (weak zero-knowledge). Let (P,V) be an interactive proof system for a language L. We say that (P,V) is *weak zero-knowledge* if for every PPT adversary V^* and every PPT distinguisher D, there exists a PPT simulator S and a negligible function $\nu(\cdot)$ such that for every $n \in \mathbb{N}$, $x \in L \cap \{0,1\}^n$, and $z \in \{0,1\}^*$, we have

$$|\Pr[D(x,z,Out_{V^*}[P(x) \leftrightarrow V^*(x,z)]) = 1] - \Pr[D(x,z,S(x,z)) = 1]| \leq \nu(n).$$

We now show that under reasonable cryptographic assumptions, weak zero-knowledge is not equivalent to zero-knowledge.

Theorem 9. *Under reasonable cryptographic assumptions, there exists an interactive proof system (P,V) for an NP language L such that (P,V) is weak zero-knowledge but not zero-knowledge.*

Proof idea. The proof *roughly* works as follows. Suppose that for $i = 1, \ldots, \log^2 n$, we have a two-round "timed" commitment scheme Com_i that is hard to break in $p(n)^{i-1}$ steps (where $p(\cdot)$ is some polynomial), but can always be broken in $p(n)^i$ steps to obtain the committed value (e.g., one can get such timed commitment schemes from a timed commitment scheme in [BN00]). Suppose also that for $i = 1, \ldots, \log^2 n$, we have a "timed" worst-case weak one-way permutation f_i that is somewhat hard to invert in $p(n)^{i+1}$ steps in the worst case (i.e., an adversary running in $p(n)^{i+1}$ steps will fail to invert some instance $f_i(x')$ with probability at least $1/poly(n)$), but can always be inverted in $p(n)^{i+2}$ steps. (Note that f_i is slightly harder to break than Com_i.) Now, let L be the trivial NP language $\{0,1\}^*$ with witness relation $R_L(x) = \{(f_1^{-1}(x), \ldots, f_{\log^2 |x|}^{-1}(x))\}$.

Let $(P(x,y), V(x))$ be the following interactive proof, where $x \in \{0,1\}^*$, $n = |x|$, $\ell = \log^2 n$, and $y = (f_1^{-1}(x), \ldots, f_\ell^{-1}(x))$:

1. The verifier V generates and sends ρ_i for $i = 1, \ldots, \ell$ to the prover, where ρ_i is the first message of an execution of Com_i.

2. The prover P sends $\mathsf{Com}_i(f_i^{-1}(x), \rho_i)$ for $i = 1, \ldots, \ell$ to the verifier, where $\mathsf{Com}_i(v, r)$ denotes the commitment of v using Com_i with first message r.

3. The verifier V accepts (i.e., outputs 1).

To see that (P, V) is weak zero-knowledge, consider any PPT verifier V^* and any PPT distinguisher D, and let $T(n)$ be a polynomial that bounds the combined running time of V^* and D. Then, a simulator S can compute the smallest positive integer j such that $p(n)^{j-1} > T(n)$, and then break $f_1^{-1}(x), \ldots, f_{j-1}^{-1}(x)$ in polynomial time. Then, the simulator S can simulate the protocol except that for $i = j, \ldots, \ell$, the simulator S sends $\mathsf{Com}_i(0^n, \rho_i)$ to V^* since S does not know $f_i^{-1}(x)$. By the hiding property of $\mathsf{Com}_j, \ldots, \mathsf{Com}_\ell$, the distinguisher D cannot distinguish between the output of the verifier V^* (in a true interaction with P) and the output of the simulator S, since D and V^* (combined) cannot break any of the commitment schemes $\mathsf{Com}_j, \ldots, \mathsf{Com}_\ell$ (since D and V^* do not run long enough).

Intuitively, (P, V) is not zero-knowledge because the existence of a (universal) simulator S would allow us to invert a worst-case weak one-way permutation f_j with overwhelming probability and in less time than what is specified in our hardness assumption for f_j. To see this, consider a PPT distinguisher D that, given x and a view of V, runs longer than S and breaks a commitment $\mathsf{Com}_j(w_j, \rho'_j)$ from the view of V such that the time needed to break f_j is much longer than the running time of the simulator S, and then verifies whether or not $f(w_j) = x$. The fact that the simulator S works for the distinguisher D will ensure that with overwhelming probability, the output of $S(x)$ will contain a commitment $\mathsf{Com}_j(w_j, \rho'_j)$ of some w_j such that $f_j(w_j) = x$. Thus, we can now construct an adversary A that inverts $f_j(w_j)$ with overwhelming probability by running the simulator S on input $f_j(w_j)$ and breaking the commitment $\mathsf{Com}_j(w_j, \rho'_j)$ in the output of S. Since breaking f_j takes longer time than running the simulator S and breaking the commitment $\mathsf{Com}_j(w_j, \rho'_j)$, the adversary A contradicts our hardness assumption for f_j. □

See Appendix A for the formal proof of Theorem 9.

3 From Weak to Strong (t, ϵ)-Zero-Knowledge

From Theorem 9, we know that zero-knowledge and weak zero-knowledge are not equivalent. Thus, we now consider relaxed notions of zero-knowledge. We first consider a concrete security variant of the notion of zero-knowledge.

Definition 10 (**(t, ϵ)-zero-knowledge).** Let (P, V) be an interactive proof system for a language L. We say that (P, V) is (t, ϵ)-*zero-knowledge* if for every PPT adversary V^*, there exists a PPT simulator S such that for every t-time distinguisher D, there exists an $n_0 \in \mathbb{N}$ such that for every $n \geq n_0$, $x \in L \cap \{0,1\}^n$, and $z \in \{0,1\}^*$, we have

$$|\Pr[D(x, z, Out_{V^*}[P(x) \leftrightarrow V^*(x, z)]) = 1] - \Pr[D(x, z, S(x, z)) = 1]| \leq \epsilon(n).$$

Similar to before, we can relax the definition of zero-knowledge by switching the order of the quantifiers $\exists S$ and $\forall D$ so that the simulator S can depend on the distinguisher D. We call the relaxed definition *weak* (t, ϵ)*-zero-knowledge*.

Definition 11 (**weak (t, ϵ)-zero-knowledge).** Let (P, V) be an interactive proof system for a language L. We say that (P, V) is *weak* (t, ϵ)*-zero-knowledge* if for every PPT adversary V^* and every t-time distinguisher D, there exists a PPT simulator S and an $n_0 \in \mathbb{N}$ such that for every $n \geq n_0$, $x \in L \cap \{0,1\}^n$, and $z \in \{0,1\}^*$, we have

$$|\Pr[D(x, z, Out_{V^*}[P(x) \leftrightarrow V^*(x, z)]) = 1] - \Pr[D(x, z, S(x, z)) = 1]| \leq \epsilon(n).$$

Note that if (P, V) is (t, ϵ)-zero-knowledge (resp. weak (t, ϵ)-zero-knowledge) for some super polynomial function t and some negligible function ϵ, then (P, V) is zero-knowledge (resp. weak zero-knowledge) in the classic sense. We now show that (t, ϵ)-zero-knowledge and weak (t, ϵ)-zero-knowledge are equivalent if we consider all polynomials t and inverse polynomials ϵ.

Theorem 12. *Let (P, V) be an interactive proof system for a language L. Then, (P, V) is weak (t, ϵ)-zero-knowledge for every polynomial t and inverse polynomial ϵ if and only if (P, V) is (t', ϵ')-zero-knowledge for every polynomial t' and inverse polynomial ϵ'.*

Proof. The "if" direction clearly holds by definition. We will now prove the "only if" direction. Suppose (P, V) is weak (t, ϵ)-zero-knowledge for every polynomial t and inverse polynomial ϵ. Let t' be any polynomial, and let ϵ' be any inverse polynomial.

Let V^* be any PPT adversary, and let $T_{V^*}(\cdot)$ be any polynomial that bounds the running time of V^*. It is not hard to see that without loss of generality, we can assume that the auxiliary input $z \in \{0,1\}^*$ in the definition of (t', ϵ')-zero-knowledge is exactly $C \cdot (T_{V^*}(n) + t'(n))$ bits long, where C is some constant ≥ 1.[5] Furthermore, it is easy to see that without loss of generality, we can also remove the absolute value $| \cdot |$ and change $\epsilon'(n)$ to $O(\epsilon'(n))$. Thus, it suffices to construct a PPT simulator S such that for every t'-time distinguisher D, there exists an $n_0 \in \mathbb{N}$ such that for every $n \geq n_0$, $x \in L \cap \{0,1\}^n$, and $z \in \{0,1\}^*$ with $|z| = C \cdot (T_{V^*}(n) + t'(n))$, we have

$$\Pr[D(x, z, Out_{V^*}[P(x) \leftrightarrow V^*(x, z)]) = 1] - \Pr[D(x, z, S(x, z)) = 1] \leq O(\epsilon'(n)).$$

We will now construct the required PPT simulator S for V^*.

High-level description of the simulator S: We first give a high-level description of the simulator S. The simulator S uses the multiplicative weights algorithm described in [FS99]. The simulator S, on input (x, z) with $n := |x|$, first runs a multiplicative weights algorithm to find a "good set" of simulator machines $\{S_1, \ldots, S_L\}$; then, the simulator S randomly and uniformly chooses one of the simulator machines in $\{S_1, \ldots, S_L\}$ to perform the simulation, i.e., S runs the chosen simulator machine on input (x, z) and outputs whatever the simulator machine outputs.

Before we describe the multiplicative weights algorithm run by the simulator S, let us introduce some notation. Given a simulator S' and a distinguisher D', let the "payoff" of S' (with respect to D') be

$$\mu(S', D') := \Pr[D'(x, z, S'(x, z)) = 1] - \Pr[D'(x, z, Out_{V^*}[P(x) \leftrightarrow V^*(x, z)]) = 1].$$

Given a simulator S' and a distribution $\mathcal{D}^{(i)}$ over distinguishers, let

$$\mu(S', \mathcal{D}^{(i)}) := \mathbb{E}_{D' \sim \mathcal{D}^{(i)}}[\mu(S', D')] = \sum_{D' \in Supp(\mathcal{D}^{(i)})} \mathcal{D}^{(i)}(D') \cdot \mu(S', D').$$

We note that we want to design the simulator S so that for every t'-time distinguisher D, we have $\mu(S, D) \geq -O(\epsilon'(n))$.

[5] This follows from standard padding techniques and the fact that the adversary V^* and the distinguisher D cannot read any of the bits after the first $T_{V^*}(n) + t'(n)$ bits of z.

Let $D_1, D_2, D_3, \ldots$ be any enumeration of the set of all (uniform) distinguishers, and let $D'_1, D'_2, D'_3, \ldots$ be the corresponding sequence where D'_j is the same as D_j except that after $t'(n)$ steps, D'_j stops and outputs 0. We note that each fixed t'-time distinguisher D will eventually appear in the set $\{D'_1, \ldots, D'_n\}$ as n gets larger.

We now describe the multiplicative weights algorithm run by S. In the multiplicative weights algorithm, S simulates L rounds (repetitions) of a zero-sum game between a "simulator player" Sim and a "distinguisher player" Adv, where the payoff function for Sim is the function $\mu(\cdot, \cdot)$ defined above. In each round i, Adv chooses a mixed strategy (i.e., a distribution) $\mathcal{D}^{(i)}$ over its set of pure strategies $\{D'_1, \ldots, D'_n\}$ (a set of distinguishers), and then Sim chooses a simulator machine $S_i := S_i(\mathcal{D}^{(i)})$ that hopefully "does well" against Adv's mixed strategy $\mathcal{D}^{(i)}$, i.e., Sim's (expected) payoff $\mu(S_i, \mathcal{D}^{(i)})$ is high.

In the first round, Adv chooses the uniform distribution $\mathcal{D}^{(1)}$ over $\{D'_1, \ldots, D'_n\}$. After each round i, Adv updates its mixed strategy to get $\mathcal{D}^{(i+1)}$ in a manner similar to the multiplicative weights algorithm described in [FS99], which involves the payoff function μ. However, we here use an approximation $\widehat{\mu}$ of the payoff function μ, since we cannot *efficiently* compute μ exactly. This is because μ involves $Out_{V^*}[P(x) \leftrightarrow V^*(x, z)])$, i.e., the output of V^* after interacting with the prover P, and P may be inefficient (or for efficient prover zero-knowledge, P would have a witness y that S does not have). However, given a distinguisher D', we can estimate $\mu(S_i, D')$ by approximating $Out_{V^*}[P(x) \leftrightarrow V^*(x, z)]$ by the output of a simulator $S_{D'}$ for V^* that is good for the distinguisher D' specifically; the existence of such a simulator is guaranteed by the weak zero-knowledge property of (P, V). There are still some issues: S might not be able to find $S_{D'}$ *efficiently and uniformly*, and $S_{D'}$ only works well for sufficiently large n. We resolve these issues by using a "universal" distinguisher that essentially takes a description of a distinguisher D' as auxiliary input and runs D', and we use a simulator that is good with respect to this universal distinguisher specifically.

Using an analysis similar to that in [FS99], we will show that if, in every round $i \in [L]$, Sim manages to choose a simulator machine S_i that does well against Adv's mixed strategy $\mathcal{D}^{(i)}$, then the uniform mixed strategy over the set $\{S_1, \ldots, S_L\}$ of chosen simulator machines does well against all the distinguishers in $\{D'_1, \ldots, D'_n\}$. To choose a simulator machine S_i that does well against Adv's mixed strategy $\mathcal{D}^{(i)}$, Sim makes use of the weak zero-knowledge property of (P, V), which guarantees that for every distinguisher D, there exists a simulator S_D that does well against D. However, there are some complications: (1) $\mathcal{D}^{(i)}$ is a *mixture* of distinguishers, not a single distinguisher; (2) Sim might not be able to *efficiently and uniformly* find the distinguisher-dependent simulator; and (3) even if Sim can efficiently and uniformly find the distinguisher-dependent simulator, the simulator depends on the mixed strategy $\mathcal{D}^{(i)}$, and the time needed to sample from $\mathcal{D}^{(i)}$ could be growing (exponentially) in each round, which in turn can potentially lead to an exponential growth in the running time of the distinguisher-dependent simulator as more rounds are performed.

The simulator overcomes these problems by again using a "universal" distinguisher D_U that takes the weights (i.e., probability masses) of a distribution $\mathcal{D}$ over $\{D'_1, \ldots, D'_n\}$ as auxiliary input, samples a distinguisher from the distribution $\mathcal{D}$, and then runs the sampled distinguisher. Let S_{D_U} be the simulator for V^* that is good with respect to D_U specifically; again, the existence of such a simulator is guaranteed by the weak zero-knowledge property of (P, V). Sim chooses S_i to be the simulator machine that runs S_{D_U} with the weights of the distribution $\mathcal{D}^{(i)}$ provided as auxiliary input. We now give a formal description of the simulator S.

The simulator S: Let $D_1, D_2, D_3, \ldots$ be any enumeration of the set of all (uniform) distinguishers, and let $D'_1, D'_2, D'_3, \ldots$ be the corresponding sequence where D'_j is the same as D_j except that after $t'(n)$ steps, D'_j stops and outputs 0.

The simulator S, on input (x, z) with $n := |x|$, proceeds as follows:

1. Let $T_{D_U}(n) = O((T_{V^*}(n) + t'(n) + n)^2)$.

   Given a distribution $\mathcal{D}$ over $\{D'_1, \dots, D'_n\}$, let $\vec{p}_{\mathcal{D}}$ denote the vector of weights (i.e., probability masses) representing $\mathcal{D}$, i.e., $\vec{p}_{\mathcal{D}} = (\mathcal{D}(D'_1), \dots, \mathcal{D}(D'_n))$.

   Let D_U be a "universal" distinguisher that, on input (x, z', v), first parses z' as $z' = z||\vec{p}_{\mathcal{D}}$, where $\vec{p}_{\mathcal{D}}$ is a vector the weights representing some distribution $\mathcal{D}$ over $\{D'_1, \dots, D'_n\}$; then, D_U samples a distinguisher D'_j from the distribution $\mathcal{D}$, and then runs D'_j on input (x, z, v), but D_U always stops after $T_{D_U}(n)$ steps regardless of whether or not D'_j finishes running.

   Let S_{D_U} be the PPT simulator for D_U that is guaranteed by the weak (T_{D_U}, ϵ')-zero-knowledge property of (P, V).

2. Let $L = \Theta(\frac{\log n}{\epsilon'(n)^2})$ and $\beta = \frac{1}{1+\sqrt{(2 \ln n)/L}}$. ($L$ is the number of rounds we will run the multiplicative weights algorithm for, and β is used in the multiplicative update rule.)

3. **Multiplicative weights algorithm:**

   Let $\mathcal{D}^{(1)}$ be the uniform distribution over $\{D'_1, \dots, D'_n\}$. (The probability mass $\mathcal{D}^{(1)}(D'_j)$ for D'_j can be thought of as the "weight" for D'_j.)

   For $i = 1, \dots, L$ **do:**

   (a) **Choosing a simulator machine S_i that does well against $\mathcal{D}^{(i)}$:**
   Let S_i be a simulator machine that, on input (x, z), outputs $S_{D_U}(x, z||\vec{p}_{\mathcal{D}^{(i)}})$.

   (b) **Weight update:**
   Compute the distribution $\mathcal{D}^{(i+1)}$ from $\mathcal{D}^{(i)}$ by letting

   $$\mathcal{D}^{(i+1)}(D'_j) \sim \beta^{\widehat{\mu}(S_i, D'_j)} \cdot \mathcal{D}^{(i)}(D'_j)$$

   for every $D'_j \in \{D'_1, \dots, D'_n\}$ (and renormalizing), where

   $$\widehat{\mu}(S_i, D'_j) := \mathrm{freq}_k[D'_j(x, z, S_i(x, z))] - \mathrm{freq}_k[D'_j(x, z, S_{D_U}(x, z||\vec{p}_{D'_j}))],$$

   where $\mathrm{freq}_k[D'_j(x, z, S_i(x, z))]$ and $\mathrm{freq}_k[D'_j(x, z, S_{D_U}(x, z||\vec{p}_{D'_j}))]$ are approximations of $\Pr[D'_j(x, z, S_i(x, z)) = 1]$ and $\Pr[D'_j(x, z, S_{D_U}(x, z||\vec{p}_{D'_j})) = 1]$ by taking $k := \Theta(\frac{\log(nL/\epsilon'(n))}{\epsilon'(n)^2})$ samples, respectively, and computing the relative frequency in which 1 is outputted. The function $\widehat{\mu}$ should be viewed as being an approximation of the payoff function μ.

   End for

4. Choose $S_i \in \{S_1, \dots, S_L\}$ uniformly at random.

5. Run the simulator S_i on input (x, z) and output $S_i(x, z)$.

We now continue with the formal proof. It can be easily verified that S runs in time $poly(n, t'(n), \frac{1}{\epsilon'(n)})$. Let D be any distinguisher whose running time is bounded by $t'(n)$. Fix an integer n that is sufficiently large so that the distinguisher D appears in $\{D_1, \dots, D_n\}$ and S_{D_U} works for the distinguisher D_U on input size n for x. We note that the distinguisher D also appears in $\{D'_1, \dots, D'_n\}$, since the running time of D is bounded by $t'(n)$. Fix $x \in L \cap \{0,1\}^n$ and $z \in \{0,1\}^*$ with $|z| = C \cdot (T_{V^*}(n) + t'(n))$. To prove the theorem, it suffices to show that

$$\mu(S, D) \geq -O(\epsilon'(n)).$$

To show this, we will proceed as follows: (1) We first show that if, in every round i the chosen simulator S_i does well against the distribution $\mathcal{D}^{(i)}$ with respect to our approximation $\widehat{\mu}$ of μ, then the simulator S does well against D with respect to $\widehat{\mu}$; this is the first lemma below; (2) We then show that the first lemma holds even if we replace $\widehat{\mu}$ with μ; this is the second lemma below; (3) Finally, we show that in each round i, the chosen simulator S_i indeed does well against the distribution $\mathcal{D}^{(i)}$ with respect to μ.

We now proceed with the proof. For $i = 1, \ldots, L$, let

$$\widehat{\mu}(S_i, \mathcal{D}^{(i)}) := \mathbb{E}_{D' \sim \mathcal{D}^{(i)}}[\widehat{\mu}(S_i, D')] = \sum_{k=1}^{n} \mathcal{D}^{(i)}(D'_k) \cdot \widehat{\mu}(S_i, D'_k).$$

One should view $\widehat{\mu}(S_i, \mathcal{D}^{(i)})$ as an approximation of $\mu(S_i, \mathcal{D}^{(i)})$.

Lemma 13. *For every distinguisher $D'_j \in \{D'_1, \ldots, D'_n\}$, if we run the simulator $S(x, z)$, then (with probability 1) $S(x, z)$ generates $\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(L)}$ and $S_1, \ldots, S_L$ such that*

$$\frac{1}{L} \sum_{i=1}^{L} \widehat{\mu}(S_i, D'_j) \geq \frac{1}{L} \sum_{i=1}^{L} \widehat{\mu}(S_i, \mathcal{D}^{(i)}) - O(\epsilon'(n)).$$

The proof of Lemma 13 is essentially the same as a lemma found in [CLP13], whose proof is very similar to the analysis of the multiplicative weights algorithm found in [FS99]. In [FS99], the multiplicative weights algorithm updates the weights of $\mathcal{D}^{(i)}$ using the exact value of $\mu(S_i, D'_j)$; here, we only have an approximation $\widehat{\mu}(S_i, D'_j)$ of $\mu(S_i, D'_j)$, but with minor changes, the analysis in [FS99] can still be used to show Lemma 13. For completeness, we provide a proof of Lemma 13 in Appendix B.

We now show that we can essentially replace the $\widehat{\mu}$ in Lemma 13 with μ.

Lemma 14. *For every $D' \in \{D'_1, \ldots, D'_n\}$, if we run the simulator $S(x, z)$, then with probability $1 - O(\epsilon'(n))$ over the random coins of S, $S(x, z)$ generates $\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(L)}$ and $S_1, \ldots, S_L$ such that*

$$\frac{1}{L} \sum_{i=1}^{L} \mu(S_i, D') \geq \frac{1}{L} \sum_{i=1}^{L} \mu(S_i, \mathcal{D}^{(i)}) - O(\epsilon'(n)).$$

The proof of Lemma 14 roughly works as follows. We take Lemma 13 and show that each time we approximate μ via $\widehat{\mu}$, the approximation is good with high probability; this follows from Chernoff bounds and the fact that S_{D_U} is a simulator for V^* that is good with respect to the "universal" distinguisher D_U. Lemma 14 then follows from the union bound. See Appendix B for the proof of Lemma 14.

To complete the proof of Theorem 12, we will now show that $\mu(S, D) \geq -O(\epsilon'(n))$. We first show that for every $i \in [L]$, we always have $\mu(S_i, \mathcal{D}^{(i)}) \geq -O(\epsilon'(n))$. Fix $i \in [L]$. Now, we observe that

$$
\begin{aligned}
& \mu(S_i, \mathcal{D}^{(i)}) \\
= \ & \sum_{j=1}^{n} \mathcal{D}^{(i)}(D'_j) \cdot \big(\Pr[D'_j(x,z,S_i(x,z))=1] - \Pr[D'_j(x,z,Out_{V^*}[P(x) \leftrightarrow V^*(x,z)])=1]\big) \\
= \ & \Pr[D_U(x,z||\vec{p}_{\mathcal{D}^{(i)}}, S_i(x,z))=1] - \Pr[D_U(x,z||\vec{p}_{\mathcal{D}^{(i)}}, Out_{V^*}[P(x) \leftrightarrow V^*(x,z)])=1] \\
= \ & \Pr[D_U(x,z||\vec{p}_{\mathcal{D}^{(i)}}, S_{D_U}(x,z||\vec{p}_{\mathcal{D}^{(i)}}))=1] - \Pr[D_U(x,z||\vec{p}_{\mathcal{D}^{(i)}}, Out_{V^*}[P(x) \leftrightarrow V^*(x,z||\vec{p}_{\mathcal{D}^{(i)}})])=1] \\
\geq \ & -\epsilon'(n), \qquad (2)
\end{aligned}
$$

where the second equality follows from the definition of D_U, the third equality follows from the definition of S_i and the fact that $V^*(x,z) = V^*(x,z||\vec{p}_{\mathcal{D}^{(i)}})$ (since $|z| \geq T_{V^*}(n)$), and the last inequality follows from the fact that S_{D_U} is a simulator for D_U in the weak (t',ϵ')-zero-knowledge property of (P,V).

Now, combining Lemma 14 and (2), we have that with probability $1 - O(\epsilon'(n))$ over the randomness of S, $S(x,z)$ generates $S_1,\ldots,S_L$ such that

$$\frac{1}{L}\sum_{i=1}^{L} \mu(S_i, D) \geq -O(\epsilon'(n)). \qquad (3)$$

Now, recall that after generating $S_1,\ldots,S_L$, the simulator $S(x,z)$ chooses a uniformly random $S_i \in \{S_1,\ldots,S_L\}$ and runs $S_i(x,z)$. Thus, conditional on $S(x,z)$ generating a particular sequence $S_1,\ldots,S_L$, we have $\mu(S,D) = \sum_{i=1}^{L} \frac{1}{L} \cdot \mu(S_i,D)$. Combining this with (3) (which holds with probability $1 - O(\epsilon'(n))$ over the randomness of S), we get

$$\mu(S,D) \geq -O(\epsilon'(n)) - O(\epsilon'(n)) = -O(\epsilon'(n)),$$

as required. This completes the proof of Theorem 12. □

4 From Super-Weak to Strong Distributional (T,t,ϵ)-Zero-Knowledge

In this section we consider a "super-weak" notion of zero-knowledge, where not only do we allow the simulator to depend on the distinguisher, but also, we only require that the simulator manages to make the distinguisher output 1 with *at least as high probability* (minus some "small" gap) as the probability that the distinguisher outputs 1 when given a true view of V^*. That is, we only consider "one-sided" indistinguishability. (Such a notion was previously considered in [HP10].)

In Appendix D, we show that super-weak (t,ϵ)-zero-knowledge is not equivalent to weak (t,ϵ)-zero-knowledge. Thus, we here consider an alternative "distributional" notion of zero-knowledge (a la [Gol93]) where indistinguishability of the simulation is only required for any distribution over statements (and auxiliary inputs). Additionally, we here model both the distinguisher and the simulator as non-uniform algorithms (as opposed to uniform ones). (The combinations of these variants was previously considered by [DNRS03]). For concreteness, we also add a parameter T to the definition and require that the simulator is of size at most $T(n)$, and thus we also bound the size of the malicious verifier V^* by $t(n)$.

Definition 15 (distributional (T,t,ϵ)-zero-knowledge). Let (P,V) be an interactive proof system for a language L. We say that (P,V) is *distributional (T,t,ϵ)-zero-knowledge* if for every $n \in \mathbb{N}$, every joint distribution (X_n, Y_n, Z_n) over $(L \cap \{0,1\}^n) \times \{0,1\}^* \times \{0,1\}^*$, and every $t(n)$-size adversary V^*, there exists a $T(n)$-size simulator S such that for every $t(n)$-size distinguisher D, we have

$$|\Pr[D(X_n, Z_n, Out_{V^*}[P(X_n, Y_n) \leftrightarrow V^*(X_n, Z_n)]) = 1] - \Pr[D(X_n, Z_n, S(X_n, Z_n)) = 1]| \leq \epsilon(n).$$

In the above definition, if L is an NP-language, then we require (i.e., assume) Y_n to be a witness of X_n (this also applies to the corresponding definition below). *Weak distributional* (T, t, ϵ)*-zero-knowledge* can be defined in an analogous way by switching the ordering of the quantifiers $\exists S$ and $\forall D$. We now turn to define *super-weak distributional* (T, t, ϵ)*-zero-knowledge.*

Definition 16 (super-weak distributional (T, t, ϵ)-zero-knowledge). Let (P, V) be an interactive proof system for a language L. We say that (P, V) is *super-weak distributional* (T, t, ϵ)*-zero-knowledge* if for every $n \in \mathbb{N}$, every joint distribution (X_n, Y_n, Z_n) over $(L \cap \{0,1\}^n) \times \{0,1\}^* \times \{0,1\}^*$, every $t(n)$-size adversary V^*, and every $t(n)$-size distinguisher D, there exists a $T(n)$-size simulator S such that

$$\Pr[D(X_n, Z_n, Out_{V^*}[P(X_n, Y_n) \leftrightarrow V^*(X_n, Z_n)]) = 1] - \Pr[D(X_n, Z_n, S(X_n, Z_n)) = 1] \leq \epsilon(n).$$

We now show that super-weak distributional (T, t, ϵ)-zero-knowledge is equivalent to distributional (T, t, ϵ)-zero-knowledge if we consider all polynomials for T and t and all inverse polynomials for ϵ. In fact, we prove a more general theorem that also describes the loss in the parameters T, t, and ϵ.

Theorem 17. *Let (P, V) be an interactive proof system for a language L, and suppose (P, V) is super-weak distributional (T, t, ϵ)-zero-knowledge. Then, (P, V) is also distributional $(T', t', 2\epsilon)$-zero-knowledge, where $t'(n) = \Omega(\epsilon(n)\sqrt{t(n)} - n)$ and $T'(n) = O(\frac{t'(n)\ln(n+t'(n))}{\epsilon(n)^2}) \cdot T(n)$.*

Proof. Let $n \in \mathbb{N}$, let (X_n, Y_n, Z_n) be any joint distribution over $(L \cap \{0,1\}^n) \times \{0,1\}^* \times \{0,1\}^*$, and let V^* be any $t(n)$-size adversary. It is easy to see that w.l.o.g., we can assume that the length of Z_n is always bounded by $t'(n)$, and we can remove the absolute value $|\cdot|$ in the definition of distributional $(T', t', 2\epsilon)$-zero-knowledge. Thus, it suffices to show the following claim:

Claim 18. *There exists a $T'(n)$-size simulator S such that for every $t'(n)$-size distinguisher D,*

$$\Pr[D(X_n, Z_n, S(X_n, Z_n)) = 1] - \Pr[D(X_n, Z_n, Out_{V^*}[P(X_n, Y_n) \leftrightarrow V^*(X_n, Z_n)]) = 1] \geq -2\epsilon(n).$$

We now proceed to showing the above claim. We define a two-player zero-sum game between a "simulator player" Sim and a "distinguisher player" Adv. The set $Strat_{\mathsf{Sim}}$ of pure strategies for Sim is the set of all $T(n)$-size simulators, and the set $Strat_{\mathsf{Adv}}$ of pure strategies for Adv is the set of all $t'(n)$-size distinguishers. The payoff for Sim when Sim chooses a simulator $S \in Strat_{\mathsf{Sim}}$ and Adv chooses a distinguisher $D \in Strat_{\mathsf{Adv}}$ is

$$\mu_n(S, D) := \Pr[D(X_n, Z_n, S(X_n, Z_n)) = 1] - \Pr[D(X_n, Z_n, Out_{V^*}[P(X_n, Y_n) \leftrightarrow V^*(X_n, Z_n)]) = 1].$$

For mixed strategies (i.e., distributions) $\mathcal{S}$ over $Strat_{\mathsf{Sim}}$, and $\mathcal{D}$ over $Strat_{\mathsf{Adv}}$, we define

$$\mu_n(\mathcal{S}, \mathcal{D}) := \mathbb{E}_{S \leftarrow \mathcal{S}, D \leftarrow \mathcal{D}}[\mu_n(S, D)].$$

Later, we will use the following lemma, which essentially states that we can approximate a distribution over circuits by a relatively small circuit, which is constructed by sampling from the distribution and providing a "good" set of samples to the circuit as non-uniform advice.

Lemma 19 (Approximating a distribution over circuits by a small circuit obtained via sampling)**.** *Let X and A be finite sets, let Y be any random variable with finite support, let $\mathcal{C}$ be any distribution over s-size randomized circuits of the form $C : X \times Supp(Y) \to A$, and let U be any finite set of randomized circuits of the form $u : X \times Supp(Y) \times A \to \{0,1\}$. Then, for every $\epsilon > 0$, there exists a randomized circuit $\widehat{C}$ of size $O(\frac{\log|X|+\log|U|}{\epsilon^2} \cdot s)$ such that for every $u \in U$ and $x \in X$, we have*

$$|\mathbb{E}_{C \leftarrow \mathcal{C}}[u(x,Y,C(x,Y))] - \mathbb{E}[u(x,Y,\widehat{C}(x,Y))]| \leq \epsilon.$$

The lemma follows easily from a Chernoff bound and a union bound; see Appendix C for the proof. This proof of the main theorem follows from three relatively simple steps, which we now describe at a high level.

Step 1. We first show that for any mixed strategy $\mathcal{D}$ for Adv (i.e., any distribution over $t'(n)$-size distinguishers), there exists a $T(n)$-size simulator $S_{\mathcal{D}} \in Strat_{\mathsf{Sim}}$ such that $\mu_n(S_{\mathcal{D}}, \mathcal{D}) \geq -3\epsilon(n)/2$. By Lemma 19, we can approximate $\mathcal{D}$ by a $t(n)$-size distinguisher $\widehat{D}$, and then use the super-weak distributional (T,t,ϵ)-zero-knowledge property of (P,V) to get a $T(n)$-size simulator $S_{\widehat{D}}$ for $\widehat{D}$ such that $\mu_n(S_{\widehat{D}}, \widehat{D}) \geq -\epsilon(n)$. Since $\widehat{D}$ approximates $\mathcal{D}$ to within $\epsilon(n)/2$, we have $\mu_n(S_{\widehat{D}}, \mathcal{D}) \geq -3\epsilon(n)/2$, as required.

Step 2. We now apply the minimax theorem to the result of Step 1 to get a mixed strategy $\mathcal{S}$ for Sim (i.e., a distribution over $T(n)$-size simulators) such that for every $t'(n)$-size distinguisher $D \in Strat_{\mathsf{Adv}}$, we have $\mu_n(\mathcal{S}, D) \geq -3\epsilon(n)/2$.

Step 3. By Lemma 19, we can approximate $\mathcal{S}$ (from Step 2) by a $T'(n)$-size simulator $\widehat{S}$ so that $\mu_n(\widehat{S}, D) \geq -2\epsilon(n)$ for every $t'(n)$-size distinguisher $D \in Strat_{\mathsf{Adv}}$.

The result of Step 3 shows Claim 18, which completes the proof of the theorem. We now provide the details for Steps 1 and 3.

Details of Step 1. By Lemma 19 (in the statement of the lemma, we let $X = Supp(X_n) \times Supp(Z_n) \times \{0,1\}^{t'(n)}$, $A = \{0,1\}$, $Y = 0$, $\mathcal{C} = \mathcal{D}$, U be a set containing only the circuit $(x,y,a) \mapsto a$, and $\epsilon = \epsilon(n)/2$), there exists a distinguisher $\widehat{D}$ of size $O((n+t'(n))^2/\epsilon(n)^2) = t(n)$ such that for every $x \in X_n$, $z \in Z_n$, and $v \in \{0,1\}^{t'(n)}$, we have $|\Pr_{D \leftarrow \mathcal{D}}[D(x,z,v) = 1] - \Pr[\widehat{D}(x,z,v) = 1]| \leq \epsilon(n)/2$. Since (P,V) is super-weak distributional (T,t,ϵ)-zero-knowledge, there exists a $T(n)$-size simulator $S_{\widehat{D}}$ such that $\mu_n(S_{\widehat{D}}, \widehat{D}) \geq -\epsilon(n)$. From the result above and the definition of μ_n, we have $|\mu_n(S_{\widehat{D}}, \mathcal{D}) - \mu_n(S_{\widehat{D}}, \widehat{D})| \leq \epsilon(n)/2$, so $\mu_n(S_{\widehat{D}}, \mathcal{D}) \geq -3\epsilon(n)/2$, as required.

Details of Step 3. By Lemma 19, there exists a simulator $\widehat{S}$ of size $O((\log|Strat_{\mathsf{Adv}}|/\epsilon(n)^2) \cdot T(n))$ such that for every $t'(n)$-size distinguisher $D \in Strat_{\mathsf{Adv}}$, we have $|\Pr_{S \leftarrow \mathcal{S}}[D(X_n, Z_n, S(X_n, Z_n)) = 1] - \Pr[D(X_n, Z_n, \widehat{S}(X_n, Z_n)) = 1]| \leq \epsilon(n)/2$, which implies $|\mu_n(\mathcal{S}, D) - \mu_n(\widehat{S}, D)| \leq \epsilon(n)/2$. Combining this with the result of Step 2, we have $\mu_n(\widehat{S}, D) \geq -2\epsilon(n)$ for every $t'(n)$-size distinguisher $D \in Strat_{\mathsf{Adv}}$. Furthermore, the simulator $\widehat{S}$ has size at most $T'(n)$, since there are at most $O(q(n)+t'(n))^{O(t'(n))}$ circuits of size $t'(n)$ on $q(n)$ input bits, so $|Strat_{\mathsf{Adv}}| \leq O(n+t'(n))^{O(t'(n))}$. □

4.1 Laconic Prover Implies Distributional (T, t, ϵ)-Zero-Knowledge

In this section, we first use Theorem 17 to show that an interactive proof with short prover communication complexity implies distributional (T,t,ϵ)-zero-knowledge. We then describe applications of this result.

Theorem 20. *Let (P,V) be an interactive proof system for a language L, and suppose that the prover P has communication complexity $\ell(n)$, i.e., the total length of the messages sent by P is $\ell(n)$, where n is the length of the common input x. Then, for every function $t(n) \geq \Omega(n)$ and $\epsilon(n)$, (P,V) is distributional (T,t,ϵ)-zero-knowledge, where $T(n) = O\left(2^{\ell(n)} \cdot \frac{t(n)^3 \ln(t(n))}{\epsilon(n)^6} \cdot \log(\frac{1}{\epsilon(n)})\right)$.*

Proof. By Theorem 17, it suffices to show that (P,V) is super-weak distributional $(T',t',\epsilon/2)$-zero-knowledge, where $t'(n) = \Theta(\frac{t(n)^2}{\epsilon(n)^2})$ and $T'(n) = O(2^{\ell(n)} \cdot \frac{t(n)^2}{\epsilon(n)^4} \cdot \log(\frac{1}{\epsilon(n)}))$. Let $n \in \mathbb{N}$, let (X_n, Y_n, Z_n) be a joint distribution over $(L \cap \{0,1\}^n) \times \{0,1\}^* \times \{0,1\}^*$, let V^* be any $t'(n)$-size adversary, and let D be any $t'(n)$-size distinguisher. For a sequence of messages $(m_1,\ldots,m_k)$, let $(m_1,\ldots,m_k) \leftrightarrow V^*(x,z)$ denotes the protocol where the prover sends the message m_i to V^* in round i.

Let S be the simulator that, on input (x,z), does the following for each of the $2^{\ell(n)}$ possible sequences of messages $(m_1,\ldots,m_k)$ of total length $\ell(n)$ (that the prover P may possibly send): S estimates $\Pr[D(x,z,Out_{V^*}[(m_1,\ldots,m_k) \leftrightarrow V^*(x,z)]) = 1]$ by running V^* and D $O(\frac{1}{\epsilon(n)^2} \cdot \log(\frac{1}{\epsilon(n)}))$ many times so that with probability at least $1-\epsilon(n)/4$, the error of the estimate is at most $\epsilon(n)/4$. Then, S outputs $Out_{V^*}[m^* \leftrightarrow V^*(x,z)]$, where m^* is the sequence of messages $(m_1,\ldots,m_k)$ that had the highest estimated value for $\Pr[D(x,z,Out_{V^*}[(m_1,\ldots,m_k) \leftrightarrow V^*(x,z)]) = 1]$. It is easy to see that for every $(x,y,z) \in (X_n,Y_n,Z_n)$, we have

$$\Pr[D(x,z,Out_{V^*}[P(x) \leftrightarrow V^*(x,z)]) = 1] - \Pr[D(x,z,S(x,z)) = 1] \leq \epsilon(n)/2.$$

Furthermore, we note that the size of the simulator S is $O(2^{\ell(n)} \cdot \frac{1}{\epsilon(n)^2} \log(\frac{1}{\epsilon(n)}) \cdot t'(n)) = T'(n)$. Thus, (P,V) is super-weak distributional $(T',t',\epsilon/2)$-zero-knowledge, as required. □

Let us now provide a few corollaries of Theorem 20. The first two are new proofs of old theorems (with some new generalizations). The third one is a new result on 2-round zero-knowledge.

4.1.1 Application 1: Leakage Lemma of Gentry-Wichs

Roughly speaking, the "Leakage Lemma" of Gentry-Wichs [GW11] states that for every joint distribution $(X,\pi(X))$, where $|\pi(x)| = O(\log|x|)$ (π should be thought of as leakage on X), and for every distribution Y that is indistinguishable from X, there exists some leakage $\widetilde{\pi}$ such that the joint distributions $(X,\pi(X))$ and $(Y,\widetilde{\pi}(Y))$ are indistinguishable. We now show that this result follows as a simple corollary of Theorem 20.

Two distributions X and Y are (s,ϵ)-*indistinguishable* if every s-size circuit C can only distinguish X from Y by at most ϵ, i.e., $|\Pr[C(X)=1] - \Pr[C(Y)=1]| \leq \epsilon$.

Corollary 21 (The leakage lemma of Gentry-Wichs [GW11]). *Let $(X,\pi(X))$ be any joint distribution, where $|\pi(X)| \leq \ell$. Let Y be any distribution that is (s,ϵ)-indistinguishable from X. Then, there exists a joint distribution $(Y,\widetilde{\pi}(Y))$ such that $(X,\pi(X))$ and $(Y,\widetilde{\pi}(Y))$ are $(s',2\epsilon)$-indistinguishable, where $s' = \Omega\left(\sqrt[3]{\frac{\epsilon^6 \cdot s}{2^\ell \cdot \ln(1/\epsilon) \cdot \ln(s)}}\right)$.*

Proof. Let $L = \{0,1\}^*$ be the trivial language with the trivial witness relation $R_L(x) = \{0,1\}^*$. Let (P,V) be an interactive proof system for L where the prover P, on input a statement x with witness y, simply sends the first ℓ bits of y to the verifier V, who simply always accepts. By Theorem 20, (P,V) is distributional (T,s',ϵ)-zero-knowledge, where $T \leq s/2$. By considering the statement distribution X with witness distribution $\pi(X)$, it follows that there exists a T-size simulator S such that $(X,\pi(X))$ and $(X,S(X))$ are (s',ϵ)-indistinguishable. Also, $(X,S(X))$ and $(Y,S(Y))$ are $(s/2,\epsilon)$-indistinguishable, since X and Y are (s,ϵ)-indistinguishable and $T \leq s/2$. It follows that $(X,\pi(X))$ and $(Y,S(Y))$ are $(s',2\epsilon)$-indistinguishable, so letting $\widetilde{\pi} = S$ yields the result. □

Let us note that our proof of the leakage lemma yields an even stronger statement—namely, we have found an efficient simulator $\widetilde{\pi}$; such a version of the leakage lemma was recently established by Jetchev and Pietrzak [JP12]. (As an independent contribution, our proof of Theorem 4 is actually significantly simpler than both the proof of [GW11] and [JP12].) Additionally, since our result on zero-knowledge applies also to *interactive* protocols, we directly also get an interactive version of the leakage lemma.

4.1.2 Application 2: Dense Model Theorem

We proceed to show that the *dense model theorem* (e.g., see [RTTV08, TTV09, DP08]) follows as a corollary of Theorem 20. A distribution R is δ*-dense* in a distribution X if for every r, $\Pr[R = r] \leq \frac{1}{\delta}\Pr[X = r]$. Equivalently, R is δ-dense in X if there exists a joint distribution $(X, B(X))$ with $\Pr[B(X) = 1] \geq \delta$ such that $R = X|(B(X) = 1)$. Let U_n be the uniform distribution over $\{0,1\}^n$.

Corollary 22 (**The dense model theorem**). *Let X be any distribution over $\{0,1\}^n$ that is (s,ϵ)-indistinguishable from U_n, and suppose R is δ-dense in X. Then, there exists a distribution M that is $(\delta - 2\epsilon)$-dense in U_n, and M and R are $(s', \frac{2\epsilon}{\delta})$-indistinguishable, where $s' = \Omega\left(\sqrt[3]{\frac{\epsilon^6 \cdot s}{\ln(1/\epsilon)\cdot\ln(s)}}\right)$.*

Proof. Since R is δ-dense in X, there exists a joint distribution $(X, B(X))$ with $\Pr[B(X) = 1] \geq \delta$ such that $R = X|(B(X) = 1)$. Without loss of generality, we can assume that $B(X)$ is always either 0 or 1. Let $L = \{0,1\}^*$ be the trivial language with the trivial witness relation $R_L(x) = \{0,1\}^*$. Let (P,V) be an interactive proof system for L where the prover P, on input a statement x with witness y, simply sends the first bit of y to the verifier V, who simply always accepts. By Theorem 20, (P,V) is distributional $(T, 2s', \epsilon)$-zero-knowledge, where $T \leq s/2$. By considering the statement distribution X with witness distribution $B(X)$, it follows that there exists a T-size simulator S such that $(X, B(X))$ and $(X, S(X))$ are $(2s', \epsilon)$-indistinguishable. Also, $(X, S(X))$ and $(U_n, S(U_n))$ are $(s/2, \epsilon)$-indistinguishable, since X and U_n are (s,ϵ)-indistinguishable and $T \leq s/2$. It follows that $(X, B(X))$ and $(U_n, S(U_n))$ are $(2s', 2\epsilon)$-indistinguishable. Thus, $\Pr[S(U_n) = 1] \geq \delta - 2\epsilon$ (since $\Pr[B(X) = 1] \geq \delta$), so $U_n|(S(U_n) = 1)$ is $(\delta - 2\epsilon)$-dense in U_n. Also, $X|(B(X) = 1)$ and $U_n|(S(U_n) = 1)$ are $(s', 2\epsilon/\delta)$-indistinguishable, so letting $M = U_n|(S(U_n) = 1)$ yields the result. □

4.1.3 Application 3: 2-Round ZK

A final corollary of Theorem 20 is that the classic two-round graph non-isomorphism protocol (which is only known to be honest-verifier zero-knowledge) is also distributional (T, t, ϵ)-zero-knowledge for $T(n) = polylog(t(n), \frac{1}{\epsilon(n)})$.[6] In fact, by the complete problem for SZK [SV03], we can show that every language in SZK has a 2-round distributional (T, t, ϵ) zero-knowledge proof for $T(n) = polylog(t(n), \frac{1}{\epsilon(n)})$.

Theorem 23. *For every language $L \in SZK$ and every function $\delta(n) \geq \frac{1}{2^{poly(n)}}$, there exists a two-round interactive proof (P,V) for L with completeness $1 - negl(n)$ and soundness error $\delta(n)$ such that for every function t and ϵ, (P,V) is distributional (T, t, ϵ)-zero-knowledge, where $T(n) = polylog(\frac{1}{\delta(n)}, t(n), \frac{1}{\epsilon(n)})$.*

Proof. From [SV03], there exists a two-round interactive proof (P', V') for a complete problem L_{SZK} for SZK with completeness negligibly close to 1 and soundness error negligibly close to $\frac{1}{2}$,

[6] Recall that in the classic GNI protocol the prover sends just a single bit.

and the prover P' only sends a single bit to the verifier V'. By repeating the proof in parallel $O(\log \frac{1}{\delta(n)})$ times, we get a two-round interactive proof for L_{SZK} with completeness negligibly close to 1 and soundness error $\delta(n)$, and the prover only sends $O(\log \frac{1}{\delta(n)})$ bits to the verifier. Then, by Theorem 20, this interactive proof for L_{SZK} is distributional (T, t, ϵ)-zero-knowledge, where $T(n) = polylog(\frac{1}{\delta(n)}, t(n), \frac{1}{\epsilon(n)})$. Since L_{SZK} is a complete problem for SZK, the theorem follows. □

In Theorem 23, if we choose $\delta(n) = \frac{1}{2^n}$, $t(n) = n^{\log n}$, and $\epsilon(n) = \frac{1}{n^{\log n}}$, then every language in SZK has a 2-round "quasi-polynomial-time simulatable" distributional zero-knowledge proof (i.e., $T(n)$ is a quasi-polynomial) with completeness $1 - negl(n)$ and negligible soundness error. Alternatively, if we choose $\delta(n) = \frac{1}{poly(n)}$, $t(n) = poly(n)$, and $\epsilon(n) = \frac{1}{poly(n)}$, then every language in SZK has a 2-round "polynomial-time simulatable" (T, t, ϵ)-distributional zero-knowledge proof (i.e., $T(n)$ is a polynomial) with completeness $1 - negl(n)$ and soundness error $\frac{1}{poly(n)}$.

4.2 Necessity of Non-Black-Box Simulation

The universal simulator in Theorem 23 is obtained via Theorem 20, which uses Theorem 17, so the universal simulator makes non-black-box usage of V^*. We remark that this non-black-box usage is also necessary (assuming standard complexity theoretic assumptions): We will show that black-box simulation techniques cannot be used to demonstrate distributional (T, t, ϵ)-zero-knowledge for 2-round proof systems for languages that are hard-on-average. Thus, as long as SZK contains a problem that is hard-on-average, our non-black-box techniques are necessary. Let us first give the definition of black-box distributional (T, t, ϵ)-zero-knowledge.

Definition 24 (black-box distributional (T, t, ϵ)-zero-knowledge). Let (P, V) be an interactive proof system for a language L. We say that (P, V) is *black-box distributional (T, t, ϵ)-zero-knowledge* if for every $n \in \mathbb{N}$ and every joint distribution (X_n, Y_n, Z_n) over $(L \cap \{0,1\}^n) \times \{0,1\}^* \times \{0,1\}^*$, there exists a $T(n)$-size simulator S such that for every $t(n)$-size adversary V^* and every $t(n)$-size distinguisher D, we have

$$|\Pr[D(X_n, Z_n, Out_{V^*}[P(X_n, Y_n) \leftrightarrow V^*(X_n, Z_n)]) = 1] - \Pr[D(X_n, Z_n, S^{V^*(X_n, Z_n)}(X_n, Z_n)) = 1]| \leq \epsilon(n).$$

where $S^{V^*(X_n, Z_n)}$ means that S is given oracle access to the verifier $V^*(X_n, Z_n)$.

For any language L and any $x \in \{0,1\}^*$, let $L(x) = 1$ if $x \in L$, and $L(x) = 0$ otherwise. We now show that any 2-round interactive proof for a language L with "hard-on-average" instances is not black-box distributional zero-knowledge.

Theorem 25. *Let L be any language with hard-on-average instances, i.e., there exists an ensemble $\{X_n\}_{n \in \mathbb{N}}$ of distributions X_n over $\{0,1\}^n$ such that for every non-uniform PPT algorithm A and for sufficiently large $n \in \mathbb{N}$, we have $\Pr[A(X_n) = L(X_n)] \leq \frac{1}{2} + \epsilon(n)$, where ϵ is any function such that $\epsilon(n) < \frac{1}{12}$ for sufficiently large $n \in \mathbb{N}$.*

Then, there exists a polynomial t such that any 2-round interactive proof (P, V) for L with completeness $\frac{2}{3}$ and soundness error at most $\frac{1}{3}$ is not black-box (T, t, ϵ)-distributional zero-knowledge for any polynomial T.

Proof. Let $t(n) = O(T_V(n))$, where $T_V(n)$ is a polynomial bound on the running time of V on instances x of length n. To obtain a contradiction, suppose (P, V) is black-box (T, t, ϵ)-distributional zero-knowledge for some polynomial T. Let $n \in \mathbb{N}$, let X'_n be X_n conditioned on the event $X_n \in L$, let X''_n be X_n conditioned on the event $X_n \notin L$, let Y_n always be the empty string, and let Z_n be the uniform distribution over $\{0,1\}^{t(n)}$. Then, there exists a polynomial-size simulator S such that for every $t(n)$-size adversary V^* and every $t(n)$-size distinguisher D, we have

$$|\Pr[D(X'_n, Z_n, Out_{V^*}[P(X'_n, Y_n) \leftrightarrow V^*(X'_n, Z_n)]) = 1] - \Pr[D(X'_n, Z_n, S^{V^*(X'_n, Z_n)}(X'_n)) = 1]| \leq \epsilon(n). \tag{1}$$

Let V^* be the verifier that, on input (x, z), runs the honest verifier $V_z(x)$ with random tape z to interact with the prover, and then outputs the message a received from the prover. Let D be the distinguisher that, on input (x, z, a), outputs 1 if $V_z(x, a) = 1$, and 0 otherwise, where $V_z(x, a)$ represents the output of $V(x)$ with random tape z and with message a received from the prover.

Claim 26. *$\Pr[D(X'_n, Z_n, S^{V^*(X'_n, Z_n)}(X'_n)) = 1] \geq \frac{2}{3} - \epsilon(n)$.*

Proof of claim. Since (P, V) has completeness $\frac{2}{3}$, we have

$$\begin{aligned} & \Pr[D(X'_n, Z_n, Out_{V^*}[P(X'_n, Y_n) \leftrightarrow V^*(X'_n, Z_n)]) = 1] \\ = \ & \Pr[Out_V[P(X'_n, Y_n) \leftrightarrow V(X'_n)] = 1] \\ \geq \ & \frac{2}{3}. \end{aligned}$$

Now, combining this with (1), we have

$$\Pr[D(X'_n, Z_n, S^{V^*(X'_n, Z_n)}(X'_n)) = 1] \geq \frac{2}{3} - \epsilon(n),$$

as required. This completes the proof of the claim. □

Claim 27. *$\Pr[D(X''_n, Z_n, S^{V^*(X''_n, Z_n)}(X''_n)) = 0] \geq \frac{2}{3} - \epsilon(n)$.*

Proof of claim. To obtain a contradiction, suppose $\Pr[D(X''_n, Z_n, S^{V^*(X''_n, Z_n)}(X''_n)) = 0] < \frac{2}{3} - \epsilon(n)$. We note that the event $D(X''_n, Z_n, S^{V^*(X''_n, Z_n)}(X''_n)) = 0$ occurs if and only if the event $V_{Z_n}(X''_n, S^{V^*(X''_n, Z_n)}(X''_n)) = 0$ occurs, where $V_{Z_n}(X''_n, S^{V^*(X''_n, Z_n)}(X''_n))$ represents the output of $V(X''_n)$ with random tape Z_n and with message $S^{V^*(X''_n, Z_n)}(X''_n)$ received from the prover. Thus, we have $\Pr[V_{Z_n}(X''_n, S^{V^*(X''_n, Z_n)}(X''_n)) = 0] < \frac{2}{3} - \epsilon(n)$.

Now, consider an adversarial prover P^* that, on input x and upon receiving a message c from the verifier V, simulates $S(x)$ while responding to oracle queries with the message c, and then sends the output of $S(x)$ to V. Now, we note that the event $V_{Z_n}(X''_n, S^{V^*(X''_n, Z_n)}(X''_n)) = 0$ occurs if and only if the event $Out_V(P^*(X''_n, Y_n) \leftrightarrow V_{Z_n}(X''_n)) = 0$ occurs. Thus, we have

$$\Pr[Out_V(P^*(X''_n, Y_n) \leftrightarrow V_{Z_n}(X''_n)) = 0] < \frac{2}{3} - \epsilon(n),$$

and since we always have $X''_n \notin L$, this contradicts the assumption that (P, V) has soundness error at most $\frac{1}{3}$. This completes the proof of the claim. □

Now, using the polynomial-size simulator S and the $t(n)$-size distinguisher D, we will construct a non-uniform PPT algorithm A that contradicts the assumption that L has hard-on-average instances, i.e., for infinitely many $n \in \mathbb{N}$, we have

$$\Pr[A(X_n) = L(X_n)] > \frac{1}{2} + \epsilon(n).$$

Let A be the non-uniform PPT algorithm that, on input $x \in \{0,1\}^n$, samples a uniformly random z from Z_n, computes $S^{V^*(x,z)}(x)$ (while simulating the oracle $V^*(x,z)$ for $S(x)$) and outputs $D(x,z,S^{V^*(x,z)}(x))$. Then, for infinitely many $n \in \mathbb{N}$, we have

$$\begin{aligned}
&\Pr[A(X_n) = L(X_n)] \\
=\ &\Pr[D(X_n, Z_n, S^{V^*(X_n,Z_n)}(X_n)) = L(X_n)] \\
=\ &\Pr[X_n \in L] \cdot \Pr[D(X'_n, Z_n, S^{V^*(X'_n,Z_n)}(X'_n)) = 1] + \Pr[X_n \notin L] \cdot \Pr[D(X''_n, Z_n, S^{V^*(X''_n,Z_n)}(X''_n)) = 0] \\
\geq\ &\Pr[X_n \in L] \cdot (2/3 - \epsilon(n)) + \Pr[X_n \notin L] \cdot (2/3 - \epsilon(n)) \\
=\ &\frac{2}{3} - \epsilon(n),
\end{aligned}$$

where the inequality follows from the two claims above. This contradicts the assumption that L has hard-on-average instances. This completes the proof. □

References

[Bar01] Boaz Barak, *How to go beyond the black-box simulation barrier*, FOCS, 2001, pp. 106–115.

[BN00] Dan Boneh and Moni Naor, *Timed commitments*, Proceedings of the 20th Annual International Cryptology Conference on Advances in Cryptology, CRYPTO '00, Springer-Verlag, 2000, pp. 236–254.

[BP12] Nir Bitansky and Omer Paneth, *From the impossibility of obfuscation to a new non-black-box simulation technique*, FOCS, 2012.

[BP13] ———, *On the impossibility of approximate obfuscation and applications to resettable cryptography*, STOC, 2013.

[BS05] Boaz Barak and Amit Sahai, *How to play almost any mental game over the net - concurrent composition via super-polynomial simulation*, FOCS, 2005, pp. 543–552.

[BSW03] Boaz Barak, Ronen Shaltiel, and Avi Wigderson, *Computational analogues of entropy*, RANDOM-APPROX, 2003, pp. 200–215.

[CKLR11] Kai-Min Chung, Yael Tauman Kalai, Feng-Hao Liu, and Ran Raz, *Memory delegation*, Proceedings of the 31st annual conference on Advances in cryptology (Berlin, Heidelberg), CRYPTO'11, Springer-Verlag, 2011, pp. 151–165.

[CLP13] Kai-Min Chung, Edward Lui, and Rafael Pass, *Can theories be tested? A cryptographic treatment of forecast testing*, ITCS, 2013, pp. 47–56.

[CPS13] Kai-Min Chung, Rafael Pass, , and Karn Seth, *Non-black-box simulation from one-way functions and applications to resettable security*, STOC, ACM, 2013.

[DGS09] Y. Deng, V. Goyal, and A. Sahai, *Resolving the simultaneous resettability conjecture and a new non-black-box simulation strategy*, Foundations of Computer Science, 2009. FOCS'09. 50th Annual IEEE Symposium on, IEEE, 2009, pp. 251–260.

[DNRS03] Cynthia Dwork, Moni Naor, Omer Reingold, and Larry Stockmeyer, *Magic functions: In memoriam: Bernard m. dwork 1923–1998*, J. ACM **50** (2003), no. 6, 852–921.

[DP08] Stefan Dziembowski and Krzysztof Pietrzak, *Leakage-resilient cryptography*, In 49th FOCS, IEEE Computer Society Press, 2008, pp. 293–302.

[FK99] Alan Frieze and Ravi Kannan, *Quick approximation to matrices and applications*, Combinatorica **19** (1999), no. 2, 175–220.

[FS99] Yoav Freund and Robert E. Schapire, *Adaptive game playing using multiplicative weights*, Games and Economic Behavior **29** (1999), no. 1, 79–103.

[GMR85] S Goldwasser, S Micali, and C Rackoff, *The knowledge complexity of interactive proof-systems*, Proceedings of the seventeenth annual ACM symposium on Theory of computing, STOC '85, ACM, 1985, pp. 291–304.

[GMW91] Oded Goldreich, Silvio Micali, and Avi Wigderson, *Proofs that yield nothing but their validity or all languages in np have zero-knowledge proof systems*, J. ACM **38** (1991), no. 3, 690–728.

[Gol93] Oded Goldreich, *A uniform-complexity treatment of encryption and zero-knowledge*, Journal of Cryptology **6** (1993), 21–53.

[GVW01] Oded Goldreich, Salil P. Vadhan, and Avi Wigderson, *On interactive proofs with a laconic prover*, Proceedings of the 28th International Colloquium on Automata, Languages and Programming, (London, UK, UK), ICALP '01, Springer-Verlag, 2001, pp. 334–345.

[GW11] Craig Gentry and Daniel Wichs, *Separating succinct non-interactive arguments from all falsifiable assumptions*, Proceedings of the 43rd annual ACM symposium on Theory of computing (New York, NY, USA), STOC '11, ACM, 2011, pp. 99–108.

[HP10] Joe Halpern and Rafael Pass, *Game theory with costly computation: formulation and application to protocol security*, Proceedings of the Behavioral and Quantitative Game Theory: Conference on Future Directions (New York, NY, USA), BQGT '10, ACM, 2010, pp. 89:1–89:1.

[Imp95] R. Impagliazzo, *Hard-core distributions for somewhat hard problems*, Proceedings of the 36th Annual Symposium on Foundations of Computer Science, FOCS '95, IEEE Computer Society, 1995, pp. 538–.

[JP12] Dimitar Jetchev and Krzysztof Pietrzak, *How to fake auxiliary input*, Presentation, 2012.

[Pas04] Rafael Pass, *Bounded-concurrent secure multi-party computation with a dishonest majority*, STOC '04, 2004, pp. 232–241.

[PR03] Rafael Pass and Alon Rosen, *Bounded-concurrent secure two-party computation in a constant number of rounds*, FOCS, 2003, pp. 404–413.

[PR05] ———, *New and improved constructions of non-malleable cryptographic protocols*, STOC '05, 2005, pp. 533–542.

[RTTV08] Omer Reingold, Luca Trevisan, Madhur Tulsiani, and Salil Vadhan, *Dense subsets of pseudorandom sets*, Proceedings of the 2008 49th Annual IEEE Symposium on Foundations of Computer Science, FOCS '08, 2008, pp. 76–85.

[SV03] Amit Sahai and Salil Vadhan, *A complete problem for statistical zero knowledge*, J. ACM **50** (2003), no. 2, 196–249.

[TTV09] Luca Trevisan, Madhur Tulsiani, and Salil Vadhan, *Regularity, boosting, and efficiently simulating every high-entropy distribution*, Proceedings of the 2009 24th Annual IEEE Conference on Computational Complexity, CCC '09, IEEE Computer Society, 2009, pp. 126–136.

[Yao82] Andrew C. Yao, *Theory and application of trapdoor functions*, Proceedings of the 23rd Annual Symposium on Foundations of Computer Science, SFCS '82, 1982, pp. 80–91.

Appendix A Proof of Theorem 9

In this section, we prove Theorem 9, which we restate for convenient reference.

Theorem 9. *Under reasonable cryptographic assumptions, there exists an interactive proof system (P, V) for an NP language L such that (P, V) is weak zero-knowledge but not zero-knowledge.*

We begin by describing the assumptions we make regarding the existence of certain cryptographic primitives. We first make the following assumption regarding the existence of a collection of two-round "timed" commitment schemes (see [BN00]) satisfying certain properties:

- There exists a polynomial $p(\cdot)$ and a negligible function $\nu(\cdot)$ such that for every sufficiently large $n \in \mathbb{N}$, there exists a collection of two-round commitment schemes $\{\mathsf{Com}_i\}_{i\in[\ell]}$, where $\ell = \log^2 n$, such that for every $i \in [\ell]$, Com_i is hiding with respect to adversaries running in $p(n)^{i-1}$ steps, but Com_i can always be broken in $p(n)^i$ steps, i.e., the following two properties hold:
  - [Hiding for adversaries running in $p(n)^{i-1}$ steps] For every adversary D running in $p(n)^{i-1}$ steps, for every $x, x' \in \{0,1\}^n$, and for every possible first message ρ for Com_i, we have
  $$|\Pr[D(\mathsf{Com}_i(x,\rho)) = 1] - \Pr[D(\mathsf{Com}_i(x',\rho)) = 1]| \le \nu(n),$$
  where $\mathsf{Com}_i(v, r)$ denotes the commitment of v using Com_i with first message r.
  - [Can always be broken in $p(n)^i$ steps] There exists an algorithm A running in $p(n)^i$ steps such that for every $x \in \{0,1\}^n$ and every ρ, we have $\Pr[A(\mathsf{Com}_i(x,\rho)) = x] = 1$.

One can use the timed commitment scheme in [BN00] to get such a collection of commitment schemes. Let $p(\cdot)$ be the polynomial described above. We also make the following assumption that there exists a collection of (length-preserving) "timed" worst-case weak one-way permutations satisfying certain properties:

- There exists a polynomial $q(\cdot)$ such that for every sufficiently large $n \in \mathbb{N}$, there exists a collection of worst-case weak one-way permutations $\{f_i : \{0,1\}^n \to \{0,1\}^n\}_{i\in[\ell]}$, where $\ell = \log^2 n$, such that for every $i \in \mathbb{N}$, f_i is somewhat hard to break in $p(n)^{i+1}$ steps in the worst case, but can always be broken in $p(n)^{i+2}$ steps, i.e., the following two properties hold:
  - [Somewhat hard to break in $p(n)^{i+1}$ steps in the worst case] For every adversary A running in $p(n)^{i+1}$ steps, there exists an $x \in \{0,1\}^n$ such that $\Pr[A(f_i(x)) = x] \le 1 - \frac{1}{q(n)}$.
  - [Can always be broken in $p(n)^{i+2}$ steps] There exists an algorithm A running in $p(n)^{i+2}$ steps such that for every $x \in \{0,1\}^n$, we have $\Pr[A(f_i(x)) = x] = 1$.

We note that f_i is slightly harder to break than Com_i, which is a property we will use later in our proof. We now describe a language L and an interactive proof system (P, V) for L that is weak zero-knowledge but not zero-knowledge. Let L be the trivial language $\{0,1\}^*$ with witness relation R_L defined by $R_L(x) = \{(f_1^{-1}(x), \ldots, f_\ell^{-1}(x)) : \ell = \log^2 |x|\}$ for every $x \in \{0,1\}^*$.

Let $(P(x, y), V(x))$ be the following interactive proof, where $x \in \{0,1\}^*$, $n = |x|$, $\ell = \log^2 n$, and $y = (f_1^{-1}(x), \ldots, f_\ell^{-1}(x))$:

1. The verifier V generates and sends ρ_i for $i = 1, \ldots, \ell$ to the prover, where ρ_i is the first message of an execution of Com_i.

2. The prover P sends $\mathsf{Com}_i(f_i^{-1}(x), \rho_i)$ for $i = 1, \ldots, \ell$ to the verifier, where $\mathsf{Com}_i(v, r)$ denotes the commitment of v using Com_i with first message r.

3. The verifier V accepts (i.e., outputs 1).

We first show that (P, V) is weak zero-knowledge.

Lemma 28. *The interactive protocol (P, V) is weak zero-knowledge.*

Proof. Let V^* be any PPT adversary, and let D be any PPT distinguisher. Let T_{V^*} and T_D be polynomials that bound the running time of V^* and D, respectively, and let $T = O(T_{V^*} + T_D)$. Let S be a PPT simulator that does the following on input (x, z), where $x, z \in \{0,1\}^*$, $n = |x|$, and $\ell = \log^2 n$:

1. Run $V^*(x, z)$ to get ρ_i for $i = 1, \ldots, \ell$.

2. Let j be the smallest integer such that $p(n)^{j-1} > T(n)$. For $i = 1, \ldots, j-1$, use $p(n)^{i+2} = poly(n)$ steps to break f_i to get $f_i^{-1}(x)$.

3. For $i = 1, \ldots, j-1$, send $\mathsf{Com}_i(f_i^{-1}(x), \rho_i)$ to V^*. For $i = j, \ldots, \ell$, send $\mathsf{Com}_i(0^n, \rho_i)$ to V^*.

4. Continue running $V^*(x, z)$ and output whatever V^* outputs.

It is easy to see that the simulator S runs in polynomial time. We now claim that the simulator S works. To see this, consider a "hybrid" simulator S', where S' is the same as S except that for $i = j, \ldots, \ell$, S' sends $\mathsf{Com}_i(f_i^{-1}(x), \rho_i)$ to V^* instead of $\mathsf{Com}_i(0^n, \rho_i)$.

Consider the probability $\Pr[D(x, z, S(x, z)) = 1]$. Since the hiding property of Com_i for $i = j, \ldots, \ell$ is hard to break in $p(n)^{i-1} \geq p(n)^{j-1}$ steps, and since $T(n) < p(n)^{j-1}$, it is easy to verify that the probability $\Pr[D(x, z, S(x, z)) = 1]$ is negligibly close to $\Pr[D(x, z, S'(x, z)) = 1]$. Now, we note that the message sent by S' to V^* has the exact same distribution as the message sent by the prover P. Thus, $\Pr[D(x, z, S'(x, z)) = 1]$ is equal to $\Pr[D(x, z, Out_{V^*}[P(x) \leftrightarrow V^*(x, z)]) = 1]$, so $\Pr[D(x, z, S(x, z)) = 1]$ is negligibly close to $\Pr[D(x, z, Out_{V^*}[P(x) \leftrightarrow V^*(x, z)]) = 1]$. Thus, (P, V) is weak zero-knowledge, as required. □

We now show that (P, V) is not zero-knowledge.

Lemma 29. *The interactive protocol (P, V) is not zero-knowledge.*

Proof. Let V^* be the same as V except that at the end, V^* outputs its view. To obtain a contradiction, suppose that a PPT simulator S for V^* exists, and suppose the running time of S is bounded by n^d for some constant $d \geq 1$. Now, let D be the distinguisher that, on input $x \in \{0,1\}^n$, $z \in \{0,1\}^*$, and a view of V^*, does the following:

1. Let j be the smallest positive integer such that $p(n)^j$ is greater than Cn^d, where $C \geq 2$ is some universal constant.

2. Use $p(n)^j = \text{poly}(n)$ steps to break $\mathsf{Com}_j(w_j, \rho_j)$ in the view of V^* to get w_j.

3. Output 1 if $f_j(w_j) = x$, and output 0 otherwise.

Using the simulator S, we will construct an adversary A that breaks f_j. The fact that S works for V^* and the distinguisher D will ensure that with overwhelming probability, the output of $S(x, z)$ will contain a commitment $\mathsf{Com}_j(w_j, \rho'_j)$ such that $f_j(w_j) = x$. The adversary A, on input $f_j(w)$ for some w, will run the simulator $S(f_j(w))$ to get this commitment $\mathsf{Com}_j(w_j, \rho'_j)$, and then break it using $p(n)^j$ steps to get w_j, and then output w_j. Now, it is easy to verify that the adversary A contradicts our assumption that f_j is somewhat hard to break in $p(n)^{j+1}$ steps in the worst case. □

This completes the proof of the theorem.

Appendix B Missing Proofs for Theorem 12

Lemma 13. *For every distinguisher $D'_j \in \{D'_1, \ldots, D'_n\}$, if we run the simulator $S(x, z)$, then (with probability 1) S generates $D^{(1)}, \ldots, D^{(L)}$ and $S_1, \ldots, S_L$ such that*

$$\frac{1}{L}\sum_{i=1}^{L} \widehat{\mu}(S_i, D'_j) \geq \frac{1}{L}\sum_{i=1}^{L} \widehat{\mu}(S_i, \mathcal{D}^{(i)}) - O(\epsilon'(n)).$$

Proof. Recall that given two distributions X and Y, the Kullback-Leibler divergence (also called the relative entropy) of X and Y, denoted $KL(X||Y)$, is defined by

$$KL(X||Y) = \sum_{x \in Supp(X)} \Pr[X = x] \cdot \ln\left(\frac{\Pr[X = x]}{\Pr[Y = x]}\right).$$

Consider a distinguisher $D'_j \in \{D'_1, \ldots, D'_n\}$. Fix the random tape of the simulator S, and consider running $S(x, z)$ with the fixed random tape. Then, all the random variables (e.g., the $\mathcal{D}^{(i)}$'s) that appear in the simulator algorithm $S(x, z)$ become fixed. We first show that for every $i \in [L]$, we have

$$KL(D'_j||\mathcal{D}^{(i+1)}) - KL(D'_j||\mathcal{D}^{(i)}) \leq (\ln\frac{1}{\beta}) \cdot \widehat{\mu}(S_i, D'_j) - (1-\beta)\sum_{k=1}^{n} \Pr[\mathcal{D}^{(i)} = D'_k] \cdot \widehat{\mu}(S_i, D'_k). \quad (1)$$

Fix an $i \in [L]$. Then, we have

$$\begin{aligned}
KL(D'_j||\mathcal{D}^{(i+1)}) - KL(D'_j||\mathcal{D}^{(i)}) &= \ln \frac{1}{\Pr[\mathcal{D}^{(i+1)} = D'_j]} - \ln \frac{1}{\Pr[\mathcal{D}^{(i)} = D'_j]} \\
&= \ln \frac{\Pr[\mathcal{D}^{(i)} = D'_j]}{\Pr[\mathcal{D}^{(i+1)} = D'_j]} \\
&= \ln \frac{Z_i}{\beta^{\widehat{\mu}(S_i, D'_j)}}, \text{ where } Z_i := \sum_{k=1}^{n} \beta^{\widehat{u}(S_i, D'_k)} D^{(i)}(D'_k) \\
&= (\ln \frac{1}{\beta}) \cdot \widehat{\mu}(S_i, D'_j) + \ln \sum_{k=1}^{n} \beta^{\widehat{\mu}(S_i, D'_j)} \Pr[\mathcal{D}^{(i)} = D'_k] \\
&\leq (\ln \frac{1}{\beta}) \cdot \widehat{\mu}(S_i, D'_j) + \ln(1 - (1-\beta) \sum_{k=1}^{n} \Pr[\mathcal{D}^{(i)} = D'_k] \cdot \widehat{\mu}(S_i, D'_k)) \\
&\leq (\ln \frac{1}{\beta}) \cdot \widehat{\mu}(S_i, D'_j) - (1-\beta) \sum_{k=1}^{n} \Pr[\mathcal{D}^{(i)} = D'_k] \cdot \widehat{\mu}(S_i, D'_k),
\end{aligned}$$

where the first inequality follows from the fact that $\beta^x \leq 1 - (1-\beta)x$ for $\beta \geq 0$ and $x \in [0,1]$, and the second inequality follows from the fact that $\ln(1-x) \leq -x$ for $x < 1$. Thus, we have shown (1).

Now, summing inequality (1) over $i = 1, \ldots, L$, we have

$$KL(D'_j||\mathcal{D}^{(L+1)}) - KL(D'_j||\mathcal{D}^{(1)}) \leq (\ln \frac{1}{\beta}) \cdot \sum_{i=1}^{L} \widehat{\mu}(S_i, D'_j) - (1-\beta) \sum_{i=1}^{L} \sum_{k=1}^{n} \Pr[\mathcal{D}^{(i)} = D'_k] \cdot \widehat{\mu}(S_i, D'_k).$$

Now, using the inequalities $KL(D'_j||\mathcal{D}^{(L+1)}) \geq 0$, $KL(D'_j||\mathcal{D}^{(1)}) \leq \ln n$, and $\ln \frac{1}{\beta} \leq (1-\beta^2)/(2\beta)$ (which holds for every $\beta \in (0,1]$), we get

$$-\ln n \leq \frac{1-\beta^2}{2\beta} \sum_{i=1}^{L} \widehat{\mu}(S_i, D'_j) - (1-\beta) \sum_{i=1}^{L} \sum_{k=1}^{n} \Pr[\mathcal{D}^{(i)} = D'_k] \cdot \widehat{\mu}(S_i, D'_k).$$

Rearranging the inequality and using the fact that $\beta = \frac{1}{1+\sqrt{(2\ln n)/L}}$, we have

$$\begin{aligned}
\sum_{i=1}^{L} \sum_{k=1}^{n} \Pr[\mathcal{D}^{(i)} = D'_k] \cdot \widehat{\mu}(S_i, D'_k) &\leq \frac{1-\beta^2}{2\beta(1-\beta)} \sum_{i=1}^{L} \widehat{\mu}(S_i, D'_j) + \frac{1}{1-\beta} \ln n \\
&= \frac{1+\beta}{2\beta} \sum_{i=1}^{L} \widehat{\mu}(S_i, D'_j) + \frac{1}{1-\beta} \ln n \\
&= \sum_{i=1}^{L} \widehat{\mu}(S_i, D'_j) + (\frac{1+\beta}{2\beta} - 1) \sum_{i=1}^{L} \widehat{\mu}(S_i, D'_j) + \frac{\sqrt{2L\ln n}}{2} + \ln n \\
&\leq \sum_{i=1}^{L} \widehat{\mu}(S_i, D'_j) + \frac{1-\beta}{2\beta} \cdot L + \frac{\sqrt{2L\ln n}}{2} + \ln n \\
&= \sum_{i=1}^{L} \widehat{\mu}(S_i, D'_j) + \sqrt{2L\ln n} + \ln n.
\end{aligned}$$

Finally, dividing both sides by L and rearranging the inequality yields the result. □

Lemma 14. *For every $D' \in \{D'_1, \ldots, D'_n\}$, if we run the simulator $S(x,z)$, then with probability $1 - O(\epsilon'(n))$ over the random coins of S, $S(x,z)$ generates $\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(L)}$ and $S_1, \ldots, S_L$ such that*

$$\frac{1}{L}\sum_{i=1}^{L} \mu(S_i, D') \geq \frac{1}{L}\sum_{i=1}^{L} \mu(S_i, \mathcal{D}^{(i)}) - O(\epsilon'(n)).$$

Proof. We first show that for every $D'_j \in \{D'_1, \ldots, D'_n\}$ and every $i \in [L]$, with probability $1 - O(\frac{\epsilon'(n)}{nL})$ (over the random coins of S), we have

$$|\widehat{\mu}(S_i, D'_j) - \mu(S_i, D'_j)| \leq O(\epsilon'(n)). \tag{1}$$

Fix $D'_j \in \{D'_1, \ldots, D'_n\}$ and $i \in [L]$. Let $\widetilde{\mu}(S_i, D'_j)$ be defined by

$$\widetilde{\mu}(S_i, D'_j) = \Pr[D'_j(x, z, S_i(x,z)) = 1] - \Pr[D'_j(x, z, S_{D_U}(x, z||\vec{p}_{D'_j})) = 1].$$

One should view $\widetilde{\mu}$ as a "hybrid" between $\widehat{\mu}$ and μ. We now have the following equalities:

$$\begin{aligned}
\widehat{\mu}(S_i, D'_j) &= \mathrm{freq}_k[D'_j(x, z, S_i(x,z))] - \mathrm{freq}_k[D'_j(x, z, S_{D_U}(x, z||\vec{p}_{D'_j}))] \\
\widetilde{\mu}(S_i, D'_j) &= \Pr[D'_j(x, z, S_i(x,z)) = 1] - \Pr[D'_j(x, z, S_{D_U}(x, z||\vec{p}_{D'_j})) = 1] \\
\mu(S_i, D'_j) &= \Pr[D'_j(x, z, S_i(x,z)) = 1] - \Pr[D'_j(x, z, Out_{V^*}[P(x) \leftrightarrow V^*(x,z)]) = 1]
\end{aligned}$$

We note that $|\widehat{\mu}(S_i, D'_j) - \widetilde{\mu}(S_i, D'_j)| \leq O(\epsilon'(n))$ with probability $1 - O(\frac{\epsilon'(n)}{nL})$ by two applications of a Chernoff bound, the union bound, and the triangle inequality. Thus, to prove (1), it suffices to show that $|\widetilde{\mu}(S_i, D'_j) - \mu(S_i, D'_j)| \leq O(\epsilon'(n))$ with probability 1. Observe that

$$\begin{aligned}
&|\widetilde{\mu}(S_i, D'_j) - \mu(S_i, D'_j)| \\
&= |\Pr[D'_j(x, z, S_{D_U}(x, z||\vec{p}_{D'_j})) = 1] - \Pr[D'_j(x, z, Out_{V^*}[P(x) \leftrightarrow V^*(x,z)]) = 1]| \\
&= |\Pr[D_U(x, z||\vec{p}_{D'_j}, S_{D_U}(x, z||\vec{p}_{D'_j})) = 1] - \Pr[D_U(x, z||\vec{p}_{D'_j}, Out_{V^*}[P(x) \leftrightarrow V^*(x, z||\vec{p}_{D'_j})]) = 1]| \\
&\leq \epsilon'(n),
\end{aligned}$$

where the second equality follows from the definition of D_U and the fact that $V^*(x,z) = V^*(x, z||\vec{p}_{D'_j})$ (since $|z| \geq T_{V^*}(n)$), and the last inequality follows from the fact that S_{D_U} is a simulator for D_U in the weak (T_{D_U}, ϵ')-zero-knowledge property of (P, V), as required.

Now, by the union bound, with probability $1 - nL \cdot O(\frac{\epsilon'(n)}{nL}) = 1 - O(\epsilon'(n))$, we have

$$|\widehat{\mu}(S_i, D'_j) - \mu(S_i, D'_j)| \leq O(\epsilon'(n)) \tag{2}$$

for every $D'_j \in \{D'_1, \ldots, D'_n\}$ and every $i \in [L]$. Thus, for every $D' \in \{D'_1, \ldots, D'_n\}$, with probability $1 - O(\epsilon'(n))$, we have

$$\begin{aligned}
\frac{1}{L}\sum_{i=1}^{L} \mu(S_i, D') &\geq \frac{1}{L}\sum_{i=1}^{L} \widehat{\mu}(S_i, D') - O(\epsilon'(n)) \\
&\geq \frac{1}{L}\sum_{i=1}^{L}\sum_{j=1}^{n} \mathcal{D}^{(i)}(D'_j) \cdot \widehat{\mu}(S_i, D'_j) - O(\epsilon'(n)) \\
&\geq \frac{1}{L}\sum_{i=1}^{L}\sum_{j=1}^{n} \mathcal{D}^{(i)}(D'_j) \cdot \big(\mu(S_i, D'_j) - O(\epsilon'(n))\big) - O(\epsilon'(n)) \\
&= \frac{1}{L}\sum_{i=1}^{L} \mu(S_i, \mathcal{D}^{(i)}) - O(\epsilon'(n)),
\end{aligned}$$

where the first and third inequalities follow from (2), and the second inequality follows from Lemma 13. This completes the proof of the lemma. □

Appendix C Missing Proofs for Theorem 17

Lemma 19 (Approximating a distribution over circuits by a small circuit obtained via sampling)**.** *Let X and A be finite sets, let Y be any random variable with finite support, let $\mathcal{C}$ be any distribution over s-size randomized circuits of the form $C : X \times Supp(Y) \to A$, and let U be any finite set of randomized circuits of the form $u : X \times Supp(Y) \times A \to \{0,1\}$. Then, for every $\epsilon > 0$, there exists a randomized circuit $\widehat{C}$ of size $O(\frac{\log|X|+\log|U|}{\epsilon^2} \cdot s)$ such that for every $u \in U$ and $x \in X$, we have*

$$|\mathbb{E}_{C\leftarrow\mathcal{C}}[u(x,Y,C(x,Y))] - \mathbb{E}[u(x,Y,\widehat{C}(x,Y))]| \le \epsilon.$$

Proof. Fix $\epsilon > 0$. Let $C_1, \ldots, C_k \leftarrow \mathcal{C}$ be k circuits drawn independently from $\mathcal{C}$, where $k \ge 1$ will be specified later. By a Chernoff bound, for every $x \in X$ and $u \in U$, we have

$$\Pr_{C_1,\ldots,C_k\leftarrow\mathcal{C}}\left[\left|\mathbb{E}_{C\leftarrow\mathcal{C}}[u(x,Y,C(x,Y))] - \mathbb{E}_{i\leftarrow[k]}[u(x,Y,C_i(x,Y))]\right| > \epsilon\right] \le 2e^{-2k\epsilon^2}.$$

By a union bound over $x \in X$ and $u \in U$, we have

$$\begin{aligned}&\Pr_{C_1,\ldots,C_k\leftarrow\mathcal{C}}\left[\exists x \in X, u \in U : \left|\mathbb{E}_{C\leftarrow\mathcal{C}}[u(x,Y,C(x,Y))] - \mathbb{E}_{i\leftarrow[k]}[u(x,Y,C_i(x,Y))]\right| > \epsilon\right]\\ &\le |X| \cdot |U| \cdot 2e^{-2k\epsilon^2}.\end{aligned}$$

Now, we choose $k = O(\frac{\log|X|+\log|U|}{\epsilon^2})$ so that $|X| \cdot |U| \cdot 2e^{-2k\epsilon^2}$ in the above expression is strictly less than 1. Then, there exist $C_1, \ldots, C_k \in Supp(\mathcal{C})$ such that for every $x \in X$ and $u \in U$, we have

$$\left|\mathbb{E}_{C\leftarrow\mathcal{C}}[u(x,Y,C(x,Y))] - \mathbb{E}_{i\leftarrow[k]}[u(x,Y,C_i(x,Y))]\right| \le \epsilon.$$

Now, the lemma follows by choosing $\widehat{\mathcal{C}}$ to be a $O(\frac{\log|X|+\log|U|}{\epsilon^2} \cdot s)$-size circuit that chooses a circuit in $\{C_1, \ldots, C_k\}$ uniformly at random and then runs the chosen circuit on the input. □

Appendix D Separation of Super-Weak and Weak (t,ϵ)-Zero-Knowledge

In this section we separate the notion of super-weak and weak (t,ϵ)-zero-knowledge. First, let us formally define super-weak (t,ϵ)-zero-knowledge.

Definition 30 (super-weak (t,ϵ)-zero-knowledge). Let (P,V) be an interactive proof system for a language L. We say that (P,V) is *super-weak (t,ϵ)-zero-knowledge* if for every PPT adversary V^* and every t-time distinguisher D, there exists a PPT simulator S and an $n_0 \in \mathbb{N}$ such that for every $n \ge n_0$, $x \in L \cap \{0,1\}^n$, and $z \in \{0,1\}^*$, we have

$$\Pr[D(x,z,Out_{V^*}[P(x) \leftrightarrow V^*(x,z)]) = 1] - \Pr[D(x,z,S(x,z)) = 1] \le \epsilon(n).$$

Theorem 31. *There exists an interactive proof system (P,V) for an NP language L, such that (P,V) is super-weak (t,ϵ)-zero-knowledge for every polynomial t and inverse polynomial ϵ, but (P,V) is not weak $(t', \frac{1}{3})$-zero-knowledge for some polynomial t'.*

Proof. Let L be the trivial language $\{0,1\}^*$ with witness relation $R_L(x) = \{0,1\}$ for every $x \in \{0,1\}^*$, and let (P,V) be the interactive proof system where the prover P, on auxiliary input a bit y, sends the bit y to the verifier V, who simply outputs 1 (accepts). We first show that (P,V) is super-weak (t,ϵ)-zero-knowledge for every polynomial t and inverse polynomial ϵ. Let V^* be any PPT adversary, and let D be any t-time distinguisher. Let S be the PPT simulator that, on input (x,z), estimates $\Pr[D(x,z,Out_{V^*}[0 \leftrightarrow V^*(x,z)]) = 1]$ and $\Pr[D(x,z,Out_{V^*}[1 \leftrightarrow V^*(x,z)]) = 1]$ by running V^* and D sufficiently (polynomially) many times so that with probability $1 - \mathrm{negl}(n)$, the error is at most $\frac{1}{2\epsilon(n)}$, where $b \leftrightarrow V^*(x,z)$ denotes the protocol where the prover sends the bit b to V^*; then, S outputs $Out_{V^*}[b^* \leftrightarrow V^*(x,z)]$, where b^* is the bit b that had the higher estimated value for $\Pr[D(x,z,Out_{V^*}[b \leftrightarrow V^*(x,z)]) = 1]$. It is easy to see that with probability $1 - \mathrm{negl}(n)$, we have

$$\Pr[D(x,z,Out_{V^*}[P(x,y) \leftrightarrow V^*(x,z)]) = 1] - \Pr[D(x,z,S(x,z)) = 1] \le \frac{1}{\epsilon(n)}.$$

Thus, (P,V) is super-weak (t,ϵ)-zero-knowledge.

Let $t'(n) = O(n)$. We now show that (P,V) is not weak $(t', \frac{1}{3})$-zero-knowledge. Let V^* be the PPT adversary that simply outputs whatever the prover sends, and let D be the t'-time distinguisher that, on input $D(x,z,s)$, simply outputs the first bit of s. Now, we note that

$$\Pr[D(x,z,Out_{V^*}[P(x,0) \leftrightarrow V^*(x,z)]) = 1] = \Pr[D(x,z,0) = 1] = 0$$

and

$$\Pr[D(x,z,Out_{V^*}[P(x,1) \leftrightarrow V^*(x,z)]) = 1] = \Pr[D(x,z,1) = 1] = 1.$$

Since $\Pr[D(x,z,S(x,z)) = 1]$ cannot be simultaneously close to both 0 and 1, we see that (P,V) is not weak $(t', \frac{1}{3})$-zero-knowledge. □

The above theorem uses an NP language L with *non-unique* witnesses. However, under standard cryptographic assumptions, we can still prove the same result for an NP language L with *unique* witnesses.

Theorem 32. *Suppose there exists a one-way permutation $f : \{0,1\}^* \to \{0,1\}^*$ with a hard-core predicate $\phi : \{0,1\}^* \to \{0,1\}$. Then, there exists an interactive proof system (P,V) for an NP language L with unique witnesses, such that (P,V) is super-weak (t,ϵ)-zero-knowledge for every polynomial t and inverse polynomial ϵ, but (P,V) is not weak $(t', \frac{1}{3})$-zero-knowledge for some polynomial t'.*

Proof. Let L be the trivial language $\{0,1\}^*$ with unique witness relation $R_L(x) = f^{-1}(x)$. Let (P,V) be the interactive proof system where the prover P, on input (x,y), sends the bit $\phi(y)$ to the verifier V, who simply outputs 1 (accepts). We first show that (P,V) is super-weak (t,ϵ)-zero-knowledge for every polynomial t and inverse polynomial ϵ. Let V^* be any PPT adversary, and let D be any t-time distinguisher. Let S be the PPT simulator that, on input (x,z), first estimates $\Pr[D(x,z,Out_{V^*}[0 \leftrightarrow V^*(x,z)]) = 1]$ and $\Pr[D(x,z,Out_{V^*}[1 \leftrightarrow V^*(x,z)]) = 1]$ by running V^* and D sufficiently (polynomially) many times so that with probability $1 - \mathrm{negl}(n)$, the error is at most $\frac{1}{2\epsilon(n)}$, where $b \leftrightarrow V^*(x,z)$ denotes the protocol where the prover sends the bit b to V^*. Then, S outputs $Out_{V^*}[b^* \leftrightarrow V^*(x,z)]$, where b^* is the bit b that had the higher estimated value for $\Pr[D(x,z,Out_{V^*}[b \leftrightarrow V^*(x,z)]) = 1]$. It is easy to see that with probability $1 - \mathrm{negl}(n)$, we have

$$\Pr[D(x,z,Out_{V^*}[P(x) \leftrightarrow V^*(x,z)]) = 1] - \Pr[D(x,z,S(x,z)) = 1] \le \frac{1}{\epsilon(n)}.$$

Thus, (P, V) is super-weak (t, ϵ)-zero-knowledge.

Let $t'(n) = O(n)$. We now show that (P, V) is not weak $(t', \frac{1}{3})$-zero-knowledge. Let V^* be the PPT adversary that simply outputs whatever the prover sends, and let D be the distinguisher that, on input $D(x, z, s)$, simply outputs the first bit of s. Now, we note that

$$\Pr[D(x, z, Out_{V^*}[P(x, f^{-1}(x)) \leftrightarrow V^*(x, z)]) = 1] = \Pr[D(x, z, \phi(f^{-1}(x))) = 1] = \phi(f^{-1}(x)).$$

Now, suppose that (P, V) is weak $(t', \frac{1}{3})$-zero-knowledge. Then, there exists a PPT simulator S and an $n_0 \in \mathbb{N}$ such that for every $n \geq n_0$, $x \in L \cap \{0, 1\}^n$, and $z \in \{0, 1\}^*$, we have

$$|\Pr[D(x, z, Out_{V^*}[P(x, f^{-1}(x)) \leftrightarrow V^*(x, z)]) = 1] - \Pr[D(x, z, S(x, z)) = 1]| \leq \frac{1}{3},$$

which is equivalent to

$$|\phi(f^{-1}(x)) - \Pr[D(x, z, S(x, z)) = 1]| \leq \frac{1}{3}.$$

Now, using the simulator S and the distinguisher D, it is easy to construct an adversary A that computes the hard-core predicate with non-negligible probability. This contradicts the assumption that ϕ is a hard-core predicate for the one-way permutation f. □