Structure Versus Hardness Through the Obfuscation Lens

SIAM J. COMPUT.
Vol. 50, No. 1, pp. 98–144

STRUCTURE VERSUS HARDNESS THROUGH THE OBFUSCATION LENS*

NIR BITANSKY†, AKSHAY DEGWEKAR‡, AND VINOD VAIKUNTANATHAN§

Abstract. Much of modern cryptography, starting from public-key encryption and going beyond, is based on the hardness of structured (mostly algebraic) problems like factoring, discrete log, or finding short lattice vectors. While structure is perhaps what enables advanced applications, it also puts the hardness of these problems in question. In particular, this structure often puts them in low (and so-called structured) complexity classes such as $\mathsf{NP} \cap \mathsf{coNP}$ or statistical zero-knowledge (SZK). Is this structure really necessary? For some cryptographic primitives, such as one-way permutations and homomorphic encryption, we know that the answer is *yes*—they imply hard problems in $\mathsf{NP} \cap \mathsf{coNP}$ and SZK, respectively. In contrast, one-way functions do *not* imply such hard problems, at least not by *black-box reductions*. Yet, for many basic primitives such as public-key encryption, oblivious transfer, and functional encryption, we do not have any answer. We show that the above primitives, and many others, do *not* imply hard problems in $\mathsf{NP} \cap \mathsf{coNP}$ or SZK via black-box reductions. In fact, we first show that even the very powerful notion of indistinguishability obfuscation (IO) does not imply such hard problems, and then deduce the same for a large class of primitives that can be constructed from IO.

Key words. indistinguishability obfuscation, statistical zero-knowledge, $\mathsf{NP} \cap \mathsf{coNP}$, structured hardness, collision-resistant hashing

AMS subject classification. 94A60

DOI. 10.1137/17M1136559

1. Introduction.

In the last four decades, cryptography has produced a host of fantastic objects, from one-way functions (OWFs) and one-way permutations (OWPs) to public-key encryption [DH76, RSA78, GM82] and zero-knowledge proofs [GMR85] in the 1980s, all the way to fully homomorphic encryption [RAD78, Gen09, BV11] and indistinguishability obfuscation (IO) [BGI+01, GGH+13a] in the modern day.

The existence of all these objects requires at the very minimum that $\mathsf{NP} \not\subseteq \mathsf{BPP}$, but that is hardly ever enough. While OWFs, the most basic cryptographic object, do not seem to require much structure, as we advance up the ranks, we seem to require that certain *structured problems are hard*. For example, conjectured hard problems commonly used in cryptography (especially the public-key kind), such as factoring, discrete logarithms, and shortest (or closest) vectors in lattices, all have considerable

*Received by the editors June 28, 2017; accepted for publication (in revised form) September 17, 2020; published electronically January 13, 2021. An extended abstract of this paper appeared in *Proceedings of the IACR Advances in Cryptology Conference*, 2017.

https://doi.org/10.1137/17M1136559

Funding: This research was supported in part by NSF grants CNS-1350619 and CNS-1414119, an Alfred P. Sloan Research Fellowship, a Microsoft Faculty Fellowship, the NEC Corporation, and the Steven and Renee Finn Career Development Chair from MIT. This work was also sponsored in part by the Defense Advanced Research Projects Agency (DARPA) and the U.S. Army Research Office under contract W911NF-15-C-0226.

†Tel Aviv University, Tel Aviv 69978, Israel (nirbitan@csail.mit.edu). Member of the Check Point Institute of Information Security. Supported by the Alon Young Faculty Fellowship, by Len Blavatnik and the Blavatnik Family foundation,and an ISF grant 18/484. Part of this research was done while at MIT.

‡Two Sigma Investments, LP, New York, NY 10013 USA (degwekarakshay@gmail.com). The views expressed herein are solely the views of the author(s) and are not necessarily the views of Two Sigma Investments, LP or any of its affiliates. They are not intended to provide, and should not be relied upon for, investment advice.

§MIT, Cambridge, MA 02142 USA (vinodv@csail.mit.edu).

algebraic structure. On one hand, it is this structure that enables strong applications such as public-key and homomorphic encryption. On the other hand, this structure is also what puts their hardness in question and is exactly what algorithms may try to exploit in order to solve these problems. There is of course the fear that this structure will (eventually, if not today) deem these problems *easy*. Or, as Barak says more eloquently [Bar13],

> *based on the currently well studied schemes, structure is strongly associated with (and perhaps even implied by) public key cryptography. This is troubling news, since it makes public key crypto somewhat of an "endangered species" that could be wiped out by a surprising algorithmic advance. Therefore the question of whether structure is inherently necessary for public key crypto is not only of mathematical interest but also of practical importance as well.*

Thus, a fundamental question in cryptography is *what type of structure is necessary for different primitives?* Indeed, the answer to this question may be crucial to our understanding of what are the minimal assumptions required to construct these primitives. While there may be different ways of approaching this question, one main approach, which is also taken in this work, has been through the eyes of complexity theory. That is, we wish to understand which cryptographic primitives require hardness in low (and so-called structured) complexity classes such as $\mathsf{NP} \cap \mathsf{coNP}$, TFNP (the class of total NP search problems), or SZK (the class of problems with statistical zero-knowledge proofs).

Aiming to answer this question, one line of research demonstrates that, for some cryptographic primitives, hardness in structured complexity classes is indeed necessary. The existence of OWPs requires a hard problem in $\mathsf{NP} \cap \mathsf{coNP}$ [Bra79]; the same holds for restricted cases of public-key encryption schemes satisfying specific structural properties (e.g., ciphertext certification) [Bra79, GG98]; homomorphic encryption schemes and noninteractive computational private information retrieval (PIR) schemes imply hard problems in SZK [BL13, LV16]; and IO schemes imply a hard problem in $\mathsf{PPAD} \subseteq \mathsf{TFNP}$ (assuming $\mathsf{NP} \not\subseteq \mathsf{ioBPP}$) [BPR15].

Yet, for many primitives such hardness is not known to be inherent. While this is perhaps expected for OWFs, it is also the case for seemingly structured primitives such as collision-resistant hash functions, oblivious transfer, and general public-key encryption schemes. *Do these primitives require hardness in structured complexity classes? Can we prove that they do or that they don't?*

Black-box separations. Formalizing this question in a meaningful way requires care. Indeed, it may be easy to formalize a statement of the form "the existence of crypto primitive $\mathcal{P}$ *implies* hardness in a complexity class $\mathcal{C}$": one has to show that the ability to solve every problem in $\mathcal{C}$ implies breaking any instantiation of primitive $\mathcal{P}$. However, it is not clear how to prove statements of the form "the existence of crypto primitive $\mathcal{P}$ does *not* imply hardness in a complexity class $\mathcal{C}$." For example, it is commonly believed that $\mathsf{NP} \cap \mathsf{coNP}$ *does* contain hard problems. So in a trivial logical sense the existence of such problems is implied by any primitive $\mathcal{P}$. Instead, we follow the methodology of black-box separations, whose study in cryptography was pioneered in a remarkable work by Impagliazzo and Rudich [IR89]. Faced with a similar problem of how to show that a primitive $\mathcal{P}$ (OWFs) cannot be used to construct another primitive $\mathcal{P}'$ (public-key encryption), they prove this cannot be shown through *black-box reductions*—cryptography's de facto technique for showing such implications.

A bit more elaborately, a *fully black-box reduction* [RTV04] of a primitive (or, in our case, a problem) $\mathcal{P}'$ to a primitive $\mathcal{P}$ consists of a black-box *construction* and a black-box *security reduction*. The construction of $\mathcal{P}'$ from $\mathcal{P}$ does not exploit the ac-

tual implementation of primitive $\mathcal{P}$, but rather just its input-output interface. The security reduction can use any adversary that breaks (or, in our case, solves) $\mathcal{P}'$ to break $\mathcal{P}$ and is oblivious to the implementation of the adversary (as well as of that of $\mathcal{P}$).

Following [IR89], there has been a rich study of black-box separations in cryptography (see, e.g., [Rud91, Sim98, KST99, GKM+00, GT00, GMR01, BT03, RTV04, HR04, GGKT05, Pas06, GMM07, BM09, HH09, KSS11, BKSY11, DLMM11, GKLM12, DHT12, BBF13, Fis12, Pas13, BB15, HHRS15] and many others). Most of this study has been devoted to establishing separations between different cryptographic primitives. (In particular, the most relevant to us are the recent works of Asharov and Segev [AS15, AS16] that study black-box separations for IO, which we elaborate on below.) Some of this study puts limitations on basing cryptographic primitives on NP-hardness [GG98, AGGM06, BL13, BB15, LV16].

Going back to our main question of which primitives require structured hardness, we know the following:

- As described above, OWPs imply a hard problem in $\mathsf{NP} \cap \mathsf{coNP}$ [Bra79], homomorphic encryption and PIR imply hard problems in SZK [BL13, LV16] and IO (with OWFs) implies a hard problem in PPAD [BPR15] via *black-box reductions*.
- On the flip side, we know that there are no black-box reductions from hard problems in $\mathsf{NP} \cap \mathsf{coNP}$ to OWFs [BI87, Rud88], and from hard-on-average problems in SZK to OWPs (corollary from [Ost91, OV08, HHRS15, BHKY19]).[1]

For more advanced primitives, most notably (general) public-key encryption, we do not have results in either direction. In fact, many existing constructions are based on problems in $\mathsf{NP} \cap \mathsf{coNP}$ or SZK. We are thus left with (quite basic) primitives at an unclear state; as far as we know, they may very well imply hard problems in structured complexity classes, even by black-box reductions.

1.1. Our results. We revisit the relationship between two structured complexity classes, SZK and $\mathsf{NP} \cap \mathsf{coNP}$, and cryptographic primitives. In broad strokes, we show that there are no fully black-box reductions of hard problems in these classes to any one of a variety of cryptographic primitives, including (general) public-key encryption, oblivious transfer, deniable encryption, and functional encryption. More generally, we separate SZK and $\mathsf{NP} \cap \mathsf{coNP}$ from IO. Then, leveraging on the fact that IO can be used to construct a wide variety of cryptographic primitives in a black-box way, we derive corresponding separations for these primitives.[2] One complexity-theoretic corollary of this result is a separation between SZK and $\mathsf{NP} \cap \mathsf{coNP}$ from the class PPAD [MP91] that captures the complexity of computing Nash equilibria.

We now go into more detail on each of the results.

Statistical zero-knowledge and cryptography. The notion of SZK proofs was introduced in the seminal work of Goldwasser, Micali, and Rackoff [GMR85]. The class of *promise problems* with SZK proofs can be characterized by several complete problems, such as *statistical difference* [SV03] and *entropy difference* [GV99]. SZK hardness is known to follow from various number-theoretic problems that are commonly used in

[1] Specifically, there exists a fully black-box reduction of constant-round statistically hiding commitments to average-case hardness in SZK [Ost91, OV08, BHKY19], whereas a fully black-box reduction of the latter primitive to OWPs is ruled out in [HHRS15]. Together, these rule out a fully black-box reduction of average-case SZK hardness to OWPs.

[2] More accurately, these primitives follow from IO and OWFs, and accordingly our separation addresses IO and OWFs in conjunction. The concept of a black-box reduction from IO and OWF requires clarification and discussion. Here we will follow the framework of Asharov and Segev [AS15]. We elaborate below.

cryptography, such as discrete logarithms [GK93], quadratic residuosity [GMR85], and lattice problems [GG98, MV03], as well as problems like graph isomorphism [GMW91]. As mentioned, we also know that a handful of cryptographic primitives such as homomorphic encryption [BL13], PIR [LV16], and rerandomizable encryption imply hardness in SZK. (On the other hand, $\mathsf{SZK} \subseteq \mathsf{AM} \cap \mathsf{coAM}$ [For89, AH91], and thus, SZK cannot contain NP-hard problems, unless the polynomial hierarchy collapses [BHZ87].)

We ask more generally which cryptographic primitives can be shown to imply such hardness, with the intuition that such primitives are *structured* in a certain way. In particular, whereas one may not expect a seemingly unstructured object like OWFs to imply such hardness, what can we say, for instance, about OWPs, public-key encryption, or even IO (which has proven to be powerful enough to yield almost any known cryptographic goal)?

We prove that none of these primitives imply such hardness through black-box reductions.

THEOREM 1.1 (informal). *There is no fully black-box reduction of any (even worst-case) hard problem in* SZK *to IO and OWPs.*

COROLLARY 1.2 (from [SW14, Wat15], informal). *There is no such reduction to (general) public-key encryption, oblivious transfer, deniable encryption, functional encryption, or any other object that has a black-box reduction to IO and OWPs.*

We would like to elaborate a bit more on what a black-box construction of a hard problem in SZK means. We shall focus on the characterization of SZK by the *statistical difference* promise problem [SV03]. In this problem, an instance is a pair of circuit samplers $C_0, C_1 : \{0,1\}^n \to \{0,1\}^m$ which induce distributions $\boldsymbol{C}_0$ and $\boldsymbol{C}_1$ where the distribution $\boldsymbol{C}_b$ is obtained by evaluating the circuit C_b on a uniformly random input. The promise is that the statistical distance $s = \Delta(\boldsymbol{C}_0, \boldsymbol{C}_1)$ of the corresponding distributions is either large (say, $s \geq 2/3$) or small (say, $s \leq 1/3$). The problem, named $\mathbf{SD}^{1/3,2/3}$ (or just $\mathbf{SD}$), is to decide which is the case.

Let us look at a specific example of the construction of such a problem from *rerandomizable encryption*. In a (say, symmetric-key) rerandomizable encryption scheme, on top of the usual encryption and decryption algorithms $(\mathsf{Enc}, \mathsf{Dec})$ there is a ciphertext rerandomization algorithm ReRand that can statistically refresh ciphertexts. Namely, for any ciphertext CT encrypting a bit b, $\mathsf{ReRand}(\mathsf{CT})$ produces a ciphertext that is statistically close to a fresh encryption $\mathsf{Enc}_{\mathsf{sk}}(b)$. This immediately gives rise to a hard statistical difference problem [BL13]: given a pair of ciphertexts $(\mathsf{CT}_0, \mathsf{CT}_1)$, decide whether the corresponding rerandomized distributions given by the circuits $(C_0(\cdot), C_1(\cdot)) := (\mathsf{ReRand}(\mathsf{CT}_0; \cdot), \mathsf{ReRand}(\mathsf{CT}_1; \cdot))$ are statistically far or close. Indeed, this corresponds to whether they encrypt the same bit or not, which is hard to decide by the security of the encryption scheme.

A feature of this reduction of hard statistical difference instances to rerandomizable encryption is that, similarly to most reductions in cryptography, it is *fully black-box* [RTV04] in the sense that the circuits C_0, C_1 only make black-box use of the encryption scheme's algorithms, and can in fact be represented as oracle-aided circuits $(C_0^{\mathsf{ReRand}(\cdot)}, C_1^{\mathsf{ReRand}(\cdot)})$. Furthermore, "hardness" can be shown by a black-box security proof that can use any decider for the problem in a black-box way to break the underlying encryption scheme. More generally, one can consider the statistical difference problem relative to different oracles implementing different cryptographic primitives and ask when hardness can be shown based on a black-box reduction. Theorem 1.1 rules out such reductions relative to IO and OWPs (and everything that follows from these in a fully black-box way). For more details, see subsection 1.1 and section 3.

$\mathsf{NP}\cap\mathsf{coNP}$ *and cryptography.* Hard (on average) problems in $\mathsf{NP}\cap\mathsf{coNP}$ are known to follow based on several number-theoretic problems in cryptography, such as discrete log, factoring, and lattice problems [Has88, LLJS90, AR04]. As in the previous section for SZK, we are interested in understanding which cryptographic primitives would imply such hardness, again with the intuition that this implies structure. For instance, it is known [Bra79] that any OWP $f : \{0,1\}^n \to \{0,1\}^n$ implies a hard problem in $\mathsf{NP}\cap\mathsf{coNP}$, e.g., given an index $i \in [n]$ and an image $f(x)$ find the ith preimage bit x_i. In contrast, Blum and Impagliazzo [BI87] and Rudich [Rud88] proved that seemingly unstructured objects like OWFs do not imply hardness in $\mathsf{NP} \cap \mathsf{coNP}$ by fully black-box reductions. In this context, a fully black-box reduction essentially means that the nondeterministic verifiers only make black-box use of the OWF (or OWP in the previous example) and the reduction establishing the hardness is also black-box (in both the decider and the OWF).[3]

But what about more structured primitives such as public-key encryption, oblivious transfer, or even IO? We rule out fully black-box reductions from OWFs (or even *injective OWFs* (IOWFs)) and IO to hard problems in $\mathsf{NP} \cap \mathsf{coNP}$, hence, also for the other primitives, which can be constructed from IO (with OWFs) in a fully black-box way.

THEOREM 1.3 (informal). *There is no fully black-box reduction of any (even worst-case) hard problem in* $\mathsf{NP} \cap \mathsf{coNP}$ *to IO and IOWFs.*

COROLLARY 1.4 (from [SW14, Wat15], informal). *There is no such reduction to (general) public-key encryption, oblivious transfer, deniable encryption, functional encryption, or any other object that has a black-box reduction to IO and OWFs.*

Our approach also gives a new (rather different) proof to the original separation between OWFs and $\mathsf{NP} \cap \mathsf{coNP}$ [BI87, Rud88]. For more details, see subsection 1.1 and section 4.

We remark that unlike our result for SZK (which ruled out hard *promise problems*), the above result only rules out hard *languages* in $\mathsf{NP} \cap \mathsf{coNP}$. Indeed, Even, Selman, and Yacobi [ESY84] demonstrated promise problems in $\mathsf{NP} \cap \mathsf{coNP}$ that are NP-hard. Hence even the assumption $\mathsf{P} \neq \mathsf{NP}$ (let alone OWFs) gives us hard promise problems in $\mathsf{NP} \cap \mathsf{coNP}$. (See [Gol06] for further reading.)

Relation to the work of Asharov and Segev. The flood of IO applications starting from [GGH$^+$13b, SW14] has lead many to conjecture that IO may be "complete for cryptography" (assuming also OWFs, or just $\mathsf{NP} \not\subseteq \mathsf{ioBBP}$ [KMN$^+$14]). Nevertheless, some cryptographic goals could not be constructed based on IO.

Asharov and Segev [AS15, AS16] were the first to initiate a formal study to understand *the limits of IO*. Our separations for IO are based on their framework [AS15]. We aim to draw the complexity-theoretic boundaries of IO. Indeed, black-box separations from IO require some care, given that the typical use of IO makes non-black-box use of the circuits it obfuscates and thus any associated cryptographic primitive such as OWFs. The Asharov–Segev framework considers obfuscators that take as input circuits with OWF (or OWP) gates. They observe, most known IO-based constructions fall into this category. Thus, a separation in this model allows deriving the corresponding separations between SZK or $\mathsf{NP} \cap \mathsf{coNP}$ and a wide variety of cryptographic primitives. See subsection 1.1 for more details.

[3] Roughly speaking, [BI87] rule out *perfectly correct constructions*, where the $\mathsf{NP}\cap\mathsf{coNP}$ structure is guaranteed for any implementation of the OWF oracle. In [Rud88], this is generalized also to *almost perfectly correct constructions* that only work for an overwhelming fraction of OWF oracles. We also rule out constructions that are perfectly correct.

In terms of results, they show that collision-resistant hashing and (domain invariant) OWPs do not have black-box reductions to IO (and OWFs). Our separation of IO and $\mathsf{NP} \cap \mathsf{coNP}$ is more general and implies their previous result for OWPs (and gives a rather different proof for this fact). Their result for collision-resistant hashing is not captured by our results (indeed collision-resistance is not known to imply hardness in either SZK or $\mathsf{NP} \cap \mathsf{coNP}$). We also stress that our separation of SZK from IO and OWPs does not follow from their results; indeed, SZK-hardness is not known to imply collision-resistance.

Indistinguishability obfuscation: Perspective. Since the breakthrough of [GGH$^+$13b], the notion of IO has been extensively studied. While we already understand that IO has far-reaching implications, our understanding of how it can be constructed and under what assumptions is still at an early stage. Indeed, basing IO on solid foundations is one of cryptography's greatest challenges today. In this context, we stress that the results presented in this work hold regardless of the state of existing candidates. In fact, even if it turned out that there is no secure realization of IO, the separation of SZK and $\mathsf{NP} \cap \mathsf{coNP}$ from primitives such as public-key encryption, which follow from IO, still holds. The expressiveness of IO (established in [GGH$^+$13b, SW14] and onward) allows us to prove many separations in one shot. (Indeed, three years ago we would have probably addressed each primitive separately.)

As for the search for candidates itself, while at this point candidates are based on lattice-related problems that do break in SZK, our work suggests the theoretical possibility that IO candidates may not require such structure. A similar conclusion is true of course for the much more basic and long-studied question of public-key encryption. Almost all known public-key encryption candidates rely on very algebraic assumptions (that do break in SZK or $\mathsf{NP} \cap \mathsf{coNP}$). Constructing public key encryption from less structured assumptions remains a fascinating open question. While there have been initial steps trying to diverge from such structure [Ale03, ABW10], there is yet a long way to go.

On TFNP versus $\mathsf{NP} \cap \mathsf{coNP}$. One of the corollaries of our result is a separation between SZK and $\mathsf{NP} \cap \mathsf{coNP}$ from the complexity class PPAD. PPAD, a subclass of total NP search problems called TFNP [MP91], was defined by Papadimitriou [Pap94] and has been shown to capture the complexity of computing Nash equilibria [DGP06, CDT09]. It was recently shown [BPR15] that IO and IOWFs can be used (in a black-box way) to construct hard problems in PPAD. Put together with our separation, we get that there is no black-box construction of an SZK (resp., $\mathsf{NP} \cap \mathsf{coNP}$) hard problem from PPAD-hardness.[4]

Given that TFNP, which contains PPAD, is commonly thought of as a search version of $\mathsf{NP} \cap \mathsf{coNP}$, it is interesting to note that the result shows that hardness in $\mathsf{NP} \cap \mathsf{coNP}$ (of decisional problems) does not follow from hardness in TFNP (a.k.a., hardness of search problems) in a black-box way. Namely, there is no black-box "search-to-decision reduction" between these classes.

Subsequent work. Following the publication of the conference version of this work, Komargodski and Yogev [KY18] showed that Simon's oracle [Sim98] can be used to decide SZK. Combining their work with that of Asharov and Segev [AS15] gives another, quite different, proof separating average-case hardness in SZK from IO and OWPs.

[4] We note that in concurrent and independent work, Rosen, Shahaf, and Segev [RSS16] show that OWFs do not have black-box reductions to PPAD-hardness, which combined with [Ost91] also yields a separation between SZK and PPAD.

Bitansky and Degwekar [BD19], based on the coupling-based approach presented in this work, showed that collision-resistant hash functions do not imply hardness in SZK, and also gave a new proof of the separation of IO and collision-resistant hash functions previously shown by Asharov and Segev [AS15].

Organization. We give an overview of the methodology and techniques used in subsection 1.1. Section 2 provides required preliminaries. The black-box separation between SZK and IO (plus OWPs) is given in section 3. The separation between $\mathsf{NP} \cap \mathsf{coNP}$ and IO (plus IOWFs) is given in section 4.

1.2. Overview of techniques. We now give an overview of our approach and main ideas. We start by discussing how to capture fully black-box constructions in the context of IO following [AS15]. We then recall the common methodology for ruling out black-box constructions [IR89, RTV04, BBF13] and explain the main ideas behind our impossibility results for SZK and $\mathsf{NP} \cap \mathsf{coNP}$.

Indistinguishability obfuscation and black-box constructions. Traditionally, when thinking about a *black-box construction* of one cryptographic primitive $\mathcal{P}'$ (e.g., a pseudorandom generator) from a primitive $\mathcal{P}$ (e.g., an OWF), we mean that all algorithms in the construction of $\mathcal{P}'$ invoke $\mathcal{P}$ as a black-box, oblivious of its actual implementation. This is hardly the case in constructions based on IO where circuits that explicitly invoke the primitive $\mathcal{P}$ may be obfuscated.

Nonetheless, as observed by Asharov and Segev [AS15], in almost all existing constructions, the code implementing $\mathcal{P}$ is used in a very restricted manner. Typically, obfuscated circuits can be implemented as oracle-aided circuits $C^{\mathcal{P}}$ that are completely black-box in $\mathcal{P}$, where $\mathcal{P}$ is some low-level primitive, such as an OWF. Indeed, in most cases the circuits obfuscated are symmetric-key primitives, such as puncturable pseudorandom functions [SW14], which can be constructed in a black-box way from OWFs (in some constructions more structured low-level primitives may be used, like IOWFs, or injective OWPs). Furthermore, in these constructions, the obfuscator $i\mathcal{O}$ itself is also treated as a black-box.

Accordingly, almost all existing constructions based on IO can be cast into a model in which IO exists for oracle-aided circuits $C^{\mathcal{P}}$, where $\mathcal{P}$ is, say, an OWF, and both $\mathcal{P}$ and the obfuscator $i\mathcal{O}$ can only be accessed as black-boxes. On top of that, they can be proven secure in this model by a *black-box reduction* that makes black-box use of $(\mathcal{P}, i\mathcal{O})$ and any attacker against the constructed primitive $\mathcal{P}'$. Such constructions where both the construction itself and the reduction are black-box are called *fully black-box constructions* [RTV04]. Following Asharov and Segev [AS15, AS16], we shall prove our results in this model, ruling out black-box constructions of hard problems in SZK and $\mathsf{NP} \cap \mathsf{coNP}$ based on IO for oracle-aided circuits. This approach traces back to the work of Brakerski et al. [BKSY11] in the context of zero-knowledge proofs (rather than IO) and was further extended by Garg, Mahmoody, and Mohammed [GMM17]. Further details follow.

Ruling out black-box reductions. We prove our results in the model described above following the methodology of oracle separations (see, e.g., [IR89, Sim98, RTV04, HR04]). Concretely, to prove that there is no fully black-box construction of a primitive $\mathcal{P}'$ from primitive $\mathcal{P}$, we demonstrate oracles $(\Psi, \mathcal{A})$ such that

- relative to Ψ, there exists a construction $C^{\Psi}_{\mathcal{P}}$ realizing $\mathcal{P}$ that is secure in the presence of $\mathcal{A}$,
- but *any* construction $C^{\Psi}_{\mathcal{P}'}$ realizing $\mathcal{P}'$ can be broken using $\mathcal{A}$.

Indeed, if such oracles $(\Psi, \mathcal{A})$ exist, then no efficient reduction will be able to use (as a black-box) the attacker $\mathcal{A}$ against $\mathcal{P}'$ to break $\mathcal{P}$ (as the construction of $\mathcal{P}$ is secure

in the presence of $\mathcal{A}$). In our case, we would like to apply this paradigm to rule out black-box constructions of hard instances in either SZK or $\mathsf{NP} \cap \mathsf{coNP}$ from IO for oracle-aided circuits and a low-level primitive (e.g., an OWF). We next outline the main ideas behind the construction and analysis of the oracles $(\Psi, \mathcal{A})$ in each of the two cases.

Ruling out black-box constructions of hard SZK *problems.* As explained in the previous section, we focus on the characterization of SZK by its complete problem: the statistical difference problem $\mathbf{SD}$ [SV03]. We demonstrate oracles $(\Psi, \mathcal{A})$ such that relative to Ψ there exist constructions of OWPs and IO for circuits with OWP gates, and these constructions are secure in the presence of $\mathcal{A}$. At the same time, $\mathcal{A}$ will decide (in the worst-case) $\mathbf{SD}^{\Psi}$. Since $\mathbf{SD}$ is complete for SZK in a relativizing manner, deciding $\mathbf{SD}^{\Psi}$ suffices to break SZK^{Ψ}. That is, $\mathcal{A}$ will decide *all* instances (C_0^{Ψ}, C_1^{Ψ}) of circuit samplers that only use the IO and OWPs realized by Ψ in a black-box manner. We next explain how each of the two are constructed.

The construction of Ψ follows a general recipe suggested in [AS15, AS16]. The oracle consists of three parts $(f, \mathcal{O}, \mathsf{Eval}^{f,\mathcal{O}})$, where

1. f is a random permutation, realizing the OWP primitive;
2. $\mathcal{O}$ is a random injective function, realizing the obfuscation algorithm, and it takes as input an oracle-aided circuit $C^{(\cdot)}$ along with randomness r and outputs an obfuscation $\widehat{C} = \mathcal{O}(C, r)$;
3. $\mathsf{Eval}^{\mathcal{O},f}$ realizes evaluation of obfuscated circuits. On input $(\widehat{C}, x)$, it inverts $\mathcal{O}$ to find (C, r) and outputs $C^f(x)$. If $\widehat{C}$ is not in the image of $\mathcal{O}$, it returns $\perp$.

The above construction readily satisfies the syntactic (or "functionality") requirements of OWPs and IO. Furthermore, using standard techniques, it is not hard to show that relative to Ψ, the function f is one-way and $\mathcal{O}$ satisfies the IO indistinguishability requirement. The challenge is to now come up with an oracle $\mathcal{A}$ that, on one hand, will decide $\mathbf{SD}^{\Psi}$ but, on the other, will not compromise the security of the latter primitives.

Recall that deciding $\mathbf{SD}^{\Psi}$ means that given two oracle-aided circuit samplers (C_0, C_1) such that the statistical distance of the corresponding distributions $(\boldsymbol{C}_0^{\Psi}, \boldsymbol{C}_1^{\Psi})$ is $s = \Delta(\boldsymbol{C}_0^{\Psi}, \boldsymbol{C}_1^{\Psi}) \in [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$, the oracle $\mathcal{A}$ must decide in which of the two intervals s lies, whereas if the promise is not satisfied and $s \in (\frac{1}{3}, \frac{2}{3})$, there is no requirement whatsoever. With this in mind, a first naive attempt would be the following. $\mathcal{A}$ having unbounded access to Ψ, and given a query (C_0, C_1) as input, computes the statistical distance $s = \Delta(\boldsymbol{C}_0, \boldsymbol{C}_1)$ and then output whether $s < \frac{1}{2}$ or $s \geq \frac{1}{2}$. While such an oracle would definitely decide $\mathbf{SD}^{\Psi}$, it is not too hard to show that it is simply too powerful and would not only break IO and OWPs but would, in fact, allow solving any problem in NP^{Ψ} (or even in PP^{Ψ}). Other naive attempts, such as refusing to answer outside the promise intervals, encounter a similar problem.

At a high level, the problem with such oracles is that solutions to hard problems can be easily correlated with "tiny" differences in the statistical distance of the two input circuits, whereas the above oracle may reflect tiny changes when the statistical distance is close to some threshold (1/2 in the above example) on which the oracle changes its behavior. This motivates our actual definition of $\mathcal{A}$ as a *noisy oracle* that produces its answer, not according to some fixed threshold, but according to a random threshold, chosen afresh for each and every query. Concretely, the oracle, which we call StaDif^{Ψ}, for any query (C_0, C_1), chooses a uniformly random threshold $t \leftarrow (\frac{1}{3}, \frac{2}{3})$ and answers accordingly:

$$\mathsf{StaDif}^{\Psi}(C_0, C_1) = \begin{cases} Y & \text{if } s \geq t \text{ (far distributions)}, \\ N & \text{if } s < t \text{ (similar distributions)}. \end{cases}$$

The main challenge in proving that the security of the IO and OWPs realized by $\mathcal{A}$ is not compromised by this oracle is that StaDif^{Ψ} has the power to query Ψ on exponentially many points in order to compute s. For instance, it may query Ψ on the preimage of an OWP challenge $f(x)$ or of a given obfuscation $\mathcal{O}(C, r)$. The key observation behind the proof is that the oracle's final answer still does not reflect how Ψ behaves locally on random points.

Intuitively, choosing the threshold t at random, for each query (C_0, C_1), guarantees that with high probability t is "far" from the corresponding statistical distance $s = \Delta(C_0^{\Psi}, C_1^{\Psi})$. Thus, changing the oracle Ψ on, say, a single input x, such as the preimage of an OWP challenge $f(x)$, should not significantly change s and will not affect the oracle's answer, that is, unless the circuits query Ψ on x with high probability to begin with. We give a reduction showing that we can always assume that (C_0, C_1) are "smooth," in the sense that they do not make any specific query to Ψ with too high probability.

Following this intuition, we are able to show that through such local changes that go undetected by StaDif^{Ψ}, we can move to an ideal world where inverting the OWP or breaking IO can be easily shown to be impossible. We refer the reader to section 3 for further details.

Ruling out black-box constructions of hard $\mathsf{NP} \cap \mathsf{coNP}$ *problems.* As mentioned earlier, a fully black-box construction of hard problems in $\mathsf{NP} \cap \mathsf{coNP}$ is actually known assuming OWPs and cannot be ruled out as in the case of SZK. Instead, we rule out constructions from (nonsurjective) IOWFs and IO for circuits with IOWF gates. This generalizes several previous results by Blum and Impagliazzo [BI87] and Rudich [Rud88], showing that OWFs do not give hardness in $\mathsf{NP} \cap \mathsf{coNP}$, by Matsuda and Matsuura [MM11], showing that IOWFs do not give OWPs (which are a special case of hardness $\mathsf{NP} \cap \mathsf{coNP}$), and by Asharov and Segev [AS16], showing that OWFs and IO for circuits with OWF gates do not give OWPs. In fact, our approach yields a new (and rather different) proof for each one of these results.

We follow a methodology similar to the one we used for the case of SZK. That is, we would like to come up with oracles $(\Psi, \mathcal{A})$ such that Ψ realizes IOWFs and IO for circuits with IOWF gates, which are both secure in the presence of $\mathcal{A}$, whereas black-box constructions of problems in $\mathsf{NP} \cap \mathsf{coNP}$ from these primitives can be easily solved by $\mathcal{A}$. By black-box constructions here we mean a pair of efficient oracle-aided nondeterministic verifiers $V_0^{(\cdot)}, V_1^{(\cdot)}$ that for every oracle Ψ implementing IOWFs and IO yield co-languages $\overline{L}^{\Psi}, L^{\Psi}$ in $\mathsf{NP}^{\Psi} \cap \mathsf{coNP}^{\Psi}$.

The requirement that V_0, V_1 give a language in $\mathsf{NP} \cap \mathsf{coNP}$ for *every* oracle implementing IOWFs and IO follows previous modeling [BI87][5] and aligns with how we usually think about *correctness* of black-box constructions of cryptographic primitives. For instance, the construction of public-key encryption from trapdoor permutations is promised to be correct, for all oracles implementing the trapdoor permutation. Similarly, the construction of hard $\mathsf{NP} \cap \mathsf{coNP}$ languages from OWPs gives an $\mathsf{NP} \cap \mathsf{coNP}$ language for any oracle implementing a permutation.[6]

[5] Rudich [Rud88] also considered a slight relaxation of constructions that are correct for an overwhelming fraction of oracles rather than all.

[6] We note that this issue does not come up for black-box constructions of SZK *promise* problems, because the construction is allowed to yield instances that do not obey the promise; there correctness is always guaranteed, and the only question is whether the instances that do satisfy the promise are hard to decide.

We stress that a construction where correctness is only guaranteed for particular (even if natural) oracles may definitely exist. This is, for example, the case if we only consider implementations of IO similar to those presented above in the context of SZK. Indeed, in that construction the implementation of IO has an additional property—it allows identifying *invalid obfuscations* (the Eval oracle would simply return $\perp$ on such obfuscations). This "verifiability" property coupled with the injectivity of obfuscators actually implies a hard problem in $\mathsf{NP} \cap \mathsf{coNP}$ in a black-box way.[7] Our separation thus leverages the fact that IO need not necessarily be verifiable and rules out constructions that are required to be correct for any implementation of IO, even a nonverifiable one.

Accordingly, the oracles $\Psi = (f, \mathcal{O}, \mathsf{Eval}^{f,\mathcal{O}})$ that we consider are a tweaked version of the oracles considered in the SZK case. Now f is a random injective function that is expanding, rather than a permutation, the oracle $\mathcal{O}$ is defined as before, and the oracle $\mathsf{Eval}^{f,\mathcal{O}}$ is defined as before for valid obfuscations $\widehat{C} \in \mathrm{Image}(\mathcal{O})$ *but is allowed to act arbitrarily for invalid obfuscations.* As for $\mathcal{A}$, this time it is trivially implemented by an oracle Decide^{Ψ} that, given input x, simply returns the unique bit b such that $V_b(x) = 1$, namely it just decides the corresponding language L^{Ψ}.

In the results mentioned above [Rud88, MM11, AS16], it is actually shown that deciding $\mathsf{NP}^{\Psi} \cap \mathsf{coNP}^{\Psi}$ does not require an explicit Decide oracle. Rather, it is possible to simulate any decision making a small number of queries to Ψ.[8] We do not show such a simulation process. Instead, we take a different approach inspired by our proof for the SZK setting described above. Roughly speaking, we show that somewhat similarly to our statistical difference oracle StaDif^{Ψ}, the oracle Decide^{Ψ} is also rather robust to random local changes. The main observation here is that for any fixed yes-instance $x \in L^{\Psi}$, tweaking Ψ at a random input into a new oracle Ψ', it is likely that x will still be a yes-instance in $L^{\Psi'}$, as long as Ψ' is in our allowed family of oracles and $L^{\Psi'}$ is indeed in $\mathsf{NP}^{\Psi'} \cap \mathsf{coNP}^{\Psi'}$ (and the same is true for no-instances).

In slightly more detail, fixing a witness w such that $V_1^{\Psi}(x, w) = 1$, we can show that since V_1 makes a small number of oracle calls, with high probability tweaking the oracle Ψ at a random place will not affect these oracle calls and thus $V_1^{\Psi'}(x, w) = V_1^{\Psi}(x, w) = 1$. Then, assuming $L^{\Psi'}$ is guaranteed to be in $\mathsf{NP} \cap \mathsf{coNP}$, we can deduce that x must still be a yes-instance (other witnesses for this fact may be added or disappear, but this does not change the oracle's answer). In the body, we argue that indeed $L^{\Psi'} \in \mathsf{NP}^{\Psi'} \cap \mathsf{coNP}^{\Psi'}$, where we strongly rely on the fact that arbitrary behavior of Eval is permitted on invalid obfuscations.

Once again, we show that through local changes that go undetected by Decide^{Ψ}, we can move to an ideal world where inverting the IOWF or breaking IO can be easily shown to be impossible. We refer the reader to section 4 for further details.

Implied separations. As a result of the two separations discussed above, we can rule out black-box constructions of hard problems in SZK or $\mathsf{NP} \cap \mathsf{coNP}$ from various cryptographic primitives or complexity classes. This applies to almost all primitives that have so far been constructed from OWPs (or IOWFs) and IO for circuits with OWP (or IOWF) gates. This includes public-key encryption, oblivious

[7] For example, the language of all valid obfuscations and indices i, such that the ith bit of the obfuscated circuit is 1.

[8] More accurately, this is the case for Rudich's result for $\mathsf{NP} \cap \mathsf{coNP}$, whereas for the other results that rule out constructions of OWPs, one can simulate an analogue of Decide that inverts the permutation.

transfer, deniable encryption [SW14],[9] functional encryption [Wat15], delegation, [BGL$^+$15, CHJV15, KLW15], hard (on-average) PPAD instances [BPR15], and more.

We note that there are a few applications of IO that do not fall under this characterization. For instance, the construction of IO for Turing machines from IO-based succinct randomized encodings [BGL$^+$15, CHJV15, KLW15] involves obfuscating a circuit that itself outputs (smaller) obfuscated circuits. To capture this, we would need to extend the above model to IO for circuits that can also make IO oracle calls (on smaller circuits). Another example is the construction of noninteractive witness indistinguishable proofs from IO [BP15]. There an obfuscated circuit may get as input another obfuscated circuit and would have to internally run it; furthermore, in this application, the code of the obfuscator is used in a (non-black-box) ZAP [DN00]. Extending the above model to account for this type of IO application is an interesting question that we leave for future exploration.

2. Preliminaries. In this section, we introduce the basic definitions and notation used throughout the paper.

2.1. Conventions. For a distribution D, we denote the process of sampling from D by $x \leftarrow D$. A function $\mathsf{negl} : \mathbb{N} \to \mathbb{R}^+$ is negligible if for every constant c, there exists a constant n_c such that for all $n > n_c$ $\mathsf{negl}(n) < n^{-c}$. We refer to uniform probabilistic polynomial-time algorithms as PPT algorithms.

Randomized algorithms. As usual, for a random algorithm A, we denote by $A(x)$ the corresponding output distribution. When we want to be explicit about the algorithm using randomness r, we shall denote the corresponding output by $A(x; r)$.

Oracles. We consider *oracle-aided algorithms (or circuits)* that make repeated calls to an oracle Γ. Throughout, we will consider deterministic oracles Γ that are a priori sampled from a distribution Γ on oracles. More generally, we consider infinite oracle ensembles $\Gamma = \{\Gamma_n\}_{n\in\mathbb{N}}$, one distribution Γ_n for each security parameter $n \in \mathbb{N}$ (each defined over a finite support). For example, we may consider an ensemble $f = \{f_n\}$ where each $f_n : \{0,1\}^n \to \{0,1\}^n$ is a random function. For such an ensemble Γ and an oracle-aided algorithm (or circuit) A with finite running time, we will often abuse notation and denote by $A^\Gamma(x)$ and execution of A on input x where each of the (finite number of) oracle calls that A makes is associated with a security parameter n and is answered by the corresponding oracle Γ_n. When we write $A_1^\Gamma, \ldots, A_k^\Gamma$ for k algorithms, we mean that they all access the same realization of Γ.

2.2. Indistinguishability obfuscation for oracle-aided circuits. The notion of IO was introduced by Barak et al. [BGI$^+$01] and the first candidate construction was demonstrated in the work of Garg et al. [GGH$^+$13a]. Since then, IO has given rise to a plethora of applications in cryptography and beyond. Nevertheless, Asharov and Segev [AS15, AS16] demonstrated that IO is insufficient to achieve some cryptographic tasks, most notably (domain-invariant) OWPs, collision-resistant hashing, and as a corollary, PIR and (even additively) homomorphic encryption. To formally

[9] Formally, the construction of deniable encryption described in [SW14] does not conform with the framework we consider; however, it can be easily adapted so that it does. Specifically, the construction involves obfuscating a circuit C^E that internally makes calls to a public-key encryption circuit E. When instantiating the public-key encryption circuit E based on IO and OWFs (as in [SW14]), this leads to "double-obfuscation"—the encryption circuit E itself is already an obfuscated version of another circuit E'. The framework that we consider does not support such double obfuscation. To adapt the construction, we can simply consider the obfuscation of $C^{E'}$, namely, the external layer of obfuscation is enough; this is because any functionally preserving changes to the internal E' are functionally preserving changes to the circuit $C^{E'}$ that will be protected by the external obfuscation.

show such a statement, they introduced the IO framework for oracle-aided circuits. We follow their framework.

We begin by recalling the notion of two oracle-aided circuits being equivalent, and move on to defining IO relative to oracles.

DEFINITION 2.1. *Let C_0 and C_1 be two oracle-aided circuits and let f be a function. C_0 and C_1 are said to be* functionally equivalent relative to f*, denoted as $C_0^f \equiv C_1^f$, if for every input x, $C_0^f(x) = C_1^f(x)$.*

DEFINITION 2.2. *Let $\mathcal{C} = \{\mathcal{C}_n\}_{n\in\mathbb{N}}$ be a class of oracle-aided circuits, where each $C \in \mathcal{C}_n$ is of size n.*[10] *A* PPT *algorithm $i\mathcal{O}$ is an* indistinguishability obfuscator *for $\mathcal{C}$ relative to an oracle distribution ensemble $\Gamma = \{\Gamma_n\}_{n\in\mathbb{N}}$ if the following conditions are met:*

1. Functionality. *For all $n \in \mathbb{N}$ and for all $C \in \mathcal{C}_n$ it holds that*
$$\Pr_{\Gamma, i\mathcal{O}}\left[C^\Gamma \equiv \widehat{C}^\Gamma \mid \widehat{C} \leftarrow i\mathcal{O}^\Gamma(1^n, C)\right] = 1.$$
2. Indistinguishability. *For any nonuniform* PPT *distinguisher $D = (D_1, D_2)$ there exists a negligible function* negl *such that for all $n \in \mathbb{N}$*
$$\mathsf{Adv}^{\mathsf{iO}}_{\Gamma, i\mathcal{O}, \mathcal{C}, D}(n) = \left|\Pr\left[\mathsf{Exp}^{\mathsf{iO}}_{\Gamma, i\mathcal{O}, \mathcal{C}, D}(n) = 1\right] - \frac{1}{2}\right| \le \mathsf{negl}(n),$$
where the random variable $\mathsf{Exp}^{\mathsf{iO}}_{\Gamma, i\mathcal{O}, \mathcal{C}, D}(n)$ is defined via the following experiment:
   (a) $b \leftarrow \{0,1\}$.
   (b) $(C_0, C_1, \mathsf{state}) \leftarrow D_1^\Gamma(1^n)$ *where* $C_0, C_1 \in \mathcal{C}_n$ *and* $C_0^\Gamma \equiv C_1^\Gamma$.
   (c) $\widehat{C} \leftarrow i\mathcal{O}^\Gamma(1^n, C_b)$.
   (d) $b' = D_2^\Gamma(\mathsf{state}, \widehat{C})$.
   (e) *If $b = b'$ output 1, else output 0.*
   We further say that $i\mathcal{O}$ satisfies δ-indistinguishability if the above negligible advantage is at most δ.

We will also consider the following definition of "positive advantage" in the security game above. Our actual proofs would bound the positive advantage. This suffices due to a result by Brakerski and Goldreich [BG11] that gives a (black-box) transformation between the two notions.

DEFINITION 2.3. *For any oracle Γ and nonuniform admissible* PPT *distinguisher $D = (D_1, D_2)$, define the* positive advantage *of D, denoted* $\mathrm{PAdv}^{\mathsf{iO}}_{\Gamma, i\mathcal{O}, \mathcal{C}, D}(n)$*, as follows:*
$$\mathrm{PAdv}^{\mathsf{iO}}_{\Gamma, i\mathcal{O}, \mathcal{C}, D}(n) = \Pr\left[\mathsf{Exp}^{\mathsf{iO}}_{\Gamma, i\mathcal{O}, \mathcal{C}, D}(n) = 1\right] - \frac{1}{2},$$
where the random variable $\mathsf{Exp}^{\mathsf{iO}}_{\Gamma, i\mathcal{O}, \mathcal{C}, D}(n)$ is defined as above in Definition 2.2.

By definition, $\mathsf{Adv}^{i\mathcal{O}}_{\Gamma, O, \mathcal{C}, D}(n) = |\mathrm{PAdv}^{i\mathcal{O}}_{\Gamma, O, \mathcal{C}, D}(n)|$. Note that for a distinguisher D to have
$$\mathrm{PAdv}^{i\mathcal{O}}_{\Gamma, O, \mathcal{C}, D}(n) > \varepsilon$$
for some $\varepsilon > 0$ is a stronger condition than $\mathsf{Adv}^{i\mathcal{O}}_{\Gamma, O, \mathcal{C}, D}(n) > \varepsilon$ because this requires the distinguisher to correctly predict which circuit was obfuscated better than chance,

[10] As in [AS15], we assume throughout that the size of the obfuscated circuits equals the security parameter. This is only for simplicity of notation and is without loss of generality as the circuits can be padded up if they are too small, and the security parameter can be polynomially increased if the circuits are too large.

instead of just being sufficiently far away from a random outcome. Brakerski and Goldreich showed that there is an efficient procedure that can transform any distinguisher with nonnegligible advantage to another distinguisher with nonnegligible positive advantage. Below we state their result in the context of our application.

LEMMA 2.4 (Brakerski–Goldreich [BG11]). *Given any distinguisher* $D=(D_1, D_2)$ *such that*

$$\mathop{\mathbb{E}}_{\Gamma}\left[\mathsf{Adv}^{\mathsf{iO}}_{\Gamma,i\mathcal{O},\mathcal{C},D}(n)\right] > \varepsilon(n) \ ,$$

for some function ε*, there exists a distinguisher* $D' = (D'_1, D'_2)$ *that makes* $O(1)$ *black-box invocations of* D *such that*

$$\mathop{\mathbb{E}}_{\Gamma}\left[\mathrm{PAdv}^{\mathsf{iO}}_{\Gamma,i\mathcal{O},\mathcal{C},D'}(n)\right] > 2\varepsilon(n)^2 \ .$$

The proof follows directly from [BG11] and can be found explicitly in [BD19].

3. One-way permutations, indistinguishability obfuscation, and hardness in statistical zero knowledge. In this section, we ask which cryptographic primitives imply hardness in the class SZK. Roughly speaking, we show that OWPs and IO, for circuits with OWP gates, do not give rise to a black-box construction of hard problems in SZK. This in turn implies that many cryptographic primitives (e.g., public-key encryption, functional encryption, and delegation), and hardness in certain low-level complexity classes (e.g., PPAD), also do not yield black-box constructions of hard problems in SZK.

We first motivate and define a framework of SZK relative to oracles, define fully black-box constructions of hard SZK problems, and then move on to the actual separation.

3.1. SZK and statistical difference. The notion of SZK proofs was introduced in the seminal work of Goldwasser, Micali, and Rackoff [GMR85]. The class of promise problems with SZK proofs can be characterized by several complete problems, such as *statistical difference* [SV03] and *entropy difference* [GV99] (see also [Vad99] and references within). We shall focus on the characterization of SZK by the statistical difference problem. Here an instance is a pair of circuit samplers $C_0, C_1 : \{0,1\}^n \to \{0,1\}^m$ with the promise that the statistical distance $s = \Delta(\boldsymbol{C}_0, \boldsymbol{C}_1)$ of the corresponding distributions is either large (say, $s \geq 2/3$) or small (say, $s \leq 1/3$). The problem is to decide which is the case.

Hard statistical difference problems from cryptography: Motivation. SZK hardness, and in particular hard statistical difference problems, is known to follow from various number-theoretic and lattice problems that are commonly used in cryptography, such as decision Diffie–Hellman, quadratic residuosity, and learning with errors. We ask more generally which cryptographic primitives can be shown to imply such hardness, with the intuition that such primitives are *structured* in a certain way. In particular, whereas one would not expect a completely unstructured object like OWFs to imply such hardness, what can we say, for instance, about public-key encryption, or even IO (which has proven to be structured enough to yield almost any known cryptographic goal).

We prove that none of these primitives imply such hardness through the natural class of black-box constructions and security reductions. To understand what a black-box construction of a hard statistical difference problem means, let us look at a specific example of the construction of such a problem from *rerandomizable encryption*. In a (say, symmetric-key) rerandomizable encryption scheme, on top of the usual encryption and decryption algorithms $(\mathsf{Enc}, \mathsf{Dec})$ there is a ciphertext

rerandomization algorithm ReRand that can statistically refresh ciphertexts. Namely, for any ciphertext CT encrypting a bit b, $\mathsf{ReRand}(\mathsf{CT})$ produces a ciphertext that is statistically close to a fresh encryption $\mathsf{Enc}(b)$. Note that this immediately gives rise to a hard statistical difference problem: given a pair of ciphertexts $(\mathsf{CT}, \mathsf{CT}')$, decide whether the corresponding rerandomized distributions given by the circuits $(C_0(\cdot), C_1(\cdot)) := (\mathsf{ReRand}(\mathsf{CT}; \cdot), \mathsf{ReRand}(\mathsf{CT}'; \cdot))$ are statistically far or close. Indeed, this corresponds to whether they encrypt the same bit or not, which is hard to decide by the security of the encryption scheme.

A feature of this construction of hard statistical difference instances is that, similarly to most constructions in cryptography, it is *fully black-box* [RTV04] in the sense that the circuits C_0, C_1 only make black-box use of the encryption scheme's algorithms and can in fact be represented as oracle-aided circuits $(C_0^{\mathsf{ReRand}(\cdot)}, C_1^{\mathsf{ReRand}(\cdot)})$. Furthermore, "hardness" can be shown by a black-box reduction that can use any decider for the problem in a black-box way to break the underlying encryption scheme. More generally, one can consider the statistical difference problem relative to different oracles implementing different cryptographic primitives and ask when hardness can be shown based on a black-box reduction. We will rule out such reductions relative to IO and OWPs (and everything that follows from these in a fully black-box way).

3.2. Fully black-box constructions of hard SD problems from IO and OWPs. We start by defining the statistical difference problem relative to oracles. In the following definition, for an oracle-aided (sampler) circuit $C^{(\cdot)}$ with a k-bit input and an oracle Ψ, we denote by $\boldsymbol{C}^{\Psi}$ the output distribution $C^{\Psi}(r)$ where $r \leftarrow \{0,1\}^k$. For two distributions $\mathbf{X}$ and $\mathbf{Y}$ we denote their statistical distance by $\Delta(\mathbf{X}, \mathbf{Y})$.

Definition 3.1 (statistical difference relative to oracles). *For an oracle Ψ, the statistical difference promise problem relative to Ψ, denoted as $\mathbf{SD}^{\Psi} = (\mathbf{SD}_Y^{\Psi}, \mathbf{SD}_N^{\Psi})$, is given by*

$$\mathbf{SD}_Y^{\Psi} = \left\{ (C_0, C_1) \;\middle|\; \Delta(\boldsymbol{C}_0^{\Psi}, \boldsymbol{C}_1^{\Psi}) \geq \frac{2}{3} \right\},$$
$$\mathbf{SD}_N^{\Psi} = \left\{ (C_0, C_1) \;\middle|\; \Delta(\boldsymbol{C}_0^{\Psi}, \boldsymbol{C}_1^{\Psi}) \leq \frac{1}{3} \right\}.$$

We now formally define the class of constructions and reductions ruled out, that is, *fully black-box* constructions of hard statistical distance problems from OWPs and IO for OWP-aided circuits. The definition is similar in spirit to those in [AS15, AS16], adapted to our context of SZK-hardness.

Definition 3.2. *A fully black-box construction of a hard statistical distance problem from OWPs and IO for the class $\mathcal{C}$ of circuits with OWP gates consists of a family $\Pi = \{\Pi_n\}_{n \in \mathbb{N}}$, where each Π_n is a set of oracle-aided circuit pairs $(C_0^{(\cdot)}, C_1^{(\cdot)}) \in \{0,1\}^{n \times 2}$, and a probabilistic oracle-aided reduction $\mathcal{R}$ that satisfy the following:*

- Black-box security proof. *There exist functions $q_{\mathcal{R}}(\cdot), \varepsilon_{\mathcal{R}}(\cdot)$ such that the following holds. Let $f = \{f_n\}$ be any family of permutations and let $i\mathcal{O}$ be any function family such that $\widehat{C}^f \equiv C^f$ for any $C^{(\cdot)}$ and r, where $\widehat{C}^{(\cdot)} := i\mathcal{O}(C^{(\cdot)}, r)$. Then for any probabilistic oracle-aided $\mathcal{D}$ that* decides Π in the worst case*, namely, for all $n \in \mathbb{N}$,*

$$\Pr_{\mathcal{D}}\left[\mathcal{D}^{f, i\mathcal{O}}(C_0, C_1) = B \quad \textit{for all} \quad \begin{array}{l} (C_0, C_1) \in \Pi_n, B \in \{Y, N\} \\ \textit{such that } (C_0, C_1) \in \mathbf{SD}_B^{f, i\mathcal{O}} \end{array} \right] = 1,$$

the reduction breaks either f or $i\mathcal{O}$, *namely, for infinitely many* $n \in \mathbb{N}$ *either*

$$\Pr_{\mathcal{D},x\leftarrow\{0,1\}^n}\left[\mathcal{R}^{\mathcal{D},f,i\mathcal{O}}(f(x)) = x\right] \geq \varepsilon_{\mathcal{R}}(n)$$

or

$$\mathsf{Adv}^{\mathsf{iO}}_{\Gamma,i\mathcal{O},\mathcal{C},D}(n) \geq \varepsilon_{\mathcal{R}}(n),$$

where in both $\mathcal{R}$ *makes at most* $q_{\mathcal{R}}(n)$ *queries to any of its oracles* $(\mathcal{D}, f, i\mathcal{O})$, *and any query* $(C_0^{(\cdot)}, C_1^{(\cdot)})$ *it makes to* $\mathcal{D}$ *consists of circuits that also make at most* $q_{\mathcal{R}}(n)$ *queries to their oracles* $(f, i\mathcal{O})$. *Random variable* $\mathsf{Adv}^{\mathsf{iO}}_{(f,i\mathcal{O}),i\mathcal{O},\mathcal{C},\mathcal{R}^{\mathcal{D}}}(n)$ *represents the reductions winning probability in the IO security game (Definition* 2.2*) relative to* $(f, i\mathcal{O})$.

We make several remarks about the definition:

- *Correctness.* Typically, we also require certain *correctness* from the black-box construction. For instance, in the next section, we shall require that the construction always satisfies the $\mathsf{NP}\cap\mathsf{coNP}$ structure. In the above definition, the construction is allowed to yield instances $(C_0^{f,i\mathcal{O}}, C_1^{f,i\mathcal{O}})$ that do not satisfy the SZK promise, namely $(C_0^{f,i\mathcal{O}}, C_1^{f,i\mathcal{O}}) \notin \mathbf{SD}_Y^{f,i\mathcal{O}} \cup \mathbf{SD}_N^{f,i\mathcal{O}}$. It is natural to think of more stringent definitions that require that the corresponding problem $\Pi^{f,i\mathcal{O}}$ is nontrivial, in the sense that $\Pi^{f,i\mathcal{O}} \cap \mathbf{SD}_Y^{f,i\mathcal{O}} \neq \emptyset$ and $\Pi^{f,i\mathcal{O}} \cap \mathbf{SD}_N^{f,i\mathcal{O}} \neq \emptyset$ (which is the case for known constructions of SZK hardness from cryptographic primitives). Our impossibility is more general and would, in particular, rule out such definitions as well.
- *Worst-case versus average-case hardness.* In the above, we address *worst-case hardness*, in the sense that the reduction $\mathcal{R}$ has to break the underlying primitives only given a decider $\mathcal{D}$ that is always correct. One could further ask whether IO and OWPs even imply average-case hardness in SZK (as do many of the algebraic hardness assumptions in cryptography). Ruling out worst-case hardness (as we will do shortly) in particular rules out such average-case hardness as well.
- *IO for oracle-aided circuits.* Following [AS15, AS16], we consider IO for oracle-aided circuits C^f that can make calls to the OWP oracle. This model captures constructions where IO is applied to circuits that use pseudorandom generators, puncturable pseudorandom functions, or IOWFs as all of those have fully black-box constructions from OWPs (see further discussion in [AS15]). This includes almost all known constructions from IO, including public-key encryption, deniable encryption [SW14], functional encryption [Wat15], delegation [BGL$^+$15, CHJV15, KLW15], and hard (on average) PPAD instances [BPR15]. Accordingly, separating SZK from IO and OWPs in this model results in a similar separation between SZK and any one of these primitives.

  We note that there are a few applications though that do not fall under this model. The first is in applications where the obfuscated circuit might itself output an (smaller) obfuscated circuit, for instance, in the construction of IO for Turing machines from IO-based succinct randomized encodings [BGL$^+$15, CHJV15, KLW15]. To capture such applications, one would have to extend the model to also account for circuits with IO gates (and not only OWP gates). A second example is the construction of noninteractive witness indistinguishable proofs from IO [BP15]. There an obfuscated circuit may

get as input another obfuscated circuit and would have to internally run it; furthermore, in this application, the code of the obfuscator is used in a (non-black-box) ZAP. Extending our results (and those of [AS15, AS16]) to these models is an interesting question, left for future work.

- *Security loss.* In the above definition the functions $q_{\mathcal{R}}$ and $\varepsilon_{\mathcal{R}}$ capture the *security loss* of the reduction. Most commonly in cryptography, the query complexity is polynomial $q_{\mathcal{R}}(n) = n^{O(1)}$ and the probability of breaking the underlying primitive is inverse polynomial $\varepsilon_{\mathcal{R}}(n) = n^{-O(1)}$. Our lower bounds will in fact apply for *exponential* $q_{\mathcal{R}}, \varepsilon_{\mathcal{R}}^{-1}$. This allows capturing also constructions that rely on subexponentially secure primitives (e.g., [BGL$^+$15, CHJV15, KLW15, BPR15, BPW16]).

Ruling out fully black-box constructions: A road map. Our main result in this section is that a fully black-box construction of a hard statistical difference problem from IO and OWPs does not exist. Furthermore, this holds even if the latter primitives are exponentially secure.

THEOREM 3.3. *Any fully black-box construction of a statistical difference problem* Π *from OWPs and IO for circuits with OWP gates has an exponential security loss:* $\max(q_{\mathcal{R}}(n), \varepsilon_{\mathcal{R}}^{-1}(n)) \geq \Omega(2^{n/10})$.

The proof of the theorem follows a common methodology (applied, for instance, in [HR04, HHRS15, AS15]). We exhibit two (distributions on) oracles $(\Psi, \mathsf{StaDif}^{\Psi})$, where Ψ realizes OWPs and IO for circuits with OWP gates, and StaDif^{Ψ} that decides $\mathbf{SD}^{\Psi}$, the statistical difference problem relative to Ψ, in the worst case. Since $\mathbf{SD}$ is complete for SZK in a relativizing manner, solving $\mathbf{SD}^{\Psi}$ suffices to break SZK^{Ψ}. We then show that the primitives realized by Ψ are (exponentially) secure even in the presence of StaDif^{Ψ}. This statement is proved when the oracle Ψ is sampled at random from the constructed distribution. Since StaDif^{Ψ} solves $\mathbf{SD}^{\Psi}$, in the worst case, for every oracle Ψ, for every given reduction, there is a fixed oracle $\Gamma = (\Psi, \mathsf{StaDif}^{\Psi})$ relative to which (1) $\mathbf{SD}$ is easy to decide, (2) the reduction fails to break IO (or OWPs). This implies Theorem 3.3, ruling out fully black-box constructions with a subexponential security loss.

The rest of this section is organized according to the above plan. First, in subsection 3.3, we describe the oracle StaDif^{Ψ} (which is independent of the specific way that Ψ realizes IO and OWPs). Then, in subsections 3.4 and 3.5, we describe the oracle Ψ realizing OWPs and IO and prove its (exponential) security in the presence of StaDif^{Ψ}.

3.3. A noisy statistical-distance oracle. We now define the oracle StaDif^{Ψ} that will solve the statistical difference problem $\mathbf{SD}^{\Psi}$ in all the separations proved in this section. Our goal is to design StaDif^{Ψ} in a way that will not break the security of the cryptographic primitives realized by Ψ (OWPs in the warmups, and then OWPs and IO for circuits with OWP gates). For this purpose, in our definition of the oracle StaDif^{Ψ}, we will try to exploit the fact that statistical distance is insensitive to *local changes* in the input distributions. Then, we will show that breaking the relevant cryptographic primitives, captured by Ψ, is impossible without detecting such local changes.

The concrete way of capturing the spoken insensitivity will be to define a "noisy oracle" that would be correct on distribution pairs whose distance is within the promise range $[0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$, but would behave randomly within $(\frac{1}{3}, \frac{2}{3})$.

DEFINITION 3.4 (oracle StaDif^{Ψ}). *The oracle consists of* $\mathbf{t} = \{\mathbf{t}_n\}_{n\in\mathbb{N}}$ *where* $\mathbf{t}_n : \{0,1\}^{2n} \to (\frac{1}{3}, \frac{2}{3})$ *is a uniformly random function. Given* n*-bit descriptions of oracle-aided circuits* $C_0, C_1 \in \{0,1\}^n$*, let* $t = \mathbf{t}_n(C_0, C_1)$*, and let* $s = \Delta(\boldsymbol{C}_0^{\Psi}, \boldsymbol{C}_1^{\Psi})$*, and return*

$$\mathsf{StaDif}^{\Psi}(C_1, C_2; t) := \begin{cases} N & \text{if } s < t, \\ Y & \text{if } s \geq t. \end{cases}$$

It is immediate to see that StaDif^{Ψ} decides $\mathbf{SD}^{\Psi}$ in the worst case.

CLAIM 3.5. *For any oracle* Ψ*,*

$$\mathbf{SD}^{\Psi} \in \mathsf{P}^{\Psi, \mathsf{StaDif}^{\Psi}}.$$

The main challenge is in showing that Ψ can implement OWPs and IO (for OWP-aided circuits) that will be secure in the presence of StaDif^{Ψ}. We next develop the terminology and establish several useful properties of StaDif that will allow us to carry out the above plan.

Capturing insensitivity to local changes. We introduce two general notions of *farness* and *smoothness* that aim to capture the sense in which the statistical difference oracle StaDif^{Ψ} defined above is insensitive to local changes.

Roughly speaking *farness* says that the random threshold t used for a query (D_0, D_1) to StaDif^{Ψ} is "far" from the actual statistical distance. We will show that with high probability over the choice of random threshold $\mathbf{t}$, farness holds for all queries (D_0, D_1) made to StaDif^{Ψ} by any (relatively) efficient adversary. This intuitively means that changing the distributions $(\boldsymbol{D}_0^{\Psi}, \boldsymbol{D}_1^{\Psi})$, on sets of small density, will not change the oracle's answer.

DEFINITION 3.6 (farness). *The oracles* $(\Psi, \mathsf{StaDif}^{\Psi})$ *satisfy* δ-farness *with respect to oracle-aided circuits* $(D_0, D_1) \in \{0,1\}^n$ *if the statistical difference* $s = \Delta(\boldsymbol{D}_0^{\Psi}, \boldsymbol{D}_1^{\Psi})$ *and the threshold* $t = \mathbf{t}_n(D_0, D_1)$ *sampled by* StaDif *are* δ*-far:*

$$|s - t| \geq \delta.$$

For an adversary $\mathcal{D}$*, we denote by* $\mathbf{Far}(\mathcal{D}, \Psi, \delta)$ *the event that* $\Gamma = (\Psi, \mathsf{StaDif}^{\Psi})$ *satisfies* δ*-farness for all queries* (D_0, D_1) *made by* $\mathcal{D}$ *to* StaDif^{Ψ}*.*

CLAIM 3.7. *Fix any* Ψ *and any oracle-aided adversary* $\mathcal{D}$ *such that* $\mathcal{D}^{\Psi, \mathsf{StaDif}^{\Psi}}$ *makes at most* q *queries to* StaDif^{Ψ}*. Then*

$$\Pr_{\mathbf{t}}[\mathbf{Far}(\mathcal{D}, \Psi, \delta)] \geq 1 - 6\delta q,$$

where the probability is over the choice $\mathbf{t}$ *of random thresholds by* StaDif*.*

Proof. This follows from the fact that, for any query (D_0, D_1) to StaDif^{Ψ} with $s = \Delta(\boldsymbol{D}_0^{\Psi}, \boldsymbol{D}_1^{\Psi})$, δ-farness does not hold only if the threshold $t = \mathbf{t}(D_0, D_1)$, chosen at random for this query, happens to be in the interval $(s - \delta, s + \delta)$, which occurs with probability at most $|(s - \delta, s + \delta)| / |(\frac{1}{3}, \frac{2}{3})| = 6\delta$, since $t = \mathbf{t}(D_0, D_1)$ is sampled uniformly at random independently of (D_0, D_1). The lemma then follows by a union bound over at most q queries. ∎

We now turn to define the notion of *smoothness*. Roughly speaking, we say that an oracle-aided circuit D is smooth with respect to some oracle Ψ and a set of inputs

T if the circuit, on a random input, queries the oracle Ψ at any location $x \in T$ with low probability. In particular, for a pair of smooth circuits (D_0, D_1), changes in how the oracle Ψ behaves on T should not change significantly the statistical distance $s = \Delta(\boldsymbol{D}_0^{\Psi}, \boldsymbol{D}_1^{\Psi})$.

DEFINITION 3.8 ((Ψ, T, δ)-smoothness). *An oracle-aided circuit $D^{(\cdot)} : \{0,1\}^n \to \{0,1\}^m$ is said to be (Ψ, T, δ)-smooth if*

$$\Pr_{w \leftarrow \{0,1\}^n} \left[D^{\Psi}(w) \text{ queries } \Psi \text{ at any } x \in T \right] \leq \delta.$$

For an adversary $\mathcal{D}$, we denote by $\mathbf{Smo}(\mathcal{D}, \Psi, T, \delta)$ the event that all queries (D_0, D_1) made by $\mathcal{D}$ to StaDif^{Ψ} are (Ψ, T, δ)-smooth.

CLAIM 3.9. *Let Ψ, Ψ' be oracles that are identical everywhere outside T. Let (D_0, D_1) be (Ψ, T, δ)-smooth. Let $s = \Delta(D_0^{\Psi}, D_1^{\Psi})$ and $s' = \Delta(D_0^{\Psi'}, D_1^{\Psi'})$ then $|s - s'| \leq 2\delta$.*

Proof. For either $b \in \{0,1\}$,

$$\begin{aligned} &\Delta(D_b^{\Psi}, D_b^{\Psi'}) \\ &\quad \leq \Pr_w \left[D_b^{\Psi}(w) \neq D_b^{\Psi'}(w) \right] \\ &\quad \leq \Pr_w \left[D_b^{\Psi}(w) \text{ queries } \Psi \text{ at } x \in T \right] \leq \delta. \end{aligned}$$

The claim then follows by the fact that

$$|s - s'| := \left| \Delta(C_0^{\Psi}, C_1^{\Psi}) - \Delta(C_0^{\Psi'}, C_1^{\Psi'}) \right| \leq \Delta(C_0^{\Psi}, C_0^{\Psi'}) + \Delta(C_1^{\Psi}, C_1^{\Psi'}) \leq 2\delta. \qquad \square$$

3.4. Warmup: One-way permutations in the presence of StaDif. In this section, we show that a random permutation f is hard to invert even given access to the noisy statistical difference oracle StaDif^f. We start by defining the oracle. In what follows, $\mathbf{P}_n$ denotes the set of permutations of $\{0,1\}^n$.

DEFINITION 3.10 (the oracle f). *$f = \{f_n\}_{n \in \mathbb{N}}$ on input $x \in \{0,1\}^n$ answers with $f_n(x)$ where f_n is a random permutation $f_n \leftarrow \mathbf{P}_n$.*

Our main theorem states that f cannot be inverted, except with exponentially small probability, even given an exponential number of oracle queries to f and StaDif^f. We say that an adversary $\mathcal{D}$ is q-query if $\mathcal{D}^{f, \mathsf{StaDif}^f}$ makes at most q queries to f and q queries to StaDif^f, and any query made to StaDif^f consists of oracle-aided circuits (D_0, D_1) that make at most q queries to f, on any specific input.

THEOREM 3.11. *Let $q \leq O(2^{n/5})$. Then for any q-query adversary $\mathcal{D}$*

$$\Pr_{f, \mathsf{StaDif}, x} \left[\mathcal{D}^{f, \mathsf{StaDif}^f}(f(x)) = x \right] \leq O(2^{-n/5}),$$

where the probability is over the random choices of f, StaDif and $x \leftarrow \{0,1\}^n$.

For the black-box reduction to succeed, it has to invert the OWF f for every f where StaDif^f breaks SZK (that is, for every f). We show that even for a random permutation f, the reduction cannot invert with high probability. This is stronger than what is needed to establish the required impossibility result.

At a very high level, the proof of the theorem follows the plan outlined above, showing that in order to invert a random permutation the adversary must be able to detect certain local changes to the permutation, which the noisy statistical difference oracle is insensitive to.

TABLE 3.1
The hybrid experiments.

Hybrid	$\mathbf{H_1}$ (Real)	$\mathbf{H_2}$	$\mathbf{H_2}$	$\mathbf{H_3}$ (Ideal)
Permutation	$f_n \leftarrow \mathbf{P}_n$			
Preimage	$x \leftarrow \{0,1\}^n$			
2nd preimage	$z \leftarrow \{0,1\}^n$			
Planted image	$y \leftarrow \{0,1\}^n$			
Challenge	$f(x)$		y	
Oracle	f, StaDif^f	$f_{z\mapsto f(x)}, \mathsf{StaDif}^{f_{z\mapsto f(x)}}$	$f_{x\mapsto y}, \mathsf{StaDif}^{f_{x\mapsto y}}$	f, StaDif^f
Winning condition	Find x			

Proof. We, in fact, prove a stronger statement: the above holds when fixing the oracles $f_{-n} := \{f_k\}_{k\neq n}$. Fix a q-query adversary $\mathcal{D}$. To bound $\mathcal{D}$'s inversion probability, we consider four hybrid experiments $\{\mathbf{H_i}\}_{i\in[4]}$ given in Table 3.1. Throughout, for a permutation $f \in \mathbf{P}_n$ and $x, y \in \{0,1\}^n$, we denote by $f_{x\mapsto y}$ the function that maps x to y and is identical to f on all other inputs (in particular, $f_{x\mapsto y}$ is no longer a permutation when $x \neq f^{-1}(y)$).

Hybrid $\mathbf{H_1}$ is identical to the real world where $\mathcal{D}$ wins if it successfully inverts the permutation at a random output. We show that the probability that the simulator wins in any of the experiments is roughly the same and that in hybrid $\mathbf{H_4}$ the probability that $\mathcal{D}$ wins is tiny.

CLAIM 3.12. $|\Pr[\mathcal{D} \textit{ wins in } \mathbf{H_1}] - \Pr[\mathcal{D} \textit{ wins in } \mathbf{H_2}]| \leq O(2^{-n/5})$.

Proof. The difference between the two hybrids is in the oracle that $\mathcal{D}$ is given: simply f in the first, and its slightly tweaked version $f_{z\mapsto f(x)}$ in the second. We can bound the difference between the winning probabilities in $\mathbf{H_1}$ and $\mathbf{H_2}$ as follows:

$$\begin{aligned}&|\Pr[\mathcal{D} \text{ wins in } \mathbf{H_1}] - \Pr[\mathcal{D} \text{ wins in } \mathbf{H_2}]| \\ &\quad\leq \Pr_{\substack{\mathcal{D},x,z\\ f,\mathsf{StaDif}}}\left[\mathcal{D}^{f,\mathsf{StaDif}^f}(f(x)) \neq \mathcal{D}^{f_{z\mapsto f(x)},\mathsf{StaDif}^{f_{z\mapsto f(x)}}}(f(x))\right],\end{aligned}$$

where the probability is over the coins of $\mathcal{D}$ and StaDif and the choice of $x, z \leftarrow \{0,1\}^n, f_n \leftarrow \mathbf{P}_n$.

In what follows, we denote by $\mathbf{Hit} = \mathbf{Hit}(\mathcal{D}, f, x, z)$ the event that $\mathcal{D}^{f,\mathsf{StaDif}^f}(f(x))$ queries f on z. Also, let $\mathbf{Far} = \mathbf{Far}(\mathcal{D}(f(x)), f, 2\delta)$ be the event that 2δ-farness holds for all StaDif-queries made by $\mathcal{D}^{f,\mathsf{StaDif}^f}(f(x))$ (Definition 3.6), and $\mathbf{Smo} = \mathbf{Smo}(\mathcal{D}(f(x)), f, z, \delta)$ is the event that for every StaDif-query (D_0, D_1) made by $\mathcal{D}^{f,\mathsf{StaDif}^f}(f(x))$ is $(f, T = \{z\}, \delta)$-smooth (see Definition 3.8).

We now claim as follows.

CLAIM 3.13. *For any $\delta < 1$,*

$$\Pr_{\substack{\mathcal{D},x,z\\ f,\mathsf{StaDif}}}\left[\mathcal{D}^{f,\mathsf{StaDif}^f}(f(x)) \neq \mathcal{D}^{f_{z\mapsto f(x)},\mathsf{StaDif}^{f_{z\mapsto f(x)}}}(f(x))\right] \leq \Pr_{\substack{\mathcal{D},x,z\\ f,\mathsf{StaDif}}}\left[\mathbf{Hit} \vee \overline{\mathbf{Far}} \vee \overline{\mathbf{Smo}}\right].$$

Proof. We argue that whenever the complement $\overline{\mathbf{Hit}} \wedge \mathbf{Far} \wedge \mathbf{Smo}$ occurs then

$$\mathcal{D}^{f,\mathsf{StaDif}^f}(f(x)) = \mathcal{D}^{f_{z\mapsto f(x)},\mathsf{StaDif}^{f_{z\mapsto f(x)}}}(f(x)).$$

For any StaDif-query (C_0, C_1) made by $\mathcal{D}^{f,\mathsf{StaDif}^f}(f(x))$, $\mathbf{Smo}(\mathcal{D}(f(x)), f, z, \delta)$ implies that changing f at z does not affect the statistical distance by much. Concretely, by Claim 3.9,

$$\left|\Delta\left(\boldsymbol{D}_0^f, \boldsymbol{D}_1^f\right) - \Delta\left(\boldsymbol{D}_0^{f_{z\mapsto f(x)}}, \boldsymbol{D}_1^{f_{z\mapsto f(x)}}\right)\right| \leq 2\delta.$$

Hence, if 2δ-farness also holds for any such query (for some threshold $\mathbf{t}$ sampled by StaDif), then

$$\mathsf{StaDif}^f(C_0, C_1; \mathbf{t}) = \mathsf{StaDif}^{f_{z\mapsto f(x)}}(C_0, C_1; \mathbf{t}).$$

If in addition **Hit** does not occur, then for any f-query w made by $\mathcal{D}^{f,\mathsf{StaDif}^f}(f(x))$,

$$f(w) = f_{z\mapsto f(x)}(w).$$

It follows that the views of $\mathcal{D}^{f,\mathsf{StaDif}^f}(f(x))$ and $\mathcal{D}^{f_{z\mapsto f(x)},\mathsf{StaDif}^{f_{z\mapsto f(x)}}}(f(x))$ are identical. ∎

It is left to bound the probability of each of the events $\mathbf{Hit}, \overline{\mathbf{Far}}, \overline{\mathbf{Smo}}$. First, noting that the view of $\mathcal{D}^{f,\mathsf{StaDif}^f}(f(x))$ is independent of the random z, we can bound

$$\Pr[\mathbf{Hit}] \leq 2^{-n} \cdot \#\{f\text{-queries made by } \mathcal{D}\} \leq 2^{-n} \cdot q.$$

Furthermore, by the farness Claim 3.7,

$$\Pr\left[\overline{\mathbf{Far}}\right] \leq 12q\delta.$$

Finally, we have as follows.

CLAIM 3.14. *For any fixed f, x,*

$$\Pr_z\left[\overline{\mathbf{Smo}(\mathcal{D}(f(x)), f, z, \delta)}\right] < 2^{-n} \cdot 2q^2/\delta\ .$$

Proof. Every circuit D that makes at most q queries has at most q/δ locations queried with probability more than δ. Taking into account the $2q$ queries (D_0, D_1) made by $\mathcal{D}$, there are overall at most $2q^2/\delta$ such queries. Since z is chosen uniformly at random and independently of these queries, it hits any specific one of them with probability 2^{-n}. ∎

Overall, we can bound the difference between $\mathbf{H_1}$ and $\mathbf{H_2}$ by

$$2^{-n} \cdot 2q^2/\delta + 2^{-n} \cdot q + 12q\delta \leq O(2^{-n/5}),$$

when setting $\delta = 2^{-2n/5}$, and recalling that $q \leq O(2^{n/5})$. ∎

CLAIM 3.15. $\Pr[\mathcal{D}\ \textit{wins in}\ \mathbf{H_2}] = \Pr[\mathcal{D}\ \textit{wins in}\ \mathbf{H_3}]$.

Proof. The difference between $\mathbf{H_2}$ and $\mathbf{H_3}$ is in the input of $\mathcal{D}$, $f(x)$ in the first and a random y in the second, and in the oracle $\mathcal{D}$ is given $f_{z\mapsto f(x)}$ in the first and $f_{x\mapsto y}$ in the second. We argue, however, that the distribution $\left\{(f(x), f_{z\mapsto f(x)}, x) \;\middle|\; f \leftarrow \mathbf{P}_n, x, z \leftarrow \{0,1\}^n\right\}$ in $\mathbf{H_1}$ is identical to that of $\{(y, f_{x\mapsto y}, x) \mid f \leftarrow \mathbf{P}_n, x, z \leftarrow \{0,1\}^n\}$ are in $\mathbf{H_2}$. Indeed, in $\mathbf{H_1}$, $(f(x), x)$ are distributed uniformly and independently just as (y, x) in $\mathbf{H_2}$. Then, conditioned on any (y, x), the oracle in both distributions can be sampled as a random permutation f conditioned on $y = f(x)$ and diverting a random z from $f(z)$ to y. ∎

CLAIM 3.16. $|\Pr[\mathcal{D}\ \textit{wins in}\ \mathbf{H_3}] - \Pr[\mathcal{D}\ \textit{wins in}\ \mathbf{H_4}]| \leq O(2^{-n/5})$.

The difference between the two hybrids is in the oracle that $\mathcal{D}$ is given: simply f in the second and its slightly tweaked version $f_{x\mapsto y}$ in the first. The proof of their indistinguishability is essentially identical to that of Claim 3.12, except that here we start with $\mathbf{H_4}$, consider the notion of smoothness with respect to x, and observe that it is independent of the execution.

To conclude the proof of Theorem 3.11, we observe as follows.

CLAIM 3.17. $\Pr[\mathcal{D} \text{ wins in } \mathbf{H_4}] \leq 2^{-n}$.

Proof. The view of $\mathcal{D}$ in this hybrid is completely independent of the random choice of x. ∎

3.5. Indistinguishability obfuscation (and OWPs) in the presence of StaDif. In this section, we consider an oracle Ψ that realizes both IO and OWPs and show that neither breaks in the presence of the noisy statistical difference oracle StaDif^{Ψ}. We start by defining the oracle Ψ. In a nutshell, the oracle realizes OWPs through a random permutation oracle. IO for circuits with OWP gates is captured in a similar way to [AS15] by a random injective mapping coupled with a corresponding evaluation algorithm.

In what follows, $\mathbf{P}_n$ denotes the set of permutations of $\{0,1\}^n$, $\mathbf{F}_n^m$ denotes the set of functions mapping $\{0,1\}^n$ to $\{0,1\}^m$, and $\mathbf{I}_n^m$ denotes the set of injective functions mapping $\{0,1\}^n$ to $\{0,1\}^m$.

DEFINITION 3.18 (the oracle Ψ). *The oracle $\Psi = (f, \mathcal{O}, \mathsf{Eval}^{f,\mathcal{O}})$ consists of three parts:*

- *$f = \{f_n\}_{n\in\mathbb{N}}$ on input $x \in \{0,1\}^n$ answers with $f_n(x)$, where f_n is a random permutation $f_n \leftarrow \mathbf{P}_n$.*
- *$\mathcal{O} = \{\mathcal{O}_n\}_{n\in\mathbb{N}}$ on input $(C,r) \in \{0,1\}^{n\times 2}$ answers with $\widehat{C} := \mathcal{O}_n(C,r)$ where $\mathcal{O}_n$ is a random injective function $\mathcal{O}_n \leftarrow \mathbf{I}_{2n}^{5n}$ into $\{0,1\}^{5n}$.*
- *$\mathsf{Eval}^{f,\mathcal{O}}$ given $\widehat{C} \in \{0,1\}^{5n}, x \in \{0,1\}^\star$ computes $(C,r) = \mathcal{O}_n^{-1}(\widehat{C})$, interprets C as an oracle-aided circuit, and returns $C^f(x)$. If the input size of C is inconsistent with $|x|$, the oracle returns $\perp$. We further extend the definition of Eval to the case that $\mathcal{O}$ is not injective: If $\widehat{C}$ does not have a unique preimage, the oracle returns $\perp$.*

In the next two subsections, we show that the oracle Ψ securely realizes OWPs and IO in the presence of the noisy statistical difference oracle StaDif^{Ψ}. Throughout, we address adversaries with oracles $\Psi = (f, \mathcal{O}, \mathsf{Eval}^{\mathcal{O},f})$ and StaDif^{Ψ}. We will say that such an adversary is q-query if they

1. make only q queries to f,
2. make only q queries to either $\mathcal{O}$ or Eval, and any query $\widehat{C}$ to Eval is of size at most $5q$, and in particular, any oracle-aided circuit C that is mapped to $\widehat{C}$ by $\mathcal{O}$ is of size at most q and makes at most q queries to f,
3. make only q queries to StaDif^{Ψ}, and for any query (C_0, C_1) made to StaDif^{Ψ}, (C_0, C_1) are Ψ-aided and each of them is q-query (according to the two conditions above).

3.5.1. One-way permutations. We show that f cannot be inverted, except with exponentially small probability even given an exponential number of oracle queries to $\Psi = (f, \mathcal{O}, \mathsf{Eval}^{\mathcal{O},f})$ and StaDif^{Ψ}. This is proved by a reduction to Theorem 3.11. We show that an even stronger adversary, one that gets access to $\mathcal{O}$ completely, cannot invert f. When the adversary gets complete access to $\mathcal{O}$ it can answer $\mathcal{O}, \mathsf{Eval}$ queries by itself and reduce $\mathsf{StaDif}^{f,\mathcal{O},\mathsf{Eval}}$ queries to StaDif^{f} queries.

THEOREM 3.19. *Let $q(n) \leq O(2^{n/10})$. Then for any q-query adversary $\mathcal{D}$*

$$\Pr_{\substack{\Psi=(f,\mathcal{O},\mathsf{Eval})\\ \mathsf{StaDif},x}}\left[\mathcal{D}^{\Psi,\mathsf{StaDif}^{\Psi}}(f(x)) = x\right] \leq O(2^{-n/5}),$$

where the probability is over the random choice of Ψ, StaDif and $x \leftarrow \{0,1\}^n$.

Proof. We will, in fact, prove a stronger statement: the above holds when fixing the oracles $f_{-n} := \{f_k\}_{k\neq n}$, $\mathcal{O} = \{\mathcal{O}_n\}_{n\in\mathbb{N}}$. We prove the theorem by a reduction to the case that Ψ only consists of the permutation f (and does not include $\mathcal{O}, \mathsf{Eval}$). Concretely, fixing any q-query adversary $\mathcal{D}$ that inverts the random permutation f_n given access to $\Psi = (f, \mathcal{O}, \mathsf{Eval})$ and StaDif^{Ψ}, we show how to reduce it to a q^2-query adversary $\mathcal{B}^f(f_n(x))$ that inverts f_n for a random $x \leftarrow \{0,1\}^n$ with the same probability as $\mathcal{D}$. The proof then follows from Theorem 3.11.

The new adversary $\mathcal{B}^{f,\mathsf{StaDif}^f}(f_n(x))$ emulates $\mathcal{D}^{\Psi,\mathsf{StaDif}^{\Psi}}(f_n(x))$ answering Ψ-queries as follows:

- f queries: answered according to $\mathcal{B}$'s oracle f. This translates to at most q queries to f.
- $\mathcal{O}$ queries: answered according to the fixed oracle $\mathcal{O}$. This does not add any calls to f.
- $\mathsf{Eval}^{f,\mathcal{O}}$ queries: given query $(\widehat{C}, x)$ to Eval, invert the fixed oracle $\mathcal{O}$ to find $(C, r) = \mathcal{O}^{-1}(\widehat{C})$. If no such preimage exists, return $\perp$. If a preimage does exist, using the f-oracle, compute $C^f(x)$ and return the result. This translates to at most q^2 queries to f: q queries by C, for each of the q queries $\widehat{C}$ to Eval.
- StaDif^{Ψ} queries: given query (C_0, C_1), where C_b makes Ψ-queries translate to D_0, D_1 that only make f-queries, where each query to $\Psi = (f, \mathcal{O}, \mathsf{Eval})$ is translated to a query to f according to the previous three items. The resulting oracle-aided (D_0, D_1) may thus make up to $q + q^2$ queries f: q corresponding to the first item and q^2 corresponding to the third.[11]

Overall $\mathcal{B}^f$ is $O(q^2)$-query and perfectly emulates the view of $\mathcal{D}^{\Psi}$. The theorem now follows from Theorem 3.11. ∎

3.5.2. Indistinguishability obfuscation. We now turn to show that Ψ also realizes an indistinguishability obfuscator that does not break in the presence of StaDif^{Ψ}. We start by describing the construction, which is similar to the one in [AS15].

CONSTRUCTION 3.20 (the obfuscator $i\mathcal{O}^{\Psi}$). *Let $\Psi = (f, \mathcal{O}, \mathsf{Eval}^{f,\mathcal{O}})$. Given an oracle-aided circuit $C \in \{0,1\}^n$, $i\mathcal{O}^{\Psi}(1^n, C)$ samples a random $r \leftarrow \{0,1\}^n$, computes $\widehat{C} = \mathcal{O}(C, r)$, and returns an oracle-aided circuit $E_{\widehat{C}}$ that given input x computes $\mathsf{Eval}^{f,\mathcal{O}}(\widehat{C}, x)$.*

It is easy to see that $i\mathcal{O}^{f,\mathcal{O},\mathsf{Eval}}$ satisfies the functionality requirement of Definition 2.2 for the class $\mathcal{C}$ of f-aided circuits; indeed, this follows by the fact that $\mathcal{O}$ is injective and by the definition of $i\mathcal{O}$ and the oracles $\mathcal{O}, \mathsf{Eval}$. We now show that it also satisfies indistinguishability, with an exponentially small distinguishing gap, even given an exponential number of oracle queries to $\Psi = (f, \mathcal{O}, \mathsf{Eval}^{\mathcal{O},f})$ and the statistical difference oracle StaDif^{Ψ}.

THEOREM 3.21. *Let $q(n) \leq O(2^{n/5})$. Then for any q-query distinguisher $\mathcal{D}$*

$$\mathop{\mathbb{E}}_{\mathcal{O}}\left[\mathrm{PAdv}^{\mathsf{iO}}_{\Gamma,i\mathcal{O},\mathcal{C},\mathcal{D}}(n)\right] \leq O(2^{-n/5}),$$

where the random variable $\mathrm{PAdv}^{\mathsf{iO}}_{\Gamma,i\mathcal{O},\mathcal{C},\mathcal{D}}(n)$ denotes the adversary's positive distinguishing advantage in the IO security game (Definition 2.3) relative to $\Psi = (f, \mathcal{O}, \mathsf{Eval}^{f,\mathcal{O}})$ and StaDif^{Ψ}.

[11] We note that while there is a bound on the number of queries that they make, we do not put any restrictions on their size, which allows us to hardwire the fixed $\mathcal{O}$ and f_{-n} as required in the previous three items. Indeed, Theorem 3.11 does not put any restriction on the size of these circuits.

COROLLARY 3.22. *Let $q(n) \leq O(2^{n/5})$. Then for any q-query distinguisher $\mathcal{D}$*

$$\mathop{\mathbb{E}}_{\mathcal{O}}\left[\mathsf{Adv}^{\mathsf{iO}}_{\Gamma,i\mathcal{O},\mathcal{C},\mathcal{D}}(n)\right] \leq O(2^{-n/10})?$$

The black-box reduction has to succeed for every oracle $(f,\mathcal{O})$ where the SZK-breaker oracle works (that is, all oracles). Here we show that for a random oracle, the adversary's positive distinguishing advantage is small. This suffices to prove the result.

At a very high level, the proof of the theorem follows a similar rationale to the proof of Theorem 3.11 showing that OWPs do not break in the presence of the noisy statistical difference oracle. Roughly speaking, we show that in order to break the above construction of IO, the adversary must be able to detect local changes in the oracles realizing it, whereas the noisy statistical difference oracle is insensitive of these changes. At a technical level, the case of IO requires somewhat more care than the case of OWPs. For once, it has a more elaborate interface consisting not only of a hard-to-invert mapping $\mathcal{O}$ but also of the evaluation oracle $\mathsf{Eval}^{f,\mathcal{O}}$. In particular, a single change to $\mathcal{O}$ may introduce many changes to $\mathsf{Eval}^{f,\mathcal{O}}$, which could potentially be detected by the statistical difference oracle. Another aspect that complicates the proof is that the IO game is more interactive in its nature. In particular, we need to deal with the fact that the actual circuits of the IO challenge are chosen adaptively, after the adversary had already interacted with all the oracles. We now turn to the actual proof.

Proof. We prove a stronger statement: the above holds when fixing the oracles f and $\mathcal{O}_{-n} = \{\mathcal{O}_k\}_{k\neq n}$. Fix a q-query adversary $\mathcal{D} = (\mathcal{D}_1, \mathcal{D}_2)$. To bound $\mathcal{D}$'s advantage in breaking $i\mathcal{O}$, we consider four hybrid experiments $\{\mathbf{H_i}\}_{i\in[4]}$ given in Table 3.2.

We introduce some notation that will be useful to describe the hybrids:

- We use regular expressions to describe sets, in particular, the $*$ expression. We denote by $(*,r) = \{(C,r) : C \in \{0,1\}^*$ and, $|C| = |r|\}$ and $\mathcal{O}(*,r) = \{\mathcal{O}(C,r) : C \in \{0,1\}^*$ and, $|C| = |r|\}$. In particular, define the set $T = (*,r)$.
- For a function $\mathcal{O} = \{\mathcal{O}_k : \{0,1\}^{2k} \to \{0,1\}^{5k}\}_{k\in\mathbb{N}}$, a set $T \subseteq \{0,1\}^{n\times 2}$, we denote by $\mathcal{O}_{T\mapsto\perp}$ the function that maps $(C,r) \in T$ to $\perp$ and is otherwise identical to $\mathcal{O}$. Hence, $\mathsf{Eval}(\mathcal{O}(C,r),x) = \perp$ for all $(C,r) \in T$. We often refer to such a set T as *the punctured set.*
- For a function $\mathcal{O} = \{\mathcal{O}_k : \{0,1\}^{2k} \to \{0,1\}^{5k}\}_{k\in\mathbb{N}}$, a pair $(C,r) \in \{0,1\}^{n\times 2}$, and $\widehat{C} \in \{0,1\}^{5n}$, we denote by $\mathcal{O}_{(C,r)\mapsto\widehat{C}}$ the function that maps (C,r) to $\widehat{C}$ and is otherwise identical to $\mathcal{O}$.

TABLE 3.2
The hybrid experiments.

Hybrid	$\mathbf{H_1}$ (Real)	$\mathbf{H_2}$	$\mathbf{H_3}$	$\mathbf{H_4}$ (Ideal)
Obfuscator function	$\mathcal{O}_n \leftarrow \mathbf{I}^{5n}_{2n}$			
Challenger randomness	$b \leftarrow \{0,1\}$, $r \leftarrow \{0,1\}^n$			
Planted obfuscation	$\widehat{C} \leftarrow \{0,1\}^{5n} \setminus \mathrm{Image}(\mathcal{O}_n)$			
Prechallenge punctured set T	$(*,r)$			
Prechallenge oracle	$\Gamma(f,\mathcal{O},\mathbf{t})$	$\Gamma(f,\mathcal{O}_{T\mapsto\perp},\mathbf{t})$		$\Gamma(f,\mathcal{O},\mathbf{t})$
Challenge obfuscation	$\mathcal{O}(C_b,r)$		$\widehat{C}$	
Postchallenge oracle	$\Gamma(f,\mathcal{O},\mathbf{t})$		$\Gamma(f,\mathcal{O}_{(C_b,r)\mapsto\widehat{C}},\mathbf{t})$	$\Gamma(f,\mathcal{O},\widehat{C},C_0,\mathbf{t})$
Winning condition	The adversary outputs b.			

- For a function $\mathcal{O} = \{\mathcal{O}_k : \{0,1\}^{2k} \to \{0,1\}^{5k}\}_{k\in\mathbb{N}}$, we denote by $\Gamma(f,\mathcal{O},\mathbf{t})$ the oracle
$$\Gamma(f,\mathcal{O},\mathbf{t}) := f, \mathcal{O}, \mathsf{Eval}^{f,\mathcal{O}}, \mathsf{StaDif}^{f,\mathcal{O},\mathsf{Eval}^{f,\mathcal{O}}},$$
where StaDif uses the threshold function $\mathbf{t}$.
- For a function $\mathcal{O} = \{\mathcal{O}_k : \{0,1\}^{2k} \to \{0,1\}^{5k}\}_{k\in\mathbb{N}}$, a string $\widehat{C} \in \{0,1\}^{5n}$, and a circuit C, we denote by $\Gamma(f,\mathcal{O},\widehat{C},C,\mathbf{t})$ the oracle
$$\Gamma(f,\mathcal{O},\widehat{C},C,\mathbf{t}) := f, \mathcal{O}, \mathsf{Eval}^{f,\mathcal{O}}_{\widehat{C},C}, \mathsf{StaDif}^{f,\mathcal{O},\mathsf{Eval}^{f,\mathcal{O}}_{\widehat{C},C}},$$
where StaDif uses the threshold function $\mathbf{t}$ and $\mathsf{Eval}^{f,\mathcal{O}}_{\widehat{C},C}$ is an oracle that
  - given $(\widehat{D},x)$ where $\widehat{D} \neq \widehat{C}$, acts like $\mathsf{Eval}^{f,\mathcal{O}}(\widehat{D},x)$, namely, it computes $(D,r) = \mathcal{O}^{-1}(\widehat{D})$ and returns $D(x)$, or $\perp$ in case there is no unique preimage or the size of x does not match the input size of D;
  - given $(\widehat{C},x)$ returns $C(x)$, or $\perp$ in case $C = \perp$, or the size of x does not match the input size of C.
- Throughout all hybrids $\mathbf{t} = \{\mathbf{t}_k\}_{k\in\mathbb{N}}$ where $\mathbf{t}_k : \{0,1\}^{2k} \to \left(\frac{1}{3},\frac{2}{3}\right)$ is a random function.

CLAIM 3.23. $|\Pr[\mathcal{D} \textit{ wins in } \mathbf{H_1}] - \Pr[\mathcal{D} \textit{ wins in } \mathbf{H_2}]| \le O(2^{-n/5})$.

Proof. The difference between the two hybrids is in the oracle that $\mathcal{D}_1$ is given before the challenge phase: $\Gamma(f,\mathcal{O},\mathbf{t})$ in the first, and its tweaked version $\Gamma(f,\mathcal{O}_{T\mapsto\perp},\mathbf{t})$ in the second.

We can bound the difference between the success probability in $\mathbf{H_1}$ and $\mathbf{H_2}$ as follows:
$$|\Pr[\mathcal{D} \text{ wins in } \mathbf{H_1}] - \Pr[\mathcal{D} \text{ wins in } \mathbf{H_2}]| \le \mathop{\mathbb{E}}_{\mathcal{O}}\left[\Pr_{\substack{\mathcal{D}_1,\mathbf{t},\\ b,r}}\left[\mathcal{D}_1^{\Gamma(f,\mathcal{O},\mathbf{t})}(1^n) \neq \mathcal{D}_1^{\Gamma(f,\mathcal{O}_{T\mapsto\perp},\mathbf{t})}(1^n)\right]\right],$$
where $\mathcal{D}_1$ is the part of $\mathcal{D} = (\mathcal{D}_1,\mathcal{D}_2)$ that participates in the prechallenge phase, and the probability is over the coins of $\mathcal{D}_1$ and $\mathbf{t}$ (used by StaDif) and the choice of $r \leftarrow \{0,1\}^n$, and $\mathcal{O} \leftarrow \mathbf{I}^{5n}_{2n}$, and $b \leftarrow \{0,1\}$. We will, in fact, show that the above is bounded for any fixed $b \in \{0,1\}$. Indeed, for the rest of the claim, fix $b \in \{0,1\}$.

In what follows, let $\mathbf{Far} = \mathbf{Far}(\mathcal{D}_1,\mathcal{O},2\delta)$ be the event that 2δ-farness holds for all StaDif-queries made by $\mathcal{D}_1^{\Gamma(f,\mathcal{O},\mathbf{t})}(1^n)$ (Definition 3.6). Let $\mathbf{rHit}$ denote the event that $\mathcal{D}_1$ queries Γ on $T \cup (\mathcal{O}(T),*) = (*,r) \cup (\mathcal{O}(*,r),*)$, that is, queries $\mathcal{O}$ at (C,r) for some C or queries Eval on $(\mathcal{O}(C,r),z)$ for some C and z. Let $\mathbf{Smo}(\mathcal{D}_1,\Gamma,r,\delta)$ denote the event that all queries $\mathcal{D}_1$ makes are $(\Psi, T \cup (\mathcal{O}(T),*),\delta)$-smooth (see Definition 3.8).

We now claim as follows.

CLAIM 3.24. *For any fixed $f,\mathcal{O}$,*
$$\Pr_{\mathcal{D}_1,r,\mathbf{t}}\left[\mathcal{D}_1^{\Gamma(f,\mathcal{O},\mathbf{t})}(1^n) \neq \mathcal{D}_1^{\Gamma(f,\mathcal{O}_{T\mapsto\perp},\mathbf{t})}(1^n)\right] \le \Pr_{\mathcal{D}_1,r,\mathbf{t}}\left[\mathbf{rHit} \vee \overline{\mathbf{Far}} \vee \overline{\mathbf{Smo}}\right].$$

Proof. We argue that whenever the complement $\overline{\mathbf{rHit}} \wedge \mathbf{Far} \wedge \mathbf{Smo}$ occurs then
$$\mathcal{D}_1^{\Gamma(f,\mathcal{O},\mathbf{t})}(1^n) = \mathcal{D}_1^{\Gamma(f,\mathcal{O}_{T\mapsto\perp},\mathbf{t})}(1^n).$$
The two oracles $\Psi = (f,\mathcal{O},\mathsf{Eval}^{f,\mathcal{O}})$ and $\Psi' = (f,\mathcal{O}_{T\mapsto\perp},\mathsf{Eval}^{f,\mathcal{O}_{T\mapsto\perp}})$ behave identically on all queries outside $T \cup (\mathcal{O}(T),*)$. Hence, when the event $\mathbf{rHit}$ does not occur, the two oracles answer all queries identically.

Next, we need to show that all StaDif queries are answered identically by StaDif^{Ψ} and $\mathsf{StaDif}^{\Psi'}$. For any StaDif-query (D_0, D_1) made by $\mathcal{D}^{\Gamma(f,\mathcal{O},\mathbf{t})}(1^n)$, we know by $\mathbf{Smo}(\Psi, \delta, r)$ that changing Ψ on the set T does not affect the statistical distance by much. Concretely,

$$\left|\Delta(\boldsymbol{D}_0^{\Psi}, \boldsymbol{D}_1^{\Psi}) - \Delta(\boldsymbol{D}_0^{\Psi'}, \boldsymbol{D}_1^{\Psi'})\right| \leq 2\delta.$$

Furthermore, if 2δ-farness also holds for any such query (for some threshold $\mathbf{t}$ sampled by StaDif), then

$$\mathsf{StaDif}^{\Psi}(D_0, D_1; \mathbf{t}) = \mathsf{StaDif}^{\Psi'}(D_0, D_1; \mathbf{t}).$$

Hence, it follows that the views of $\mathcal{D}_1^{\Gamma(f,\mathcal{O},\mathbf{t})}(1^n)$ and $\mathcal{D}_1^{\Gamma(f,\mathcal{O}_{T\mapsto\perp},\mathbf{t})}(1^n)$ are identical. ∎

It is left to bound the probability of each of the events $\mathbf{rHit}, \overline{\mathbf{Far}}, \overline{\mathbf{Smo}}$. First, by noting that the view of $\mathcal{D}_1^{\Gamma(f,\mathcal{O},\mathbf{t})}(1^n))$ is independent of the random r and that $\mathcal{O}$ is injective, we can bound

$$\Pr[\mathbf{rHit}] \leq 2^{-n} \cdot \#\{\Psi\text{-queries made by } \mathcal{D}_1\} \leq 2^{-n} \cdot q.$$

Furthermore by the farness claim, Claim 3.7, we get that

$$\Pr\left[\overline{\mathbf{Far}}\right] \leq 12q\delta.$$

We now bound the probability that $\mathbf{Smo}$ does not occur: For any fixed $f, \mathcal{O}$,

$$\Pr_r\left[\overline{\mathbf{Smo}(\mathcal{D}_1, \Gamma, r, \delta)}\right] \leq 2^{-n} \cdot 2q^2/\delta\ .$$

The proof again follows from the fact that any circuit D that makes at most q queries to $(f, \mathcal{O}, \mathsf{Eval}^{f,\mathcal{O}})$ and the fact that $\mathcal{O}$ is injective, hence there are at most q/δ values of r that do not satisfy smoothness, that is, the circuit D queries $(*, r)$ or the corresponding outputs $(\mathcal{O}(C, r), *)$ with probability more than δ. Using the fact that r is independent of the execution in $\mathbf{H_1}$, the probability that it is one of these values is at most $2^{-n} \cdot q/\delta$. A union bound over q queries of the form (D_0, D_1) gives the required bound.

Overall, we can bound the difference between $\mathbf{H_1}$ and $\mathbf{H_2}$ by

$$2^{-n} \cdot q + 12q\delta + 2^{-n} \cdot 2q^2/\delta \leq O(2^{-n/5})$$

when setting $\delta = 2^{-2n/5}$ and recalling that $q \leq O(2^{n/5})$. ∎

CLAIM 3.25. $\Pr[\mathcal{D}\ \textit{wins in}\ \mathbf{H_2}] = \Pr[\mathcal{D}\ \textit{wins in}\ \mathbf{H_3}]$.

Proof. The difference between $\mathbf{H_2}$ and $\mathbf{H_3}$ is that in $\mathbf{H_3}$, in the challenge and postchallenge phases, the value $\mathcal{O}(C_b, r)$ is resampled uniformly at random from the co-image, namely, it is replaced everywhere by $\widehat{C} \leftarrow \{0,1\}^{5n} \setminus \mathrm{Image}(\mathcal{O}_n)$. We claim that this induces exactly the same distribution on $\mathcal{D}$'s view as in $\mathbf{H_2}$. Indeed, in $\mathbf{H_2}$, at the end of the prechallenge phase, fixing the view of $\mathcal{D}$, the distribution of $\mathcal{O}(C_b, r)$ is uniformly random in $S := \{0,1\}^{5n} \setminus \mathrm{Image}(\mathcal{O}_{(C_b,r)\mapsto\perp})$. In $\mathbf{H_2}$ $\mathcal{O}(C_b, r)$ is sampled uniformly at random directly from S, whereas in $\mathbf{H_3}$, we first sample a random value $\mathcal{O}(C_b, r)$ from S, and then resample $\widehat{C}$ from $S \setminus \{\mathcal{O}(C_b, r)\}$, which again gives a uniformly random value in S. ∎

CLAIM 3.26. $|\Pr[\mathcal{D}\ \textit{wins in}\ \mathbf{H_3}] - \Pr[\mathcal{D}\ \textit{wins in}\ \mathbf{H_4}]| \leq O(2^{-n/5})$.

Proof. There are two differences between the hybrids. The first is in the oracle that $\mathcal{D}_1$ is given before the challenge phase: $\Gamma(f,\mathcal{O},\mathbf{t})$ in $\mathbf{H_4}$, and its tweaked version $\Gamma(f,\mathcal{O}_{T\mapsto\perp},\mathbf{t})$ in $\mathbf{H_3}$. The second is in the oracle that $\mathcal{D}_2$ is given after the challenge phase: $\Gamma(f,\mathcal{O},\widehat{C},C_0,\mathbf{t})$ in $\mathbf{H_4}$, and $\Gamma(f,\mathcal{O}_{(C_b,r)\mapsto\widehat{C}},\mathbf{t}) = \Gamma(f,\mathcal{O}_{(C_b,r)\mapsto\widehat{C}},\widehat{C},C_b,\mathbf{t})$ in $\mathbf{H_3}$. We can thus bound the difference between the winning probabilities in $\mathbf{H_3}$ and $\mathbf{H_4}$ as follows:

$$
\begin{aligned}
&|\Pr[\mathcal{D} \text{ wins in } \mathbf{H_3}] - \Pr[\mathcal{D} \text{ wins in } \mathbf{H_4}]| \\
&\le \mathop{\mathbb{E}}_{\mathcal{O}}\left[\Pr_{\mathcal{D}_1,b,\mathbf{t}}\left[\mathsf{state} := \mathcal{D}_1^{\Gamma(f,\mathcal{O},\mathbf{t})}(1^n) \neq \mathcal{D}_1^{\Gamma(f,\mathcal{O}_{T\mapsto\perp},\mathbf{t})}(1^n)\right]\right] \\
&+ \mathop{\mathbb{E}}_{\mathcal{O}}\left[\Pr_{\mathcal{D}_1,b,\mathbf{t}}\left[\mathcal{D}_2^{\Gamma(f,\mathcal{O},\widehat{C},C_0,\mathbf{t})}(\mathsf{state},\widehat{C}) \neq \mathcal{D}_2^{\Gamma(f,\mathcal{O}_{(C_b,r)\mapsto\widehat{C}},\widehat{C},C_b,\mathbf{t})}(\mathsf{state},\widehat{C}) \,\middle|\, \mathsf{state} = \mathcal{D}_1^{\Gamma(f,\mathcal{O},\mathbf{t})}(1^n)\right]\right].
\end{aligned}
$$

As proved in Claim 3.23, the first summand is bounded by $O(2^{-n/5})$. This proof is unchanged here. We argue that a similar bound holds for the second summand as well.

CLAIM 3.27.

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{O}}\left[\Pr_{\mathcal{D}_1,b,\mathbf{t}}\left[\mathcal{D}_2^{\Gamma(f,\mathcal{O},\widehat{C},C_0,\mathbf{t})}(\mathsf{state},\widehat{C}) \neq \mathcal{D}_2^{\Gamma(f,\mathcal{O}_{(C_b,r)\mapsto\widehat{C}},\widehat{C},C_b,\mathbf{t})}(\mathsf{state},\widehat{C}) \,\middle|\, \mathsf{state} = \mathcal{D}_1^{\Gamma(f,\mathcal{O},\mathbf{t})}(1^n)\right]\right] \\
&\quad \le O(2^{-n/5}).
\end{aligned}
$$

Proof. The proof is similar to that of Claim 3.23 and is included here for completeness. Note that the postchallenge oracles $(f,\mathcal{O},\mathsf{Eval})$ are different in the following two ways: a query to $\mathcal{O}$ at (C_b,r) would output $\widehat{C}$ in $\mathbf{H_3}$ and $\mathcal{O}(C_b,r)$ in $\mathbf{H_4}$. Also, a query to Eval at $(\mathcal{O}(C_b,r),*)$ would output $\perp$ in $\mathbf{H_3}$. Note that a query of the form $(\widehat{C},*)$ would be answered identically in both $\mathbf{H_3}$ and $\mathbf{H_4}$ because of the functional equivalence of C_0^f and C_1^f and hence whether the Eval oracle answers using C_0 in $\mathbf{H_4}$ or C_b in $\mathbf{H_3}$, the answers would be identical.

In what follows, let $\mathbf{Far} = \mathbf{Far}(\mathcal{D}_2,\mathcal{O},2\delta)$ be the event that 2δ-farness holds for all StaDif-queries made by $\mathcal{D}_2^{\Gamma(f,\mathcal{O},\mathbf{t})}(1^n,\mathsf{state})$ (Definition 3.6). Let $\mathbf{rHit}$ denote the event that $\mathcal{D}_2$ queries Γ on T, i.e., queries $\mathcal{O}$ at (C,r) for some C or queries Eval on $(\mathcal{O}(C,r),z)$ for some C and z. Let $\mathbf{Smo}(\mathcal{D}_2,\Gamma,r,\delta)$ denote the event that all queries $\mathcal{D}_2$ makes are $(\Psi, T\cup(\mathcal{O}(T),*),\delta)$-smooth (see Definition 3.8) where $T=(*,r)$.

We now claim as follows.

CLAIM 3.28. *For any fixed $f,\mathcal{O}$,*

$$
\begin{aligned}
&\Pr_{\mathcal{D}_2,b,\mathbf{t}}\left[\mathcal{D}_2^{\Gamma(f,\mathcal{O},\widehat{C},C_0,\mathbf{t})}(\mathsf{state},\widehat{C}) \neq \mathcal{D}_2^{\Gamma(f,\mathcal{O}_{(C_b,r)\mapsto\widehat{C}},\widehat{C},C_b,\mathbf{t})}(\mathsf{state},\widehat{C}) \,\middle|\, \mathsf{state} = \mathcal{D}_1^{\Gamma(f,\mathcal{O},\mathbf{t})}(1^n)\right] \\
&\le \Pr_{\mathcal{D}_2,r,\mathbf{t}}\left[\mathbf{rHit} \vee \overline{\mathbf{Far}} \vee \overline{\mathbf{Smo}}\right].
\end{aligned}
$$

Proof. We argue that whenever the complement $\overline{\mathbf{rHit}} \wedge \mathbf{Far} \wedge \mathbf{Smo}$ occurs then

$$
\mathcal{D}_2^{\Gamma(f,\mathcal{O},\widehat{C},C_0,\mathbf{t})}(\mathsf{state},\widehat{C}) \neq \mathcal{D}_2^{\Gamma(f,\mathcal{O}_{(C_b,r)\mapsto\widehat{C}},\widehat{C},C_b,\mathbf{t})}(\mathsf{state},\widehat{C}).
$$

As noted above, the oracles $\Psi=(f,\mathcal{O},\mathsf{Eval}^{f,\mathcal{O}}_{\widehat{C},C})$ and $\Psi'=(f,\mathcal{O}_{(C_b,r)\mapsto\widehat{C}},\mathsf{Eval}^{f,\mathcal{O}_{(C_b,r)\mapsto\widehat{C}}})$ behave identically on all queries outside $T\cup(\mathcal{O}(T),*)$. Hence, when the event $\mathbf{rHit}$ does not occur, the two oracles answer all queries identically.

Next, we need to show that all StaDif queries are answered identically by StaDif^{Ψ} and $\mathsf{StaDif}^{\Psi'}$. For any StaDif-query (D_0,D_1) made by $\mathcal{D}_2^{\Gamma(f,\mathcal{O},\widehat{C},C_0,\mathbf{t})}(1^n,\mathsf{state})$, we

know by $\mathbf{Smo}(\Psi, \delta, r)$ that changing Ψ on the set T does not affect the statistical distance by much. Concretely,

$$\left|\Delta(\boldsymbol{D}_0^{\Psi}, \boldsymbol{D}_1^{\Psi}) - \Delta(\boldsymbol{D}_0^{\Psi'}, \boldsymbol{D}_1^{\Psi'})\right| \leq 2\delta.$$

Furthermore, if 2δ-farness also holds for any such query (for some threshold $\mathbf{t}$ sampled by StaDif), then

$$\mathsf{StaDif}^{\Psi}(D_0, D_1; \mathbf{t}) = \mathsf{StaDif}^{\Psi'}(D_0, D_1; \mathbf{t}).$$

Hence, it follows that the views of $\mathcal{D}_2^{\Gamma(f,\mathcal{O},\widehat{C},C_0,\mathbf{t})}(1^n, \mathsf{state})$ and $\mathcal{D}_2^{\Gamma(f,\mathcal{O}_{(C_b,r)\mapsto\widehat{C}},\mathbf{t})}(1^n, \mathsf{state})$ are identical. ∎

It is left to bound the probability of each of the events $\mathbf{rHit}, \overline{\mathbf{Far}}, \overline{\mathbf{Smo}}$. First, by noting that the view of $\mathcal{D}_2^{\Gamma(f,\mathcal{O},\widehat{C},C_0,\mathbf{t})}(1^n, \mathsf{state}))$ is independent of the random r and that $\mathcal{O}$ is injective, we can bound

$$\Pr[\mathbf{rHit}] = 2^{-n} \cdot \#\{\Psi\text{-queries made by } \mathcal{D}_2\} \leq 2^{-n} \cdot q.$$

Furthermore by the farness claim, Claim 3.7, we get that

$$\Pr\left[\overline{\mathbf{Far}}\right] \leq 12q\delta.$$

As before, we can bound the probability that $\mathbf{Smo}$ does not occur: For any fixed $f, \mathcal{O}$,

$$\Pr_r\left[\overline{\mathbf{Smo}(\mathcal{D}_2, \Gamma, r, \delta)}\right] \leq 2^{-n} \cdot 2q^2/\delta\ .$$

The proof is identical to that of Claim 3.14. Hence, we can bound the difference between $\mathbf{H_1}$ and $\mathbf{H_2}$ by

$$2^{-n} \cdot q + 12q\delta + 2^{-n} \cdot 2q^2/\delta \leq O(2^{-n/5})$$

when setting $\delta = 2^{-2n/5}$ and recalling that $q \leq O(2^{n/5})$. ∎

This completes the proof as both the terms are bounded by $O(2^{-n/5})$. ∎

Claim 3.29. *For any oracle $f, \mathcal{O}$, $\Pr[\mathcal{D}$ wins in $\mathbf{H_4}] = \frac{1}{2}$.*

Proof. The view of $\mathcal{D}$ in this hybrid is completely independent of the random choice of b and r. ∎

This concludes the proof of Theorem 3.21. ∎

4. One-way functions, indistinguishability obfuscation, and hardness in $\mathsf{NP} \cap \mathsf{coNP}$. In this section, we show that IOWFs and IO, for circuits with IOWF gates, do not give rise to a black-box construction of hard problems in $\mathsf{NP} \cap \mathsf{coNP}$. This can be seen as a generalization of previous separations by Rudich [Rud88], showing that OWFs do not give hardness in $\mathsf{NP} \cap \mathsf{coNP}$, by Matsuda and Matsuura [MM11], showing that IOWFs do not give OWPs (which are a special case of hardness $\mathsf{NP} \cap \mathsf{coNP}$), and by Asharov and Segev [AS16], showing that OWFs and IO do not give OWPs. As in the previous section, the result implies that many cryptographic primitives and hardness in PPAD also do not yield black-box constructions of hard problems in $\mathsf{NP} \cap \mathsf{coNP}$.

We first define the framework of $\mathsf{NP} \cap \mathsf{coNP}$ relative to oracles, define fully black-box constructions of hard $\mathsf{NP} \cap \mathsf{coNP}$ problems, and then move on to the actual separation.

4.1. $\mathsf{NP} \cap \mathsf{coNP}$. Throughout, we shall canonically represent languages $L \in \mathsf{NP} \cap \mathsf{coNP}$ by their corresponding nondeterministic polytime verifiers V_1, V_0, where

$$L = \left\{ x \in \{0,1\}^\star \;\middle|\; \exists w : V_1(x,w) = 1 \right\},$$
$$\overline{L} = \left\{ x \in \{0,1\}^\star \;\middle|\; \exists w : V_0(x,w) = 1 \right\} = \{0,1\}^\star \setminus L.$$

Hardness in $\mathsf{NP} \cap \mathsf{coNP}$ from cryptography: Motivation. Hard (on average) problems in $\mathsf{NP} \cap \mathsf{coNP}$ are known to follow based on certain number-theoretic problems in cryptography, such as discrete log and factoring. As in the previous section for SZK, we are interested in understanding which cryptographic primitives would imply such hardness, again with the intuition that these should be appropriately structured. For instance, it is known [Bra79] that any OWP $f : \{0,1\}^n \to \{0,1\}^n$ implies a hard problem in $\mathsf{NP} \cap \mathsf{coNP}$, e.g., given an index $i \in [n]$ and an image $f(x)$ find the ith preimage bit x_i. In contrast, in his seminal work, Rudich [Rud88] proved that completely unstructured objects like OWFs cannot construct even worst-case hard instances by fully black-box constructions. Here a fully black-box construction essentially means that the nondeterministic verifiers only make black-box use of the OWF (or OWP in the previous example) and the reduction establishing the hardness is also black-box (in both the adversary and the OWF).

But what about more structured primitives such as public-key encryption, oblivious transfer, or even IO? Indeed, IO (plus OWFs) has-been shown to imply hardness in PPAD and more generally in the class TFNP of total search problems, which is often viewed as the search analogue of $\mathsf{NP} \cap \mathsf{coNP}$ [MP91]. We will show, however, that fully black-box constructions do not give rise to a hard problem in $\mathsf{NP} \cap \mathsf{coNP}$ from OWFs (or even injective OWFs) and IO for circuits with OWF gates.

4.2. Fully black-box constructions of hardness in $\mathsf{NP} \cap \mathsf{coNP}$ from IO and IOWFs. We start by defining $\mathsf{NP} \cap \mathsf{coNP}$ relative to oracles [Rud88]. This, in particular, captures black-box constructions of such languages from cryptographic primitives, such as OWFs in [Rud88] or IO, which we will consider in this work.

Definition 4.1 ($\mathsf{NP} \cap \mathsf{coNP}$ relative to oracles). *Let $\mathfrak{S}$ be a family of oracles and let $V_1^{(\cdot)}, V_0^{(\cdot)}$ be a pair of oracle-aided nondeterministic polynomial-time verifiers. We say that V_1, V_0 define a collection of languages $L^{\mathfrak{S}} = \left\{ L^\Gamma \;\middle|\; \Gamma \in \mathfrak{S} \right\}$ in $\mathsf{NP} \cap \mathsf{coNP}$ relative to $\mathfrak{S}$ if for any $\Gamma \in \mathfrak{S}$, the machines V_1^Γ, V_0^Γ define a language $L^\Gamma \in \mathsf{NP}^\Gamma \cap \mathsf{coNP}^\Gamma$. That is,*

$$L^\Gamma = \left\{ x \in \{0,1\}^\star \;\middle|\; \exists w : V_1^\Gamma(x,w) = 1 \right\},$$
$$\overline{L}^\Gamma = \left\{ x \in \{0,1\}^\star \;\middle|\; \exists w : V_0^\Gamma(x,w) = 1 \right\} = \{0,1\}^\star \setminus L.$$

We now formally define the class of constructions and reductions ruled out, that is, *fully black-box* constructions of hard problems in $\mathsf{NP} \cap \mathsf{coNP}$ from IOWFs and IO for IOWF-aided circuits. The definition is similar in spirit to the definitions in [AS15, AS16] and in section 3, adapted to the context of $\mathsf{NP} \cap \mathsf{coNP}$ hardness.

Definition 4.2. *A fully black-box construction of a hard $\mathsf{NP} \cap \mathsf{coNP}$ problem L from IOWFs and IO for the class $\mathcal{C}$ of circuits with IOWF gates is given by two oracle-aided polytime machines (V_0, V_1) and a probabilistic oracle-aided reduction $\mathcal{R}$ that satisfy the following:*

1. Structure: *Let $\mathfrak{S}$ be the family of all oracles $(f, i\mathcal{O})$ such that f is injective and $i\mathcal{O}$ is a function such that $\widehat{C}^f \equiv C^f$ for any $C^{(\cdot)} \in \mathcal{C}$, r, and $\widehat{C}^{(\cdot)} := i\mathcal{O}(C, r)$.*

Then (V_0, V_1) *define a language* $L^{f,i\mathcal{O}} \in \mathsf{NP}^{f,i\mathcal{O}} \cap \mathsf{coNP}^{f,i\mathcal{O}}$ *relative to any oracle* $(f, i\mathcal{O}) \in \mathfrak{S}$ *(as per Definition* 4.1*).*

2. Black-box security proof: *There exist functions* $q_{\mathcal{R}}(\cdot), \varepsilon_{\mathcal{R}}(\cdot)$ *such that the following holds. Let* $(f, i\mathcal{O})$ *be any distribution supported on the family* $\mathfrak{S}$ *defined above. Then for any probabilistic oracle-aided* $\mathcal{A}$ *that* decides $L^{f,i\mathcal{O}}$ in the worst case*, namely, for all* $n \in \mathbb{N}$

$$\Pr_{f,i\mathcal{O},\mathcal{A}}\left[\mathcal{A}^{f,i\mathcal{O}}(x) = b \quad \text{for all} \quad \begin{array}{c} x \in \{0,1\}^n, b \in \{0,1\} \\ \text{such that } V_b(x) = 1 \end{array}\right] = 1,$$

the reduction breaks either f or $i\mathcal{O}$*, namely, for infinitely many* $n \in \mathbb{N}$ *either*

$$\Pr_{\substack{x \leftarrow \{0,1\}^n \\ f,i\mathcal{O},\mathcal{A}}}\left[\mathcal{R}^{\mathcal{A},f,i\mathcal{O}}(f(x)) = x\right] \geq \varepsilon_{\mathcal{R}}(n)$$

or

$$\mathsf{Adv}^{\mathsf{iO}}_{(f,i\mathcal{O}),i\mathcal{O},\mathcal{C},\mathcal{R}^{\mathcal{A}}}(n) \geq \varepsilon_{\mathcal{R}}(n),$$

where in both $\mathcal{R}$ *makes at most* $q_{\mathcal{R}}(n)$ *queries to any of its oracles* $(\mathcal{A}, f, i\mathcal{O})$*, and for any query* x *made to* $\mathcal{A}$*, the nondeterministic verifiers* $V_0^{f,i\mathcal{O}}(x)$*,* $V_1^{f,i\mathcal{O}}(x)$ *make at most* $q_{\mathcal{R}}(n)$ *queries to their oracles (for any nondeterministic choice of a witness* w*). The random variable* $\mathsf{Adv}^{\mathsf{iO}}_{(f,i\mathcal{O}),i\mathcal{O},\mathcal{C},\mathcal{R}^{\mathcal{A}}}(n)$ *represents the reductions winning probability in the IO security game (Definition* 2.2*) relative to* $(f, i\mathcal{O})$*.*

Remark about correct structure. We note that here we explicitly do put a *correctness* requirement, which we refer to as *structure*, namely, that the construction yields a language in $\mathsf{NP} \cap \mathsf{coNP}$ for any implementation of OWPs and IO. This is different from the setting from Definition 3.2 where we considered *promise problems* and allowed the construction not to satisfy the promise occasionally.

Concretely, we require that V_0, V_1 give a language in $\mathsf{NP} \cap \mathsf{coNP}$ for *every* oracle implementing IOWFs and IO. This follows the modeling of [BI87][12] and aligns with how we usually think about *correctness* of black-box constructions of cryptographic primitives. For instance, the construction of public-key encryption from trapdoor permutations is promised to be correct, for all oracles implementing the trapdoor permutation. Similarly, the construction of hard $\mathsf{NP} \cap \mathsf{coNP}$ languages from OWPs gives an $\mathsf{NP} \cap \mathsf{coNP}$ language for any oracle implementing a permutation.

We also note that as in section 3, our definition addresses *worst-case hardness*, which makes our impossibility result stronger. See further discussion after Definition 3.2.

Ruling out fully black-box constructions: A road map. Our main result in this section is that fully black-box constructions of a hard $\mathsf{NP} \cap \mathsf{coNP}$ problem from IO and IOWFs do not exist. Furthermore, this holds even if the latter primitives are exponentially secure.

THEOREM 4.3. *Any fully black-box construction of an* $\mathsf{NP} \cap \mathsf{coNP}$ *problem* L *from IOWFs and IO for circuits with IOWF gates has an exponential security loss:* $\max(q_{\mathcal{R}}(n), \varepsilon_{\mathcal{R}}^{-1}(n)) \geq \Omega(2^{n/6})$.

[12] Rudich [Rud88] also considered a slight relaxation of constructions that are correct for an overwhelming fraction of oracles rather than all.

The proof of the theorem follows a methodology similar to that in section 3. We exhibit two (distributions on) oracles $(\Psi, \mathsf{Decide}^{\Psi})$, where Ψ realizes IOWFs and IO for circuits with IOWF gates, and Decide^{Ψ} that decides $L^{\Psi} \in \mathsf{NP}^{\Psi} \cap \mathsf{coNP}^{\Psi}$ in the worst case. We then show that the primitives realized by Ψ are (exponentially) secure even in the presence of StaDif^{Ψ}. This statement is proved when the oracle Ψ is sampled at random from the constructed distribution. Since Decide^{Ψ} decides $\mathsf{NP}^{\Psi} \cap \mathsf{coNP}^{\Psi}$, in the worst case, for every oracle Ψ, for every given reduction, there is a fixed oracle $\Gamma = (\Psi, \mathsf{Decide}^{\Psi})$ relative to which (1) $\mathsf{NP} \cap \mathsf{coNP}$ is easy to decide, (2) the reduction fails to break IO (or IOWFs). This implies Theorem 4.3, ruling out fully black-box constructions with a subexponential security loss.

The rest of this section is organized according to the above plan. First, in subsection 4.3, we describe the oracle Decide^{Ψ}. As a warm-up, in subsection 4.4 we show that IOWFs cannot construct hard languages in $\mathsf{NP} \cap \mathsf{coNP}$ in a black-box manner. Then in subsection 4.5, we describe the oracle Ψ such that even in the presence of Decide^{Ψ}, (exponentially) secure OWFs and IO exist. This rules out *fully black-box* constructions of even *worst-case* hard problems in $\mathsf{NP} \cap \mathsf{coNP}$.

4.3. The decision oracle. In this section, we construct an oracle $\mathsf{Decide}_{\mathfrak{S}}$ that is defined with respect to a family $\mathfrak{S}$ of oracles (e.g., all oracles implementing IOWF and IO) and which given access to $\Psi \in \mathfrak{S}$ decides any language in $\mathsf{NP}^{\Psi} \cap \mathsf{coNP}^{\Psi}$.

DEFINITION 4.4 (oracle $\mathsf{Decide}^{\Psi}_{\mathfrak{S}}$). *For a family of oracles $\mathfrak{S}$, we define the* $\mathsf{Decide}_{\mathfrak{S}}$ *oracle as follows:*

- $\mathsf{Decide}_{\mathfrak{S}}$ *is given oracle access to some* Ψ.
- Decide *takes as input a pair of oracle-aided circuits* (V_0, V_1) *along with an input* z *where the circuits* V_0, V_1 *(allegedly) define a language in* $\mathsf{NP} \cap \mathsf{coNP}$ *relative to* $\mathfrak{S}$.
- $\mathsf{Decide}^{\Psi}_{\mathfrak{S}}(V_0, V_1, z)$ *does the following:*
  1. *It checks that* $V_0^{\Psi'}, V_1^{\Psi'} \in \mathsf{NP}^{\Psi'} \cap \mathsf{coNP}^{\Psi'}$ *for all* $\Psi' \in \mathfrak{S}$. *If not, output* $\perp$.
  2. *For the input* z, *it outputs the unique* b *such that there exists a witness* w *satisfying* $V_b^{\Psi}(z, w) = 1$. *(Since* V_0^{Ψ}, V_1^{Ψ} *define an* $\mathsf{NP} \cap \mathsf{coNP}$ *language such* b *indeed exists and is unique.)*

We make a few remarks about the $\mathsf{Decide}_{\mathfrak{S}}$ oracle.

1. We will use the $\mathsf{Decide}_{\mathfrak{S}}$ oracle in a similar way to the StaDif oracle in section 3. We will be interested in the family of oracles $\mathfrak{S}$ that implements a required primitive $\mathcal{P}$ (eventually IOWFs and IO). We will show a distribution Ψ supported on $\mathfrak{S}$ that securely implements $\mathcal{P}$ in the presence of $\mathsf{Decide}^{\Psi}_{\mathfrak{S}}$, whereas at the same time, $\mathsf{Decide}^{\Psi}_{\mathfrak{S}}$ will enable us to decide any language in $\mathsf{NP}^{\Psi} \cap \mathsf{coNP}^{\Psi}$ given by verifiers that define n $\mathsf{NP} \cap \mathsf{coNP}$ language relative to any oracle in $\mathfrak{S}$.
2. Queries to the oracle are represented as circuit verifiers V_0, V_1. We will consider adversaries that only produce V_0, V_1 that make some bounded number of oracle queries to Ψ.
3. The behavior of the oracle $\mathsf{Decide}^{\Psi}_{\mathfrak{S}}$ is undefined for oracles Ψ outside $\mathfrak{S}$. In our analysis, all oracles considered will be taken from the family $\mathfrak{S}$.

To rule out fully black-box constructions of hard languages in $\mathsf{NP} \cap \mathsf{coNP}$ we have to show two things. First, $\mathsf{Decide}^{\Psi}_{\mathfrak{S}}$ is sufficient to decide any $\mathsf{NP}^{\Psi} \cap \mathsf{coNP}^{\Psi}$ language given by verifiers that define an $\mathsf{NP} \cap \mathsf{coNP}$ language relative to any oracle in $\mathfrak{S}$. Second, it is not helpful in breaking IOWFs and IO.

The first part follows directly from the definition of this oracle.

CLAIM 4.5. *Let $\mathfrak{S}$ be any family and let (V_0, V_1) be any pair of polynomial-time verifiers that define a collection $L^{\mathfrak{S}} = \{L^{\Psi}\}_{\Psi\in\mathfrak{S}}$ in $\mathsf{NP}\cap\mathsf{coNP}$; then for any oracle $\Psi \in \mathfrak{S}$,*

$$L^{\Psi} \in \mathsf{P}^{\Psi,\mathsf{Decide}^{\Psi}_{\mathfrak{S}}}.$$

The second part is the more challenging one. Our proof strategy is somewhat inspired by the proof of Theorem 3.19 for the case of SZK. Roughly speaking, we will aim to show that the oracle $\mathsf{Decide}^{\Psi}_{\mathfrak{S}}$ is in some sense insensitive to *random local changes*, whereas breaking the latter cryptographic primitives does require the ability to detect such changes.

Toward fulfilling this proof strategy, we now prove a general claim that roughly says that the answers of $\mathsf{Decide}^{\Psi}_{\mathfrak{S}}$ to any specific query are always determined by the behavior of Ψ on a relatively small "critical" set. Intuitively, this means that random changes that "evade" this critical set will go undetected by the oracle.

In what follows, we call a verifier circuit $V : \{0,1\}^n \times \{0,1\}^m \to \{0,1\}$ q-query if for any $z \in \{0,1\}^n$, and any potential witness $w \in \{0,1\}^m$, the circuit $V^{\Psi}(z,w)$ makes at most q queries to Ψ. Similarly, we call a query (V_0, V_1, z) to the oracle $\mathsf{Decide}_{\mathfrak{S}}$ q-bounded if both the verifiers V_0 and V_1 are q-query verifiers.

CLAIM 4.6. *Let $\mathfrak{S}$ be any family of oracles. Consider an oracle Ψ from $\mathfrak{S}$. Consider any q-bounded query (V_0, V_1, z) to $\mathsf{Decide}^{\Psi}_{\mathfrak{S}}$. Then there exists a set of queries $\mathbf{C} = \mathbf{C}(\Psi, V_0, V_1, z)$, which we call* a critical set*, such that the following hold:*

1. *The critical set $\mathbf{C}$ is small: $|\mathbf{C}| \le q$.*
2. *Consider another oracle $\Psi' \in \mathfrak{S}$. If the two oracles agree on the set $\mathbf{C}$, then the corresponding $\mathsf{Decide}_{\mathfrak{S}}$ oracles also agree. That is, for every $\Psi' \in \mathfrak{S}$ such that $\Psi|_{\mathbf{C}} = \Psi'|_{\mathbf{C}}$,*

$$\mathsf{Decide}^{\Psi}_{\mathfrak{S}}(V_0, V_1, z) = \mathsf{Decide}^{\Psi'}_{\mathfrak{S}}(V_0, V_1, z).$$

Proof. At high level, the proof exploits the $\mathsf{NP}\cap\mathsf{coNP}$ structure; namely, for (V_0, V_1) corresponding to a language $L \in \mathsf{NP}^{\Psi}\cap\mathsf{coNP}^{\Psi}$, and any input z, if $z \in L$, then all the accepting witnesses w certify that $V_1^{\Psi}(z,w) = 1$ and no witness exists that certifies $V_0^{\Psi}(z,w) = 1$ (and vice versa, for $z \notin L$). So, as long as one witness is consistent across the oracles Ψ, Ψ', the answer of $\mathsf{Decide}_{\mathfrak{S}}$ remains invariant. The critical set $\mathbf{C}(\Psi, V_0, V_1, z)$ would simply correspond to the queries made by the verifiers for some specific witness.

Formally, consider any query (V_0, V_1, z). If (V_0, V_1) do not define a language in $\mathsf{NP}\cap\mathsf{coNP}$ relative to some oracle in $\mathfrak{S}$, then by definition $\mathsf{Decide}_{\mathfrak{S}}$ always returns $\perp$, and the claim trivially follows ($\mathbf{C}$ can be set to be the empty set). Hence, from here on, we assume that (V_0, V_1) do define a collection of languages $L^{\mathfrak{S}} = \{L_{\Psi}\}_{\Psi\in\mathfrak{S}}$ in $\mathsf{NP}\cap\mathsf{coNP}$.

Let $b := \mathsf{Decide}^{\Psi}_{\mathfrak{S}}(V_0, V_1, z)$. Consider the lexicographically first witness w which certifies this fact, namely, the first witness for which $V_b^{\Psi}(z,w) = 1$. We define $\mathbf{C} = \mathbf{C}(\Psi, V_0, V_1, z)$ to be the queries $V_b^{\Psi}(z,w)$ makes to Ψ to verify that $V_b^{\Psi}(z,w) = 1$. The bound on the size of $\mathbf{C}$ follows from the fact that V_b is a q-query verifier.

Now, we consider any $\Psi' \in \mathfrak{S}$ that agrees with Ψ on $\mathbf{C}$. Then by definition $V_b^{\Psi'}(z,w) = 1$. Since $\Psi' \in \mathfrak{S}$, the language $L^{\Psi'}$ defined by $V_0^{\Psi'}, V_1^{\Psi'}$ is in $\mathsf{NP}^{\Psi'}\cap\mathsf{coNP}^{\Psi'}$. This fixes the answer $\mathsf{Decide}^{\Psi'}_{\mathfrak{S}}(V_0, V_1, z)$ to b as required. ∎

4.4. Warmup: Injective one-way functions in the presence of $\mathsf{Decide}_{\mathfrak{S}}$. As a warmup, we consider the case where an oracle family that only implements IOWFs, and we show there is no fully black-box construction of a hard $\mathsf{NP} \cap \mathsf{coNP}$ problem from such oracles. This generalizes a result of [MM11] which shows that IOWFs cannot be used to construct OWPs in a black-box manner.[13]

Let $\mathfrak{S}$ be the family of injective one-bit expanding functions. As an implementation for the IOWF we will consider an oracle f that is sampled uniformly at random from $\mathfrak{S}$.

DEFINITION 4.7 (oracle f). *Let $\mathbf{I}_n^m$ denote the distribution on all injective functions from $\{0,1\}^n$ to $\{0,1\}^m$. The IOWF oracle is defined as $f = \{f_n\}_{n\in\mathbb{N}}$ where $f_n \leftarrow \mathbf{I}_n^{n+1}$ for all $n \in \mathbb{N}$.*

As already discussed above, the oracle $\mathsf{Decide}_{\mathfrak{S}}^f$ allows deciding any language in $\mathsf{NP}^f \cap \mathsf{coNP}^f$ given by verifiers V_0, V_1 that define an $\mathsf{NP} \cap \mathsf{coNP}$ language relative to any oracle in $\mathfrak{S}$. We will show that f is one-way, even in the presence of the oracle $\mathsf{Decide}_{\mathfrak{S}}$. We will show that this is the case even given an exponential number of queries to f and $\mathsf{Decide}_{\mathfrak{S}}^f$, and even if the queries (V_0, V_1, z) consist of verifiers that make an exponential number of queries.

In what follows, we call an adversary q-query if on any input y, the adversary makes at most q queries to either f or $\mathsf{Decide}_{\mathfrak{S}}^f$. Furthermore, each query (V_0, V_1, z) to $\mathsf{Decide}_{\mathfrak{S}}^f$ is q-bounded (as previously defined—the verifiers are circuits that make at most q queries to f).

THEOREM 4.8. *Let $q = O(2^{n/3})$. Then any q-query adverasry cannot invert f_n except with exponential small probability:*

$$\Pr_{x\leftarrow\{0,1\}^n, f}\left[\mathcal{A}^{f,\mathsf{Decide}_{\mathfrak{S}}^f}(f_n(x)) = x\right] \le O(2^{-n/3}).$$

Proof. We need to show that even given access to the Decide oracle, an adversary cannot invert f. We show this via a coupling argument. We want to look at the adversary's view in two worlds—the *real* world where the adversary gets a challenge $f(x)$ for a random x and the *ideal* world where the adversary gets a random element in the co-image $y \leftarrow \{0,1\}^{n+1} \setminus \mathrm{Image}(f)$ as the challenge that is completely independent of x. We will show that with very high probability, the adversary's view in both worlds is identical. To this end, we consider three hybrids.

A description of the hybrids is given in Table 4.1.

We will now show that the adversary cannot distinguish between the hybrids and hence cannot invert.

CLAIM 4.9. $\Pr_{f,x,y}[\mathcal{A} \text{ wins in } \mathbf{H_1}] = \Pr_{f,x,y}[\mathcal{A} \text{ wins in } \mathbf{H_2}]$.

Proof. We observe that the view of the adversary is distributed identically in the two hybrids. We are picking a random f and a random y outside the range and planting it at a random $x \in \{0,1\}^n$. The new oracle $f_{x\mapsto y}$ is also uniformly distributed in $\mathbf{I}_n^{n+1}$. Also, in both cases, conditioned on the function, y is distributed uniformly at random in $\mathrm{Image}(f) \cap \{0,1\}^{n+1}$. Overall, the views are identically distributed:

$$(f, f(x)) \equiv (f_{x\mapsto y}, y).$$

[13] [MM11] show a slightly different statement—they consider injective functions that are adaptively one-way. That is, even given the ability to invert the function at all values except the challenge, it is still hard to invert. Our proof works unchanged for this stronger definition. We omit it for simplicity of exposition.

TABLE 4.1
The hybrid experiments.

Hybrid	$\mathbf{H_1}$ (Real)	$\mathbf{H_2}$	$\mathbf{H_3}$ (Ideal)
Injective OWF	$f = \left\{f_k \leftarrow \mathbf{I}_k^{k+1}\right\}_{k\in\mathbb{N}}$		
Preimage	$x \leftarrow \{0,1\}^n$		
Planted image	$y \leftarrow \{0,1\}^{n+1} \setminus \text{Image}(f)$		
Challenge	$f(x)$	y	y
Oracle	$f, \mathsf{Decide}_{\mathfrak{S}}^{f}$	$f_{x\mapsto y}, \mathsf{Decide}_{\mathfrak{S}}^{f_{x\mapsto y}}$	$f, \mathsf{Decide}_{\mathfrak{S}}^{f}$
Winning condition	Find x		

We next show that the hybrids $\mathbf{H_2}$ and $\mathbf{H_3}$ are indistinguishable.

CLAIM 4.10. $|\Pr_{f,x,y}[\mathcal{A} \text{ wins in } \mathbf{H_2}] - \Pr_{f,x,y}[\mathcal{A} \text{ wins in } \mathbf{H_3}]| \le 2^{-n/3}$.

At high level, to show this, we note that $f_{x\mapsto y}$ and f differ in exactly one location—x. Furthermore, we know that in the ideal world ($\mathbf{H_3}$), x is completely independent of the adversary's view. It immediately follows that the probability that queries made to f coincide with x is exponentially small, and thus the answers to these queries wouldn't change in $\mathbf{H_2}$. We would then like to show that the answers given by $\mathsf{Decide}_{\mathfrak{S}}^{f}$ are also invariant with overwhelming probability. Here we shall crucially use the $\mathsf{NP} \cap \mathsf{coNP}$ structure of queries given by Claim 4.6, from which we can deduce that it suffices to show that x does not coincide some small critical set. We now turn to the formal proof.

Proof. We show that, with overwhelming probability, the adversary has the same view (and thus the same output) in both $\mathbf{H_2}$ and $\mathbf{H_3}$:

$$\Pr_{\substack{x\leftarrow\{0,1\}^n, f \\ y\leftarrow\{0,1\}^{n+1}\setminus\text{Image}(f)}}\left[\mathcal{A}^{f_{x\mapsto y},\mathsf{Decide}_{\mathfrak{S}}^{f_{x\mapsto y}}}(y) \neq \mathcal{A}^{f,\mathsf{Decide}_{\mathfrak{S}}^{f}}(y)\right] \le 2^{-n/3}.$$

To show this we prove the following claim.

CLAIM 4.11. *Fix any $f \in \mathfrak{S}$ and $y \in \{0,1\}^{n+1} \setminus \text{Image}(f)$. Then*

1. *for any query (V_0, V_1, z) that $\mathcal{A}^{f,\mathsf{Decide}_{\mathfrak{S}}^{f}}(y)$ makes to $\mathsf{Decide}_{\mathfrak{S}}^{f}$,*

$$\Pr_{x\leftarrow\{0,1\}^n}\left[\mathsf{Decide}^{f}(V_0,V_1,z) \neq \mathsf{Decide}^{f_{x\mapsto y}}(V_0,V_1,z)\right] \le 2^{-2n/3};$$

2. *for any query z that $\mathcal{A}^{f,\mathsf{Decide}_{\mathfrak{S}}^{f}}(y)$ makes to f,*

$$\Pr_{x}[f(z) \neq f_{x\mapsto y}(z)] \le 2^{-n}.$$

Proof. To prove the first part of the claim, we crucially rely on Claim 4.6 (with respect to our family $\mathfrak{S}$ of injective functions). Recall that the adversary $\mathcal{A}$ is a q-query adversary and thus the query (V_0, V_1, z) is q-bounded. Accordingly, by Claim 4.6, there exists a set of critical queries $\mathbf{C} = \mathbf{C}(f, V_0, V_1, z)$ such that for any other $f' \in \mathfrak{S}$ that agrees with f on $\mathbf{C}$,

$$\mathsf{Decide}^{f}(V_0,V_1,z) = \mathsf{Decide}^{f'}(V_0,V_1,z).$$

Thus all that we need to show is that overwhelming probability x is such that $f_{x\mapsto y}$ is injective (namely, in $\mathfrak{S}$) and agrees with f on $\mathbf{C}$. Indeed, $f_{x\mapsto y}$ is always injective since

$y \notin \text{Image}(f)$. Second, $f_{x\mapsto y}|_{\mathbf{C}} = f|_{\mathbf{C}}$ unless $x \in \mathbf{C}$. Since x is sampled independently of $\mathcal{A}$'s view in $\mathbf{H_3}$, and in particular independently of $\mathbf{C}$,

$$\Pr_{x\leftarrow\{0,1\}^n}[x \in \mathbf{C}] \leq |\mathbf{C}| \cdot 2^{-n} \leq q \cdot 2^{-n} \leq 2^{n/3} \cdot 2^{-n} = 2^{-2n/3}.$$

For the second part of the claim, note that $f(z) \neq f_{x\mapsto y}(z)$, unless $z = x$. As before, since x is sampled independently of $\mathcal{A}$'s view in $\mathbf{H_3}$, and in particular independently of z, this probability is at most 2^{-n}. ∎

Given Claim 4.11, we can take a union bound over all queries that $\mathcal{A}^{f,\mathsf{Decide}^f_{\mathfrak{S}}}(y)$ makes to deduce that the answers to all remain invariant when considering the oracles $f_{x\mapsto y}, \mathsf{Decide}^{f_{x\mapsto y}}_{\mathfrak{S}}$ except with probability

$$q \cdot \max(2^{-2n/3}, 2^{-n}) \leq 2^{n/3} \cdot 2^{-2n/3} = 2^{-n/3}.$$

This completes the proof of Claim 4.10. ∎

To complete the proof of Theorem 4.8, it is left to note that in the ideal world, the adversary cannot invert.

CLAIM 4.12. *The adversary cannot win in the ideal world. Concretely, for every fixed f,*

$$\Pr_{x,y}[\mathcal{A} \textit{ wins in } \mathbf{H_3}] = 2^{-n}.$$

Proof. In the third hybrid $\mathbf{H_3}$, the challenge y is independent of the answer x, which is chosen uniformly at random. So, with probability 2^{-n}, the adversary's response will be x. ∎

Putting all of the above claims together, the adversary inverts in the real world ($\mathbf{H_1}$) with probability at most

$$2^{-2n/3} + 2^{-n} \leq O(2^{-n/3}).$$

4.5. Indistinguishability obfuscation (and IOWFs) in the presence of Decide. In this section, we generalize Theorem 4.8 to show that IOWFs *and IO* cannot be used to construct worst-case hard $\mathsf{NP}\cap\mathsf{coNP}$ instances in a fully black-box way. We start by discussing an aspect of IO that turns out to be crucial for this separation—*verifiability*.

Verifiability of IO. Looking back at our separation for the SZK case in section 3, we observe that it, in fact, holds also for a stronger definition of IO that is verifiable and *unambiguous*; namely, it is possible to efficiently determine whether a given string is a valid obfuscation of *some* circuit, and this circuit is uniquely determined. Indeed, looking at the oracle $\Psi = (f, \mathcal{O}, \mathsf{Eval}^{\mathcal{O},f})$, implementing OWFs and IO there, it induces *valid* obfuscation, which are strings $\widehat{C} = \mathcal{O}(C, r)$ in the image of the *injective* $\mathcal{O}$, and *invalid* ones, which are strings outside the image of $\mathcal{O}$. Furthermore, it is possible to efficiently identify which is the case, since the oracle Eval would return $\perp$ on invalid obfuscations.

Going back to the case of $\mathsf{NP}\cap\mathsf{coNP}$, we observe that verifiable and unambiguous IO actually does imply hardness in $\mathsf{NP}\cap\mathsf{coNP}$ (in a fully black-box way). Indeed, consider the language including all $(\widehat{C}, i, b)$ such that $\widehat{C}$ is a valid obfuscation and b the ith bit of the unique circuit C it determines. Indeed, due to verifiablility and unambiguity, this language is in $\mathsf{NP}\cap\mathsf{coNP}$, and clearly any decider for this language

completely breaks IO. This means that we cannot hope to rule out fully black-box constructions of $\mathsf{NP} \cap \mathsf{coNP}$ hardness from a family of oracles $\mathfrak{S}$ if this family only includes verifiable and unambiguous IO constructions. Indeed, our Definition 4.2 of black-box constructions of hard $\mathsf{NP} \cap \mathsf{coNP}$ problems considers constructions that should work for the family $\mathfrak{S}$ of all IO constructions, and we will crucially (and necessarily) rely on this. (In fact, our separation would also work for the restricted family of IO constructions that are not verifiable but still unambiguous.)

Capturing nonverifiable IO. We augment our previous definition of the oracle $\Psi = (f, \mathcal{O}, \mathsf{Eval}^{\mathcal{O},f})$ in a way that allows the Eval oracle to answer arbitrarily on invalid obfuscations, which would capture nonverifable IO constructions. To this end, we consider an augmented Eval_φ parameterized by a "backup map" $\varphi : \{\varphi_n : \{0,1\}^{5n} \to \{0,1\}^n\}_n$ from obfuscations $\widehat{C}$ to circuits C. Given a query $(\widehat{C}, x)$, if the obfuscation $\widehat{C}$ is valid, Eval_φ answers it faithfully as the previously defined Eval; otherwise, Eval_φ obtains some circuit $C = \varphi(\widehat{C})$ from φ and uses it to answer the query. Indeed, this new oracle still implements IO and does so in a nonverifiable way. This is formally defined below.

DEFINITION 4.13 (oracle Ψ_φ). *The oracle $\Psi_\varphi = (f, \mathcal{O}, \mathsf{Eval}_\varphi^{f,\mathcal{O}})$ consists of three parts:*

- *$f = \{f_n\}_{n\in\mathbb{N}}$ on input $x \in \{0,1\}^n$ answers with $f_n(x)$, where f_n is a random injective one-bit expanding function $f_n \leftarrow \mathbf{I}_n^{n+1}$.*
- *$\mathcal{O} = \{\mathcal{O}_n\}_{n\in\mathbb{N}}$ on input $(C, r) \in \{0,1\}^{n\times 2}$ answers with $\widehat{C} := \mathcal{O}_n(C, r)$ where $\mathcal{O}_n$ is a random injective function $\mathcal{O}_n \leftarrow \mathbf{I}_{2n}^{5n}$ into $\{0,1\}^{5n}$.*
- *$\mathsf{Eval}_\varphi^{f,\mathcal{O}}(\widehat{C}, x)$ checks if $\widehat{C}$ is in the image of $\mathcal{O}_n$. If it is, it finds $(C, r) = \mathcal{O}_n^{-1}(\widehat{C})$ and returns the answer $C^f(x)$. If $\widehat{C}$ is not in the image, it uses φ to answer. That is,*

$$\mathsf{Eval}_\varphi^{f,\mathcal{O}}(\widehat{C}, x) = \begin{cases} C^f(x) & \text{if } \widehat{C} \in \mathrm{Image}(\mathcal{O}_n) \text{ and } \mathcal{O}_n(C, r) = \widehat{C}, \\ C_\varphi^f(x) & \text{if } \widehat{C} \notin \mathrm{Image}(\mathcal{O}_n) \text{ and } C_\varphi = \varphi(\widehat{C}). \end{cases}$$

For any choice of φ, and realization of Ψ_φ, we obtain a construction of an obfuscator similarly to Construction 3.20.

CONSTRUCTION 4.14 (obfuscator $i\mathcal{O}^{\Psi_\varphi}$). *Let $\Psi_\varphi = (f, \mathcal{O}, \mathsf{Eval}_\varphi^{f,\mathcal{O}})$. Given an oracle-aided circuit $C \in \{0,1\}^n$, $i\mathcal{O}^{\Psi}(1^n, C)$ samples a random $r \leftarrow \{0,1\}^n$, computes $\widehat{C} = \mathcal{O}(C, r)$, and returns an oracle aided circuit $E_{\widehat{C}}$ that given input x computes $\mathsf{Eval}_\varphi^{f,\mathcal{O}}(\widehat{C}, x)$.*

As in section 3, $i\mathcal{O}^{\Psi_\varphi}$ satisfies the functionality requirement of Definition 2.2 for f-aided circuits, and this is the case for any choice of mapping φ. Indeed, functionality puts no restriction on how evaluation behaves for invalid obfuscations. Accordingly, the family $\mathfrak{S}$, considered in our Definition 4.2 of black-box constructions of hard problems in $\mathsf{NP} \cap \mathsf{coNP}$, includes $i\mathcal{O}^{\Psi_\varphi}$ for all φ. From here on, we shall often abuse notation and write $\Psi_\varphi \in \mathfrak{S}$ rather than $i\mathcal{O}^{\Psi_\varphi} \in \mathfrak{S}$.

To rule out fully black-box constructions relative to the family $\mathfrak{S}$, we consider again the oracle $\mathsf{Decide}_{\mathfrak{S}}$. We show that for any specific choice of φ, in the presence of $\mathsf{Decide}_{\mathfrak{S}}^{\Psi_\varphi}$,

1. any language L^{Ψ_φ} defined by $(V_0^{\Psi_\varphi}, V_1^{\Psi_\varphi})$ is easy to decide provided that (V_0, V_1) define a language in $\mathsf{NP} \cap \mathsf{coNP}$ relative to any oracle in $\mathfrak{S}$,
2. f is an OWF,

3. $i\mathcal{O}^{\Psi_\varphi}$ is a secure IO.

The first item above indeed follows from Definition 4.4 of the oracle $\mathsf{Decide}_{\mathfrak{S}}$ as claimed in subsection 4.3. We stress the crucial reliance on the fact that V_0, V_1 define a language in $\mathsf{NP} \cap \mathsf{coNP}$ for all oracles in $\mathfrak{S}$. Indeed, the oracle $\mathsf{Decide}_{\mathfrak{S}}$ only responds on such V_0, V_1. The fact that $\mathsf{Decide}_{\mathfrak{S}}$ only responds on such queries is crucially used to prove one-wayness (the second item) and IO (the third item), where we shall use the fact $\{\Psi_\varphi\} \in \mathfrak{S}$ for all φ.

In the next two subsections, we prove the last two items. Throughout, we address adversaries with oracles $\Psi_\varphi = (f, \mathcal{O}, \mathsf{Eval}_\varphi^{\mathcal{O},f})$ and $\mathsf{Decide}_{\mathfrak{S}}^{\Psi_\varphi}$. We say that such an adversary is q-query if it

1. makes only q queries to f,
2. makes only q queries to either $\mathcal{O}$ or Eval, and any query $\widehat{C}$ to Eval is of size at most $5q$, and in particular, any oracle-aided circuit C that is mapped to $\widehat{C}$ by $\mathcal{O}$ is of size at most q, and makes at most q queries to f,
3. makes only q queries to $\mathsf{Decide}_{\mathfrak{S}}^{\Psi_\varphi}$, and for any query (V_0, V_1, z) made to $\mathsf{Decide}_{\mathfrak{S}}^{\Psi_\varphi}$, the verification circuits V_0, V_1 are Ψ_φ-aided and each of them is q-query (according to the two conditions above).

4.5.1. One-wayness. We show that f is an OWF in the presence of the $\mathsf{Decide}_{\mathfrak{S}}^{\Psi}$ oracle.

THEOREM 4.15. *Let $q(n) \le O(2^{n/6})$. Fix any φ. Then for any q-query adversary $\mathcal{A}$,*

$$\Pr_{\substack{\Psi_\varphi=(f,\mathcal{O},\mathsf{Eval}_\varphi^{f,\mathcal{O}}) \\ x}} \left[\mathcal{A}^{\Psi,\mathsf{Decide}_{\mathfrak{S}}^{\Psi}}(f(x)) = x\right] \le O(2^{-n/3}),$$

where the probability is over the randomness of Ψ_φ and $x \leftarrow \{0,1\}^n$.

Proof. We will, in fact, prove a stronger statement: the above holds when fixing the oracles $\mathcal{O}$ and $f_{-n} := \{f_k\}_{k \neq n}$. We prove the theorem by a reduction to the case that Ψ only consists of the injective function f (and does not include $\mathcal{O}, \mathsf{Eval}_\varphi^{f,\mathcal{O}}$), proven in Theorem 4.8. Concretely, fix any q-query adversary $\mathcal{A}$ that inverts the random injective function f_n given access to $\Psi = (f, \mathcal{O}, \mathsf{Eval}_\varphi^{f,\mathcal{O}})$ and $\mathsf{Decide}_{\mathfrak{S}}^{\Psi}$, we show how to reduce it to an $O(q^2)$-query adversary $\mathcal{B}^f(f_n(x))$ that inverts f_n for a random $x \leftarrow \{0,1\}^n$ with the same probability as $\mathcal{A}$. (This is done similarly to the proof of Theorem 3.11).

The new adversary $\mathcal{B}^{f,\mathsf{Decide}_{\mathfrak{S}}^{f}}(f_n(x))$ emulates $\mathcal{A}^{\Psi,\mathsf{Decide}_{\mathfrak{S}}^{\Psi}}(f_n(x))$ answering Ψ-queries as follows:

- f queries: answered according to $\mathcal{B}$'s oracle f. This translates to at most q queries to f.
- $\mathcal{O}$ queries: answered according to the fixed oracle $\mathcal{O}$. This does not add any calls to f.
- $\mathsf{Eval}_\varphi^{f,\mathcal{O}}$ queries: given query $(\widehat{C}, x)$ to Eval, invert the fixed oracle $\mathcal{O}$ to find $(C, r) = \mathcal{O}^{-1}(\widehat{C})$. If no such preimage exists, set $C = \varphi(\widehat{C})$. Using the f-oracle, compute $C^f(x)$ and return the result. This translates to at most q^2 queries to f: q queries by C, for each of at most q queries to $\mathsf{Eval}_\varphi^{f,\mathcal{O}}$.
- $\mathsf{Decide}_{\mathfrak{S}}^{\Psi_\varphi}$ queries: given query (V_0, V_1, z), where V_b makes Ψ_φ-queries, translate the query to D_0, D_1, z that only make f-queries, where each query to $\Psi_\varphi = (f, \mathcal{O}, \mathsf{Eval}_\varphi^{f,\mathcal{O}})$ is translated to a query to f according to the previous three items. The resulting oracle-aided query (D_0, D_1, z) may thus make up

to $q+q^2$ queries to f: q corresponding to the first item, and q^2 corresponding to the third. We note that while there is a bound on the number of queries that they make, we do not put any restrictions on their size, which allows us to hardwire the fixed $\mathcal{O}$ and f_{-n} as required in the previous three items. Indeed, Theorem 4.8 does not put any restriction on the size of these circuits.

Overall, $\mathcal{B}^f$ is $O(q^2)$-query and perfectly emulates the view of $\mathcal{A}^\Psi$. The theorem now follows from Theorem 4.8. ∎

4.5.2. Indistinguishability obfuscation. We now show that Construction 4.14 also satisfies indistinguishability, with an exponentially small distinguishing gap, even given an exponential number of oracle queries to $\Psi_\varphi = (f, \mathcal{O}, \mathsf{Eval}_\varphi^{\mathcal{O},f})$ and the decide oracle $\mathsf{Decide}_{\mathfrak{S}}^{\Psi_\varphi}$.

THEOREM 4.16. *Let $q(n) \le O(2^{n/3})$. Fix any φ. Then for any q-query distinguisher $\mathcal{A} = (\mathcal{A}_1, \mathcal{A}_2)$*

$$\mathop{\mathbb{E}}_{\mathcal{O}}\left[\mathrm{PAdv}^{\mathsf{iO}}_{\Gamma, i\mathcal{O}, \mathcal{A}}(n)\right] \le O(2^{-n/3}),$$

where the random variable $\mathrm{PAdv}^{\mathsf{iO}}_{\Gamma, i\mathcal{O}, \mathcal{A}}(n)$ is the adversary's winning probability in the IO security game (Definition 2.3) relative to $\Psi_\varphi = (f, \mathcal{O}, \mathsf{Eval}_\varphi^{f,\mathcal{O}})$ and $\mathsf{Decide}_{\mathfrak{S}}^{\Psi_\varphi}$.

By Lemma 2.4 we can bound the advantage in terms of the positive advantage.

COROLLARY 4.17. *Let $q(n) \le O(2^{n/3})$. Then for any q-query distinguisher $\mathcal{A}$*

$$\mathop{\mathbb{E}}_{\mathcal{O}}\left[\mathsf{Adv}^{\mathsf{iO}}_{\Gamma, i\mathcal{O}, \mathcal{C}, \mathcal{A}}(n)\right] \le O(2^{-n/6}).$$

Next we prove Theorem 4.16.

Proof of Theorem 4.16. We prove a stronger statement: the above holds when fixing the oracles f and $\mathcal{O}_{-n} = \{\mathcal{O}_k\}_{k \ne n}$. For simplicity, we often suppress oracle access to the fixed $\mathcal{O}_{-n}, f$ in our notation and only denote the oracle $\mathcal{O}_n$. Fix a q-query (without loss of generality deterministic) adversary $\mathcal{A} = (\mathcal{A}_1, \mathcal{A}_2)$. To bound $\mathcal{A}$'s advantage in breaking $i\mathcal{O}$, we rely on a proof strategy similar to the one in Theorem 4.8. We will consider an *ideal* world where the given challenge is uncorrelated to the bit b that the adversary will be required to guess. We will then show, through a sequence of hybrid experiments, that this world is indistinguishable from the *real* world, where the adversary get an obfuscation of the circuit C_b among two circuits C_0, C_1, which it chose.

We introduce some notation that will be useful to describe the hybrids:

- We use regular expressions to describe sets, in particular, the $*$ expression. We denote by $(*, r) = \left\{(C, r) : C \in \{0,1\}^* \text{ and, } |C| = |r|\right\}$.
- Let $A \xleftarrow{k} B$ denote sampling uniformly at random an ordered size-k subset $A = \{a_1, \ldots, a_k\}$ from the set B without repetition (a_i is sampled uniformly at random from $B \setminus \{a_1, \ldots, a_{i-1}\}$).
- For a function $\mathcal{O} = \{\mathcal{O}_k : \{0,1\}^{2k} \to \{0,1\}^{5k}\}_{k \in \mathbb{N}}$, a pair $(C, r) \in \{0,1\}^{n \times 2}$, and $\widehat{C} \in \{0,1\}^{5n}$, we denote by $\mathcal{O}_{(C,r) \mapsto \widehat{C}}$ the function that maps (C, r) to $\widehat{C}$ and is otherwise identical to $\mathcal{O}$. We also naturally extend this notation to *ordered* sets $T \subseteq \{0,1\}^{n \times 2}$ and $D \subseteq \{0,1\}^{5n}$ of the same size: $\mathcal{O}_{T \mapsto D}$ is a function that maps $(C, r) \in T$ to the corresponding element in D and is otherwise identical to $\mathcal{O}$.

- For a function $\varphi = \{\varphi_k : \{0,1\}^{5k} \to \{0,1\}^{2k}\}_{k\in\mathbb{N}}$, $C \in \{0,1\}^n$, and $\widehat{C} \in \{0,1\}^{5n}$, we denote by $\varphi_{\widehat{C}\mapsto C}$ the function that maps $\widehat{C}$ to C and is otherwise identical to φ.
- For functions $\mathcal{O}, \varphi = \{\mathcal{O}_k, \varphi_k : \{0,1\}^{2k} \to \{0,1\}^{5k}\}_{k\in\mathbb{N}}$, we denote by $\Gamma(f, \mathcal{O}, \varphi)$ the oracle

$$\Gamma(f,\mathcal{O},\varphi) := f, \mathcal{O}, \mathsf{Eval}_{\varphi}^{f,\mathcal{O}}, \mathsf{Decide}_{\mathfrak{S}}^{f,\mathcal{O},\mathsf{Eval}_{\varphi}^{f,\mathcal{O}}},$$

  where we extend $\mathsf{Eval}_{\varphi}^{f,\mathcal{O}}$ to also be defined for noninjective $\mathcal{O}$: given $\widehat{C} \in \{0,1\}^{5n}$ with more than a single preimage in $\{0,1\}^{2n}$, it returns $\perp$. Also, Decide when given access to any noninjective $\mathcal{O}$ outputs $\perp$ for all queries.

The hybrid experiments are formally described in Table 4.2. We briefly describe the hybrids in words below:

- $\mathbf{H_1}$: Hybrid $\mathbf{H_1}$ is the real world where $\mathcal{S}$ wins if it produces functionally equivalent circuits C_0, C_1, and it successfully guesses the bit b.
- $\mathbf{H_2}$: This hybrid changes the oracle seen by the adversary in the prechallenge phase, that is, before the adversary issues the challenge. In this hybrid, the pre-challenge obfuscation oracle is changed such that, all obfuscation queries C from the set $Gamma(*, r)$ are answered according to an independently chosen random subset D. That way the adversary S does not learn any information about the obfuscation oracle at all locations that use this chosen randomness r.
- $\mathbf{H_3}$: These two hybrids are identical except we rename the challenge from $\Gamma(C_b, r)$ to $\widehat{C}$ a randomly chosen label. These two hybrids are identical other than this renaming.
- $\mathbf{H_4}$: In this hybrid, we switch to an ideal world where the obfuscation to the challenge given to the adversary is completely independent of the challenge and the randomness. In this world, the adversary has an advantage of 0 in the obfuscation security game.

We show that the probability that the simulator wins in any of the experiments is roughly the same and that in hybrid $\mathbf{H_4}$ the probability that $\mathcal{S}$ wins is $1/2$ for each oracle $\mathcal{O}$.

In order to prove the theorem, we show that $\Pr[\mathcal{A} \text{ wins in } \mathbf{H}_i]$ does not change much between the hybrids $i \in \{1,2,3,4\}$. Then showing that in the last hybrid $\mathbf{H_4}$, the adversary's advantage in the IO game is exactly 0 for each oracle $\mathcal{O}$, we bound the positive advantage in the real world.

TABLE 4.2
The hybrid experiments.

Hybrid	$\mathbf{H_1}$ (Real)	$\mathbf{H_2}$	$\mathbf{H_3}$	$\mathbf{H_4}$ (Ideal)
Obfuscator function	$\mathcal{O}_n \leftarrow \mathbf{I}_{2n}^{5n}$			
Backup map	φ			
Challenger randomness	$b \leftarrow \{0,1\}$, $r \leftarrow \{0,1\}^n$			
Prechallenge remapped set T	$(*, r)$			
Prechallenge output set D	$D \overset{2^n}{\leftarrow} \{0,1\}^{5n} \setminus \text{Image}(\mathcal{O})$			
Planted challenge	$\widehat{C} \leftarrow \{0,1\}^{5n} \setminus (\text{Image}(\mathcal{O}) \cup D)$			
Prechallenge oracle	$\Gamma(f,\mathcal{O},\varphi)$	$\Gamma(f,\mathcal{O}_{T\mapsto D},\varphi)$		$\Gamma(f,\mathcal{O},\varphi)$
Challenge obfuscation	$\mathcal{O}(C_b, r)$	$\widehat{C}$		
Postchallenge oracle	$\Gamma(f,\mathcal{O},\varphi)$	$\Gamma(f,\mathcal{O}_{(C_b,r)\mapsto\widehat{C}},\varphi)$		$\Gamma(f,\mathcal{O},\varphi_{\widehat{C}\mapsto C_0})$
Winning condition	The adversary outputs b.			

CLAIM 4.18. $|\Pr[\mathcal{A} \text{ wins in } \mathbf{H_1}] - \Pr[\mathcal{A} \text{ wins in } \mathbf{H_2}]| \leq O(2^{-n/3})$.

Proof. The hybrids $\mathbf{H_1}$ and $\mathbf{H_2}$ differ in the prechallenge oracle, which is changed from $\Gamma(f, \mathcal{O}, \varphi)$ to $\Gamma(f, \mathcal{O}_{T \mapsto D}, \varphi)$ where $T = \{(C, r) : C \in \{0,1\}^n\}$ and D is a random ordered subset of $\{0,1\}^{5n} \setminus \text{Image}(\mathcal{O})$ of size 2^n. Note that like $\mathcal{O}$, $\mathcal{O}_{T \mapsto D}$ is injective and thus the oracles in both hybrids respect $\mathfrak{S}$.

We bound the difference by a coupling argument. Concretely,

$$\left|\Pr[\mathcal{A} \text{ wins in } \mathbf{H_1}] - \Pr[\mathcal{A} \text{ wins in } \mathbf{H_2}]\right| \leq \mathop{\mathbb{E}}_{\mathcal{O}}\left[\Pr_{r,b,D}\left[\mathcal{A}_1^{\Gamma(f,\mathcal{O},\varphi)}(1^n) \neq \mathcal{A}_1^{\Gamma(f,\mathcal{O}_{T\mapsto D},\varphi)}(1^n)\right]\right],$$

where the probability is over the choice of $r \leftarrow \{0,1\}^n$, $b \leftarrow \{0,1\}$, $\mathcal{O} \leftarrow \mathbf{I}_{2n}^{5n}$, and D. We will, in fact, show that the above is bounded for any fixed $b \in \{0,1\}, \mathcal{O}$. Indeed, for the rest of the claim, fix $b \in \{0,1\}$ and $\mathcal{O} \in \mathbf{I}_{2n}^{5n}$.

In what follows, let $\mathbf{Q}$ be the set of queries made by $\mathcal{A}_1^{\Gamma(f,\mathcal{O},\varphi)}(1^n)$ to its oracle

$$\Gamma(f, \mathcal{O}, \varphi) = \left(\Psi_\varphi, \mathsf{Decide}_{\mathfrak{S}}^{\Psi_\varphi}\right) \quad \text{where} \quad \Psi_\varphi = \left(f, \mathcal{O}, \mathsf{Eval}_\varphi^{f,\mathcal{O}}\right).$$

For any query $Q = (V_0, V_1, z) \in \mathbf{Q}$ made to $\mathsf{Decide}_{\mathfrak{S}}^{\Psi_\varphi}$, let $\mathbf{C}_Q$ denote the set of critical queries corresponding to Q, given by Claim 4.6. Note that indeed $\Psi_\varphi \in \mathfrak{S}$; namely, it is a valid oracle in the family $\mathfrak{S}$, so the claim can be applied. We denote by $\mathbf{C_Q} = \bigcup_{Q \in \mathbf{Q}} \mathbf{C}_Q$ the union of all such critical sets. We now define the events **rHit** and **DHit**, aimed at capturing the cases where the adversary's views in $\mathbf{H_1}$ and $\mathbf{H_2}$ may differ. Concretely, let $\mathbf{rHit} = \mathbf{rHit}(r)$ be the event that either of the following occurs:

1. $(C, r) \in \mathbf{Q} \cup \mathbf{C_Q}$ for some $C \in T$: the query (C, r) is made to $\mathcal{O}$.
2. $(\mathcal{O}(C, r), x) \in \mathbf{Q} \cup \mathbf{C_Q}$ for some $C \in T$ and $x \in \{0,1\}^n$: the query $(\mathcal{O}(C, r), x)$ is made to Eval_φ.

We define the event $\mathbf{DHit} = \mathbf{DHit}(D)$ that occurs when $(\widehat{D}, x) \in \mathbf{Q} \cup \mathbf{C_Q}$ for some $\widehat{D} \in D$ and $x \in \{0,1\}^n$: the query $(\widehat{D}, x)$ is made to Eval_φ. Note that for both the events, any query Q as above either is made directly to the corresponding oracle, namely $Q \in \mathbf{Q}$, or is within the critical set of queries $\mathbf{C_Q}$ (meaning that there is a query (V_0, V_1, z) to $\mathsf{Decide}_{\mathfrak{S}}$ where one of the verifiers might make the query Q on some canonical witness; see Claim 4.6).

CLAIM 4.19. $\Pr_{r,D}[\mathcal{A}_1^{\Gamma(f,\mathcal{O},\varphi)}(1^n) \neq \mathcal{A}_1^{\Gamma(f,\mathcal{O}_{T\mapsto D},\varphi)}(1^n)] \leq \Pr_{r,D}[\mathbf{rHit} \vee \mathbf{DHit}]$.

Proof. We observe that if either of **rHit** or **DHit** does not occur, then the view of $\mathcal{A}_1$ is the same for both oracles. Indeed, in $\mathbf{H_2}$, the oracle $\mathcal{O}$ is changed to $\mathcal{O}_{T \mapsto D}$ where $T = (*, r) = \{(C, r) : C \in \{0,1\}^n\}$ and D is a random subset of $\{0,1\}^{5n} \setminus \text{Image}(\mathcal{O})$ of size 2^n. In particular,

- for any query $Q \notin T$ to $\mathcal{O}$

$$\mathcal{O}_{T \mapsto D}(Q) = \mathcal{O}(Q),$$

- for any query $Q \notin \{(\widehat{D}, *) \cup (\mathcal{O}(T), *)\}$ to $\mathsf{Eval}_\varphi^{f,\mathcal{O}}$,

$$\mathsf{Eval}_\varphi^{f,\mathcal{O}_{(C_b,r)\mapsto\widehat{D}}}(Q) = \mathsf{Eval}_\varphi^{f,\mathcal{O}}(Q).$$

It follows that when neither **rHit** nor **DHit** occurs for any query in $Q \in \mathbf{Q} \cup \mathbf{C_Q}$, $\Psi_\varphi(Q) = \Psi'_\varphi(Q)$, where

$$\Psi_\varphi = f, \mathcal{O}, \mathsf{Eval}_\varphi^{f,\mathcal{O}} \quad \text{and} \quad \Psi'_\varphi = f, \mathcal{O}_{T \mapsto D}, \mathsf{Eval}_\varphi^{f,\mathcal{O}_{T\mapsto D}}.$$

This implies that all queries made by $\mathcal{A}_1$ directly to its oracle Ψ_φ in $\mathbf{H_1}$ are answered in the same way in $\mathbf{H_2}$. It is left to show that this is also the case for queries made to $\mathsf{Decide}_{\mathfrak{S}}^{\Psi_\varphi}$. For this purpose, note that $\Psi'_\varphi \in \mathfrak{S}$, namely, it is a valid oracle (indeed, the set D is chosen outside the image of $\mathcal{O}$, so injectivity of the obfuscation oracle is guaranteed as required). Furthermore, the oracles Ψ_φ and Ψ'_φ agree on all critical sets $\mathbf{C}_Q \subseteq \mathbf{C_Q}$. It follows, by Claim 4.6, that for any $Q \in \mathbf{Q}$

$$\mathsf{Decide}_{\mathfrak{S}}^{\Psi_\varphi}(Q) = \mathsf{Decide}_{\mathfrak{S}}^{\Psi'_\varphi}(Q),$$

which completes the proof of the claim. ∎

It is left to bound the probability that $\mathbf{rHit} \vee \mathbf{DHit}$ occurs. First, since r is chosen at random from $\{0,1\}^n$, for any fixed query $Q \in \mathbf{Q} \cup \mathbf{C_Q}$, the probability that either of the two cases defining $\mathbf{rHit}$ occurs is 2^{-n}. Thus, by a union bound we have

$$\Pr[\mathbf{rHit}] \leq |\mathbf{Q} \cup \mathbf{C_Q}| \cdot O(2^{-n}) \leq \left(|\mathbf{Q}| + \sum_{Q \in \mathbf{Q}} |\mathbf{C}_Q| \right) \cdot 2^{-n} \leq O(q^2 \cdot 2^{-n}) \leq O(2^{-n/3}).$$

As for $\mathbf{DHit}$, observe that the set D is a set of size 2^n picked at random from $\{0,1\}^{5n} \setminus \mathrm{Image}(\mathcal{O})$. For any fixed query $(\widehat{D}, x) \in \mathbf{Q} \cup \mathbf{C_Q}$, it holds that $\Pr[\widehat{D} \in D] \leq \frac{2^n}{2^{5n}-2^n}$. Hence,

$$\Pr[\mathbf{DHit}] \leq |\mathbf{Q} \cup \mathbf{C_Q}| \cdot \frac{2^n}{2^{5n}-2^n} < O(2^{-10n/3}) \leq O(2^{-n/3}) .$$

This completes the proof of Claim 4.18. ∎

Claim 4.20. $\Pr[\mathcal{A}$ *wins in* $\mathbf{H_2}] = \Pr[\mathcal{A}$ *wins in* $\mathbf{H_3}]$.

Proof. The difference between $\mathbf{H_2}$ and $\mathbf{H_3}$ is that in $\mathbf{H_3}$, in the challenge and postchallenge phases, the value $\mathcal{O}(C_b, r)$ is resampled uniformly at random from the co-image, namely, it is replaced everywhere by $\widehat{C} \leftarrow \{0,1\}^{5n} \setminus (\mathrm{Image}(\mathcal{O}) \cup D)$. Note that like $\mathcal{O}$, $\mathcal{O}_{(C_b,r)\mapsto\widehat{C}}$ is injective and thus the new oracle respects $\mathfrak{S}$.

We claim that the two hybrids induce exactly the same distribution on $\mathcal{A}$'s view. Indeed, in $\mathbf{H_2}$, at the end of the prechallenge phase, fixing the view of $\mathcal{A}$, the distribution of $\mathcal{O}(C_b, r)$ is uniformly random in $S := \{0,1\}^{5n} \setminus (\mathrm{Image}(\mathcal{O}_{(C_b,r)\mapsto\perp}) \cup D)$. In $\mathbf{H_2}$ $\mathcal{O}(C_b, r)$ is sampled uniformly at random directly from S, whereas in $\mathbf{H_3}$, we first sample a random value $\mathcal{O}(C_b, r)$ from S, and then resample $\widehat{C}$ from $S \setminus \{\mathcal{O}(C_b, r)\}$, which again gives a uniformly random value in S. ∎

Finally, we show that the adversary's advantage does not change much as we shift from hybrid $\mathbf{H_3}$ to $\mathbf{H_4}$. This proof is almost identical to the proof of Claim 4.18. Here we use the fact that modifying φ does not alter the fact that Ψ_φ is a valid implementation of IO.

Claim 4.21. $|\Pr[\mathcal{A}$ *wins in* $\mathbf{H_3}] - \Pr[\mathcal{A}$ *wins in* $\mathbf{H_4}]| \leq O(2^{-n/3})$.

Proof. There are two differences between the hybrids. The first is in the oracle that $\mathcal{A}_1$ is given before the challenge phase: $\Gamma(f, \mathcal{O}, \varphi)$ in $\mathbf{H_4}$, and its tweaked version $\Gamma(f, \mathcal{O}_{T\mapsto D}, \varphi)$ in $\mathbf{H_3}$. The second is in the oracle that $\mathcal{A}_2$ is given after the challenge phase: $\Gamma(f, \mathcal{O}, \varphi_{\widehat{C}\mapsto C_0})$ in $\mathbf{H_4}$, and $\Gamma(f, \mathcal{O}_{(C_b,r)\mapsto\widehat{C}}, \varphi)$ in $\mathbf{H_3}$. Note that $\mathcal{O}$ is injective, and thus the new oracles respect $\mathfrak{S}$. We bound the difference between the winning probabilities in $\mathbf{H_3}$ and $\mathbf{H_4}$ as follows:

$$\Big|\Pr[\mathcal{A} \text{ wins in } \mathbf{H_3}] - \Pr[\mathcal{A} \text{ wins in } \mathbf{H_4}]\Big|$$
$$\le \mathop{\mathbb{E}}_{\mathcal{O}}\Bigg[\Pr_{r,b,\widehat{D}}\Big[\mathsf{state} := \mathcal{A}_1^{\Gamma(f,\mathcal{O},\varphi)}(1^n) \neq \mathcal{A}_1^{\Gamma(f,\mathcal{O}_{T\mapsto D},\varphi)}(1^n)\Big]$$
$$+ \Pr_{r,b,\widehat{C}}\Big[\mathcal{A}_2^{\Gamma(f,\mathcal{O},\varphi_{\widehat{C}\mapsto C_0})}(\mathsf{state},\widehat{C}) \neq \mathcal{A}_2^{\Gamma(f,\mathcal{O}_{(C_b,r)\mapsto\widehat{C}},\varphi)}(\mathsf{state},\widehat{C})\,\Big|\,\mathsf{state} = \mathcal{A}_1^{\Gamma(f,\mathcal{O},\varphi)}(1^n)\Big]\Bigg],$$

where the probabilities are over the choice of $r \leftarrow \{0,1\}^n$, $b \leftarrow \{0,1\}$, $\mathcal{O} \leftarrow \mathbf{I}_{2n}^{5n}$, and $D \overset{\leftarrow}{2^n} \{0,1\}^{5n} \setminus \mathrm{Image}(\mathcal{O}_n)$ and $\widehat{C} \leftarrow \{0,1\}^{5n} \setminus (\mathrm{Image}(\mathcal{O}) \cup D)$.

We proved in Claim 4.18 that the first summand is bounded by $O(2^{-n/3})$. Next, we prove a similar bound for the second summand using an argument similar to that of Claim 4.18. The key difference between the two arguments is the following: in Claim 4.18, we argued that $\Gamma(f,\mathcal{O}_{T\mapsto D},\varphi)$ agrees with $\Gamma(f,\mathcal{O},\varphi)$ on the set of queries $\mathbf{Q}$ made by the adversary to the last oracle. However, this argument crucially relied on D being chosen at random independently of the adversary's view. Now, this is no longer the case; the oracle in $\mathbf{H_3}$ is $\Gamma(f,\mathcal{O}_{(C_b,r)\mapsto\widehat{C}},\varphi)$, where $\widehat{C}$ is the challenge, which is known to the adversary. Instead, we will be able to show that the last oracle agrees with $\Gamma(f,\mathcal{O},\varphi_{\widehat{C}\mapsto C_0})$ on the adversary's queries.

We will, in fact, show that the second summand is bounded for any fixed $b,\mathcal{O},\widehat{C}$, state. Indeed, for the rest of the claim, fix all of the above. In what follows, let $\mathbf{Q}$ be the set of queries made by $\mathcal{A}_2^{\Gamma(f,\mathcal{O},\varphi_{\widehat{C}\mapsto C_0})}(\mathsf{state},\widehat{C})$ to its oracle

$$\Gamma(f,\mathcal{O},\varphi_{\widehat{C}\mapsto C_0}) = \left(\Psi_{\varphi_{\widehat{C}\mapsto C_0}}, \mathsf{Decide}_{\mathfrak{S}}^{\Psi_{\varphi_{\widehat{C}\mapsto C_0}}}\right), \quad \text{where} \quad \Psi_{\varphi_{\widehat{C}\mapsto C_0}} = \left(f,\mathcal{O},\mathsf{Eval}_{\varphi_{\widehat{C}\mapsto C_0}}^{f,\mathcal{O}}\right).$$

For any query $Q = (V_0,V_1,z) \in \mathbf{Q}$ made to $\mathsf{Decide}_{\mathfrak{S}}^{\Psi_{\varphi_{\widehat{C}\mapsto C_0}}}$, let $\mathbf{C}_Q$ denote the set of critical queries corresponding to Q, given by Claim 4.6. We stress that $\Psi_{\varphi_{\widehat{C}\mapsto C_0}} \in \mathfrak{S}$; namely, it is a valid oracle in the family $\mathfrak{S}$, so the claim can be applied. (This is where we rely on the fact that $\mathfrak{S}$ includes *nonverifiable* IO constructions, where Eval may produce arbitrary answers on invalid obfuscations such as $\widehat{C} \notin \mathrm{Image}(\mathcal{O})$.)

We denote by $\mathbf{C_Q}$ the union of all such critical sets. We now define the event $\mathbf{rHit}$, aimed at capturing the cases when the adversary's views in $\mathbf{H_3}$ and $\mathbf{H_4}$ may differ. Concretely, let $\mathbf{rHit} = \mathbf{rHit}(r)$ be the event that one of the following occurs:

1. $(C_b,r) \in \mathbf{Q} \cup \mathbf{C_Q}$: the query (C_b,r) is made to $\mathcal{O}$.
2. $(\mathcal{O}(C_b,r),x) \in \mathbf{Q} \cup \mathbf{C_Q}$ for some $x \in \{0,1\}^n$: the query $(\mathcal{O}(C_b,r),x)$ is made to $\mathsf{Eval}_{\varphi_{\widehat{C}\mapsto C_0}}$.

Claim 4.22. $\Pr_r[\mathcal{A}_2^{\Gamma(f,\mathcal{O},\varphi_{\widehat{C}\mapsto C_0})}(\mathsf{state},\widehat{C}) \neq \mathcal{A}_2^{\Gamma(f,\mathcal{O}_{(C_b,r)\mapsto\widehat{C}},\varphi)}(\mathsf{state},\widehat{C})] \le \Pr_r[\mathbf{rHit}]$.

Proof. We observe that if $\mathbf{rHit}$ does not occur, then the view of $\mathcal{A}_2$ is the same for both oracles. Indeed, in $\mathbf{H_4}$, the oracle $\mathcal{O}_{(C_b,r)\mapsto\widehat{C}}$ is changed to $\mathcal{O}$ and φ is changed to $\varphi_{\widehat{C}\mapsto C_0}$. In particular,

- for any query $Q \neq (C_b,r)$ to $\mathcal{O}$
$$\mathcal{O}_{(C_b,r)\mapsto\widehat{C}}(Q) = \mathcal{O}(Q),$$
- for any query $Q = (\widehat{D},x) \notin \{(\widehat{C},x),(\mathcal{O}(C_b,r),x)\}_{x\in\{0,1\}^n}$ to $\mathsf{Eval}_{\varphi_{\widehat{C}\mapsto C_0}}^{f,\mathcal{O}}$,
$$\mathsf{Eval}_{\varphi}^{f,\mathcal{O}_{(C_b,r)\mapsto\widehat{C}}}(Q) = \mathsf{Eval}_{\varphi_{\widehat{C}\mapsto C_0}}^{f,\mathcal{O}}(Q),$$

since $(\mathcal{O}_{(C_b,r)\mapsto\widehat{C}})^{-1}(Q) = \mathcal{O}^{-1}(Q)$ and $\varphi(\widehat{D}) = \varphi_{\widehat{C}\mapsto C_0}(\widehat{D})$,

- for any $x \in \{0,1\}^n$ and query $Q = (\widehat{C}, x)$ to $\mathsf{Eval}^{f,\mathcal{O}}_{\varphi_{\widehat{C}\mapsto C_0}}$,

$$\mathsf{Eval}^{f,\mathcal{O}_{(C_b,r)\mapsto\widehat{C}}}_{\varphi}(\widehat{C},x) = C^f_b(x) = C^f_0(x) = (\varphi_{\widehat{C}\mapsto C_0}(\widehat{C}))^f(x) = \mathsf{Eval}^{f,\mathcal{O}}_{\varphi_{\widehat{C}\mapsto C_0}}(\widehat{C},x),$$

since the circuits are functionally equivalent: $C^f_0 \equiv C^f_1$.

It follows that for all queries $Q \in \mathbf{Q} \cup \mathbf{C_Q}$, $\Psi_{\varphi_{\widehat{C}\mapsto C_0}}(Q) = \Psi'_\varphi(Q)$, where

$$\Psi_{\varphi_{\widehat{C}\mapsto C_0}} = f, \mathcal{O}, \mathsf{Eval}^{f,\mathcal{O}}_{\varphi_{\widehat{C}\mapsto C_0}} \quad \text{and} \quad \Psi'_\varphi = f, \mathcal{O}_{(C_b,r)\mapsto\widehat{C}}, \mathsf{Eval}^{f,\mathcal{O}_{(C_b,r)\mapsto\widehat{C}}}_{\varphi}.$$

This implies that all queries made by $\mathcal{A}_2$ directly to its oracle $\Psi_{\varphi_{\widehat{C}\mapsto C_0}}$ in $\mathbf{H_4}$ are answered in the same way in $\mathbf{H_3}$. It is left to show that this is also the case for queries made to $\mathsf{Decide}^{\Psi_{\varphi_{\widehat{C}\mapsto C_0}}}_{\mathfrak{S}}$. For this purpose, note that $\Psi'_\varphi \in \mathfrak{S}$; namely, it is a valid oracle (indeed, $\widehat{C}$ is chosen outside the image of $\mathcal{O}$, so injectivity of the obfuscation oracle is guaranteed as required). Furthermore, the oracles $\Psi_{\varphi_{\widehat{C}\mapsto C_0}}$ and Ψ'_φ agree on all critical sets $\mathbf{C}_Q \subseteq \mathbf{C_Q}$. It follows, by Claim 4.6, that for any $Q \in \mathbf{Q}$

$$\mathsf{Decide}^{\Psi_{\varphi_{\widehat{C}\mapsto C_0}}}_{\mathfrak{S}}(Q) = \mathsf{Decide}^{\Psi'_\varphi}_{\mathfrak{S}}(Q),$$

which completes the proof of the claim. ∎

It is left to bound the probability that **rHit** occurs. Since r is chosen at random from $\{0,1\}^n$, for any fixed query $Q \in \mathbf{Q} \cup \mathbf{C_Q}$, the probability that any of the two cases defining **rHit** occurs is at most 2^{-n}. Thus, by a union bound we have

$$\Pr[\mathbf{Hit}] \le |\mathbf{Q} \cup \mathbf{C_Q}| \cdot 2^{-n} \le \left(|\mathbf{Q}| + \sum_{Q\in\mathbf{Q}} |\mathbf{C}_Q|\right) \cdot 2^{-n} \le O(q^2 \cdot 2^{-n}) \le O(2^{-n/3}).$$

This completes the proof of Claim 4.21. ∎

We now observe as follows.

Claim 4.23. *The adversary has no advantage in* $\mathbf{H_4}$*. That is, for every* $f, \mathcal{O}$*,*

$$\Pr[\mathcal{A} \textit{ wins in } \mathbf{H_4}] = \frac{1}{2}.$$

Proof. The view of $\mathcal{A}$ in this hybrid is completely independent of the random choice of b. ∎

This completes the proof of Theorem 4.16. ∎

Acknowledgments. We are grateful to the SICOMP reviewers for their careful reading and helpful comments; they have greatly contributed to the quality of this write-up. We also thank the reviewers of FOCS and CRYPTO for their valuable comments. We also thank Gil Segev, Iftach Haitner, and Mohammad Mahmoody for elaborately answering our questions regarding existing separation results in cryptography. We also thank Gil for helpful discussions regarding different aspects of this work.

REFERENCES

[ABW10] B. Applebaum, B. Barak, and A. Wigderson, *Public-key cryptography from different assumptions*, in Proceedings of the 42nd ACM Symposium on Theory of Computing, Cambridge, MA, 2010 pp. 171–180.

[AGGM06] A. Akavia, O. Goldreich, S. Goldwasser, and D. Moshkovitz, *On basing one-way functions on np-hardness*, in Proceedings of the 38th Annual ACM Symposium on Theory of Computing, Seattle, 2006, pp. 701–710.

[AH91] W. Aiello and J. Hastad, *Statistical zero-knowledge languages can be recognized in two rounds*, J. Comput. System Sci., 42 (1991), pp. 327–345.

[Ale03] M. Alekhnovich, *More on average case vs approximation complexity*, in Proceedings of the 44th Symposium on Foundations of Computer Science, Cambridge, MA, 2003, pp. 298–307.

[AR04] D. Aharonov and O. Regev, *Lattice problems in NP cap coNP*, in Proceedings of the 45th Symposium on Foundations of Computer Science, Rome, IEEE Computer Society, 2004, pp. 362–371.

[AS15] G. Asharov and G. Segev, *Limits on the power of indistinguishability obfuscation and functional encryption*, in Proceedings of the Symposium on the Foundations of Computer Science, 2015.

[AS16] G. Asharov and G. Segev, *On constructing one-way permutations from indistinguishability obfuscation*, in Theory of Cryptography, Springer, New York, 2016, pp. 512–541.

[Bar13] B. Barak, *Structure vs. Combinatorics in Computational Complexity*, http://windowsontheory.org/2013/10/07/structure-vs-combinatorics-in-computational-complexity/, 2013.

[BB15] A. Bogdanov and C. Brzuska, *On basing size-verifiable one-way functions on NP-hardness*, in Theory of Cryptography, Part I, Y. Dodis and J. Buus Nielsen, eds., Lecture Notes in Comput. Sci. 9014, Springer, New York, 2015, pp. 1–6.

[BBF13] P. Baecher, C. Brzuska, and M. Fischlin, *Notions of black-box reductions, revisited*, in Proceedings of Advances in Cryptology, 19th International Conference on the Theory and Application of Cryptology and Information Security, Bengaluru, India, Part I, 2013, pp. 296–315.

[BD19] N. Bitansky and A. Degwekar, *On the complexity of collision resistant hash functions: New and old black-box separations*, in Proceedings of 17th International Conference on Theory of Cryptography, Nuremberg, Part I, 2019, pp. 422–450.

[BG11] Z. Brakerski and O. Goldreich, *From absolute distinguishability to positive distinguishability*, in Studies in Complexity and Cryptography. Miscellanea on the Interplay Between Randomness and Computation, Springer, New York, 2011, pp. 141–155.

[BGI+01] B. Barak, O. Goldreich, R. Impagliazzo, S. Rudich, A. Sahai, S. P. Vadhan, and K. Yang, *On the (im)possibility of obfuscating programs*, in Advances in Cryptology, Joe Kilian, ed., Lecture Notes in Comput. Sci. 2139, Springer, New York, 2001, pp. 1–18.

[BGL+15] N. Bitansky, S. Garg, H. Lin, R. Pass, and S. Telang, *Succinct randomized encodings and their applications*, in Proceedings of the Symposium on Theory of Computing, 2015.

[BHKY19] N. Bitansky, I. Haitner, I. Komargodski, and E. Yogev, *Distributional collision resistance beyond one-way functions*, in Proceedings of Advances in Cryptology, 38th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Darmstadt, Part III, 2019, pp. 667–695.

[BHZ87] R. B. Boppana, J. Hastad, and S. Zachos, *Does co-NP have short interactive proofs?*, Inform. Process. Lett., 25 (1987), pp. 127–132.

[BI87] M. Blum and R. Impagliazzo, *Generic oracles and oracle classes*, in Proceedings of the 28th Annual Symposium on Foundations of Computer Science, Washington, DC, IEEE Computer Society, 1987, pp. 118–126.

[BKSY11] Z. Brakerski, J. Katz, G. Segev, and A. Yerukhimovich, *Limits on the power of zero-knowledge proofs in cryptographic constructions*, in Theory of Cryptography, Lecture Notes in Comput. Sci. 6597, Springer, New York, 2011, pp. 559–578.

[BL13] A. Bogdanov and C. H. Lee, *Limits of provable security for homomorphic encryption*, in Advances in Cryptology, Part I, R. Canetti and J. A. Garay, eds., Lecture Notes in Comput. Sci. 8042, Springer, New York, 2013, pp. 111–128.

[BM09] B. Barak and M. Mahmoody-Ghidary, *Merkle puzzles are optimal—an $o(n^2)$-query attack on any key exchange from a random oracle*, in Advances in Cryptology, Lecture Notes in Comput. Sci. 5677, S. Halevi, ed., Springer, New York, 2009, pp. 374–390.

[BP15] N. Bitansky and O. Paneth, *ZAPs and non-interactive witness indistinguishability from indistinguishability obfuscation*, in Theory of Cryptography, Part II, Y. Dodis and J. B. Nielsen, eds., Lecture Notes in Comput. Sci. 9015, Springer, New York, pp. 401–427.

[BPR15] N. Bitansky, O. Paneth, and A. Rosen, *On the cryptographic hardness of finding a nash equilibrium*, in Proceedings of the IEEE 56th Annual Symposium on Foundations of Computer Science, Berkeley, CA, V. Guruswami, ed., IEEE Computer Society, 2015, pp. 1480–1498.

[BPW16] N. Bitansky, O. Paneth, and D. Wichs, *Perfect structure on the edge of chaos—trapdoor permutations from indistinguishability obfuscation*, in Proceedings of 13th International Conference on Theory of Cryptography, Tel Aviv, Proceedings, Part I, 2016, pp. 474–502.

[Bra79] G. Brassard, *Relativized cryptography*, in Proceedings of the 20th Annual Symposium on Foundations of Computer Science, San Juan, Puerto Rico, IEEE Computer Society, 1979, pp. 383–391.

[BT03] A. Bogdanov and L. Trevisan, *On worst-case to average-case reductions for NP problems*, in Proceedings of the 44th Symposium on Foundations of Computer Science, Cambridge, MA, 2003, pp. 308–317.

[BV11] Z. Brakerski and V. Vaikuntanathan, *Efficient fully homomorphic encryption from (standard) LWE*, SIAM J. Comput., 43 (2012), pp. 831–871.

[CDT09] X. Chen, X. Deng, and S.-H. Teng, *Settling the complexity of computing two-player nash equilibria*, J. ACM, 56 (2009).

[CHJV15] R. Canetti, J. Holmgren, A. Jain, and V. Vaikuntanathan, *Succinct garbling and indistinguishability obfuscation for RAM programs*, in Proceedings of the 47th Annual ACM on Symposium on Theory of Computing, Portland, OR, 2015, pp. 429–437.

[Cra12] R. Cramer, ed., *Theory of Cryptography*, Lecture Notes in Comput. Sci. 7194, Springer, New York, 2012.

[DBL00] *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, Redondo Beach, CA, IEEE Computer Society, 2000.

[DBL03] *Proceedings of the 44th Annual Symposium on Foundations of Computer Science*, Cambridge, MA, IEEE Computer Society, 2003.

[DGP06] C. Daskalakis, P. W. Goldberg, and C. H. Papadimitriou, *The complexity of computing a nash equilibrium*, in Proceedings of the 38th Annual ACM Symposium on Theory of Computing, Seattle, 2006, pp. 71–78.

[DH76] W. Diffie and M. E. Hellman, *New directions in cryptography*, IEEE Trans. Inform. Theory, 22 (1976), pp. 644–654.

[DHT12] Y. Dodis, I. Haitner, and A. Tentes, *On the instantiability of hash-and-sign RSA signatures*, in Theory of Cryptography, R. Cramer, ed., Lecture Notes in Comput. Sci. 7194, Springer, New York, 2012, pp. 112–132.

[DLMM11] D. Dachman-Soled, Y. Lindell, M. Mahmoody, and T. Malkin, *On the black-box complexity of optimally-fair coin tossing*, in Theory of Cryptography, Lecture Notes in Comput. Sci. 6597, Springer, New York, 2011, pp. 450–467.

[DN00] C. Dwork and M. Naor, *Zaps and their applications*, in Proceedings 41st Annual Symposium on Foundations of Computer Science, IEEE, 2000.

[ESY84] S. Even, A. L. Selman, and Y. Yacobi, *The complexity of promise problems with applications to public-key cryptography*, Inform. Control, 61 (1984), pp. 159–173.

[Fis12] M. Fischlin, *Black-box reductions and separations in cryptography*, in Progress in Cryptology, A. Mitrokotsa and S. Vaudenay, eds., Lecture Notes in Comput. Sci. 7374, Springer, New York, 2012, pp. 413–422.

[For89] L. J. Fortnow, *Complexity-Theoretic Aspects of Interactive Proof Systems*, Ph.D. thesis, Massachusetts Institute of Technology, 1989.

[Gen09] C. Gentry, *Fully homomorphic encryption using ideal lattices*, in Proceedings of STOC, 2009, pp. 169–178.

[GG98] O. Goldreich and S. Goldwasser, *On the possibility of basing cryptography on the assumption that $\mathcal{P} \neq \mathcal{NP}$*, IACR Cryptology ePrint Archive, 1998:5, 1998.

[GGH+13a] S. Garg, C. Gentry, S. Halevi, M. Raykova, A. Sahai, and B. Waters, *Candidate indistinguishability obfuscation and functional encryption for all circuits*, in

Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science, Berkeley, CA, IEEE, 2013, pp. 40–49.

[GGH+13b] S. Garg, C. Gentry, S. Halevi, A. Sahai, M. Raikova, and B. Waters, *Candidate indistinguishability obfuscation and functional encryption for all circuits*, in Proceedings of FOCS, 2013.

[GGKT05] R. Gennaro, Y. Gertner, J. Katz, and L. Trevisan, *Bounds on the efficiency of generic cryptographic constructions*, SIAM J. Comput., 35 (2005), pp. 217–246.

[GK93] O. Goldreich and E. Kushilevitz, *A perfect zero-knowledge proof system for a problem equivalent to the discrete logarithm*, J. Cryptology, 6 (1993), pp. 97–116.

[GKLM12] V. Goyal, V. Kumar, S. V. Lokam, and M. Mahmoody, *On black-box reductions between predicate encryption schemes*, in Theory of Cryptography, R. Cramer, ed., Lecture Notes in Comput. Sci. 7194, Springer, New York, 2012, pp. 440–457.

[GKM+00] Y. Gertner, S. Kannan, T. Malkin, O. Reingold, and M. Viswanathan, *The relationship between public key encryption and oblivious transfer*, in Proceedings of the 41st Annual Symposium on Foundations of Computer Science, Redondo Beach, CA, 2000, pp. 325–335.

[GM82] S. Goldwasser and S. Micali, *Probabilistic encryption and how to play mental poker keeping secret all partial information*, in Proceedings of the 14th Annual ACM Symposium on Theory of Computing, H. R. Lewis, B. B. Simons, W. A. Burkhard, and L. H. Landweber, eds., San Francisco, ACM, 1982, pp. 365–377.

[GMM07] Y. Gertner, T. Malkin, and S. Myers, *Towards a separation of semantic and CCA security for public key encryption*, in Theory of Cryptography, S. P. Vadhan, ed., Lecture Notes in Comput. Sci. 4392, Springer, New York, 2007, pp. 434–455.

[GMM17] S. Garg, M. Mahmoody, and A. Mohammed, *Lower bounds on obfuscation from all-or-nothing encryption primitives*, in Proceedings of Advances in Cryptology, 37th Annual International Cryptology Conference, Santa Barbara, CA, Part I, 2017, pp. 661–695.

[GMR85] S. Goldwasser, S. Micali, and C. Rackoff, *The knowledge complexity of interactive proof-systems (extended abstract)*, in Proceedings of the 17th Annual ACM Symposium on Theory of Computing, R. Sedgewick, ed., ACM, Providence, RI, 1985, pp. 291–304.

[GMR01] Y. Gertner, T. Malkin, and O. Reingold, *On the impossibility of basing trapdoor functions on trapdoor predicates*, in Proceedings of the 42nd Annual Symposium on Foundations of Computer Science, Las Vegas, NV, IEEE Computer Society, 2001, pp. 126–135.

[GMW91] O. Goldreich, S. Micali, and A. Wigderson, *Proofs that yield nothing but their validity for all languages in NP have zero-knowledge proof systems*, J. ACM, 38 (1991), pp. 691–729.

[Gol06] O. Goldreich, *On promise problems: A survey*, in Theoretical Computer Science: Essays in Memory of Shimon Even, Lecture Notes in Comput. Sci. 3895, Springer, New York, 2006, pp. 254–290.

[GT00] R. Gennaro and L. Trevisan, *Lower bounds on the efficiency of generic cryptographic constructions*, in Proceedings of the 41st Annual Symposium on Foundations of Computer Science, Redondo Beach, CA, pp. 305–313.

[GV99] O. Goldreich and S. P. Vadhan, *Comparing entropies in statistical zero knowledge with applications to the structure of SZK*, in Proceedings of the 14th Annual IEEE Conference on Computational Complexity, Atlanta, GA, 1999, p. 54.

[Has88] J. Hastad, *Dual vectors and lower bounds for the nearest lattice point problem*, Combinatorica, 8 (1988), pp. 75–81.

[HH09] I. Haitner and T. Holenstein, *On the (im)possibility of key dependent encryption*, in Theory of Cryptography, O. Reingold, ed., Lecture Notes in Comput. Sci. 5444, Springer, New York, 2009, pp. 202–219.

[HHRS15] I. Haitner, J. J. Hoch, O. Reingold, and G. Segev, *Finding collisions in interactive protocols—tight lower bounds on the round and communication complexities of statistically hiding commitments*, SIAM J. Comput., 44 (2015), pp. 193–242.

[HR04] C.-Y. Hsiao and L. Reyzin, *Finding collisions on a public road, or do secure hash functions need secret coins?*, in Proceedings of Advances in Cryptology, 24th Annual International CryptologyConference, Santa Barbara, CA, 2004, pp. 92–105.

[IR89] R. Impagliazzo and S. Rudich, *Limits on the provable consequences of one-way permutations*, in Proceedings of the 21st Annual ACM Symposium on Theory of Computing, ACM, 1989, pp. 44–61.

[Ish11] Y. Ishai, ed., *Theory of Cryptography*, Lecture Notes in Comput. Sci. 6597, Springer, New York, 2011.

[Kle06] J. M. Kleinberg, ed., *Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, Seattle, WA, ACM, New York, 2006.

[KLW15] V. Koppula, A. Bishop Lewko, and B. Waters, *Indistinguishability obfuscation for turing machines with unbounded memory*, in Proceedings of the 47th Annual ACM on Symposium on Theory of Computing, Portland, OR, 2015, pp. 419–428.

[KMN+14] I. Komargodski, T. Moran, M. Naor, R. Pass, A. Rosen, and E. Yogev, *One-way functions and (im)perfect obfuscation*, in Proceedings of the 55th IEEE Annual Symposium on Foundations of Computer Science, Philadelphia, PA, IEEE Computer Society, 2014, pp. 374–383.

[KSS11] J. Kahn, M. E. Saks, and C. D. Smyth, *The dual BKR inequality and Rudich's conjecture*, Combin. Probab. Comput., 20 (2011), pp. 257–266.

[KST99] J. H. Kim, D. R. Simon, and P. Tetali, *Limits on the efficiency of one-way permutation-based hash functions*, in Proceedings of the 40th Annual Symposium on Foundations of Computer Science, New York, IEEE Computer Society, 1999, pp. 535–542.

[KY18] I. Komargodski and E. Yogev, *On distributional collision resistant hashing*, in Proceedings of CRYPTO, 2018.

[LLJS90] J. C. Lagarias, H. W. Lenstra, Jr., and C.-P. Schnorr, *Korkin-Zolotarev bases and successive minima of a lattice and its reciprocal lattice*, Combinatorica, 10 (1990), pp. 333–348.

[LV16] T. Liu and V. Vaikuntanathan, *On basing private information retrieval on np-hardness*, in Theory of Cryptography, E. Kushilevitz and T. Malkin, eds., Part I, Lecture Notes in Comput. Sci. 9562, Springer, New York, 2016, pp. 372–386.

[MM11] T. Matsuda and K. Matsuura, *On black-box separations among injective one-way functions*, in Theory of Cryptography, Springer, New York, 2011, pp. 597–614.

[MP91] N. Megiddo and C. H. Papadimitriou, *On total functions, existence theorems and computational complexity*, Theoret. Comput. Sci., 81 (1991), pp. 317–324.

[MV03] D. Micciancio and S. P. Vadhan, *Statistical zero-knowledge proofs with efficient provers: Lattice problems and more*, in Advances in Cryptology, D. Boneh, ed., Lecture Notes in Comput. Sci. 2729, Springer, New York, 2003, pp. 282–298.

[Ost91] R. Ostrovsky, *One-way functions, hard on average problems, and statistical zero-knowledge proofs*, in Proceedings of the 6th Annual Structure in Complexity Theory Conference, IEEE, 1991, pp. 133–138.

[OV08] S. Jin Ong and S. P. Vadhan, *An equivalence between zero knowledge and commitments*, in Proceedings of Theory of Cryptography, New York, 2008, pp. 482–500.

[Pap94] C. H. Papadimitriou, *On the complexity of the parity argument and other inefficient proofs of existence*, J. Comput. System Sci., 48 (1994), pp. 498–532.

[Pas06] R. Pass, *Parallel repetition of zero-knowledge proofs and the possibility of basing cryptography on NP-hardness*, in Proceedings of the 21st Annual IEEE Conference on Computational Complexity, Prague, Czech Republic, IEEE Computer Society, 2006, pp. 96–110.

[Pas13] R. Pass, *Unprovable security of perfect NIZK and non-interactive non-malleable commitments*, in Proceedings of the Theory of Cryptography Conference, 2013, pp. 334–354.

[RAD78] R. Rivest, L. Adleman, and M. Dertouzos, *On data banks and privacy homomorphisms*, in Foundations of Secure Computation, Academic Press, New York, 1978, pp. 169–177.

[RSA78] R. L. Rivest, A. Shamir, and L. M. Adleman, *A method for obtaining digital signatures and public-key cryptosystems*, Commun. ACM, 21 (1978), pp. 120–126.

[RSS16] A. Rosen, G. Segev, and I. Shahaf, *Can PPAD hardness be based on standard cryptographic assumptions?*, Electronic Colloquium on Computational Complexity 23:59, 2016.

[RTV04] O. Reingold, L. Trevisan, and S. P. Vadhan, *Notions of reducibility between cryptographic primitives*, in Proceedings of Theory of Cryptography, Cambridge, MA, 2004, pp. 1–20.

[Rud88] S. Rudich, *Limits on the Provable Consequences of One-Way Functions*, Ph.D. thesis, University of California, Berkeley, 1988.

[Rud91] S. Rudich, *The use of interaction in public cryptosystems (extended abstract)*, in Advances in Cryptology, J. Feigenbaum, ed., Lecture Notes in Comput. Sci. 576, Springer, New York, 1991, pp. 242–251.

[Sim98] D. R. Simon, *Finding collisions on a one-way street: Can secure hash functions be based on general assumptions?*, in Advances in Cryptology, Springer, 1998, pp. 334–345.

[SV03] A. Sahai and S. Vadhan, *A complete problem for statistical zero knowledge*, J. ACM, 50 (2003), pp. 196–249.

[SW14] A. Sahai and B. Waters, *How to use indistinguishability obfuscation: Deniable encryption, and more*, in Proceedings of the Symposium on Theory of Computing, D. B. Shmoys, ed., New York, ACM, 2014, pp. 475–484.

[Vad99] S. P. Vadhan, *A Study of Statistical Zero-Knowledge Proofs*, Ph.D. thesis, Massachusetts Institute of Technology, 1999.

[Wat15] B. Waters, *A punctured programming approach to adaptively secure functional encryption*, in Proceedings of Advances in Cryptology, Part II, 2015, pp. 678–697.