# Optimal Universal Black-Box Privacy Preservation and Tight Adversarially-Adaptive Composition for Entropic Secret

**Abstract.** Privacy entails a fundamental tradeoff between information propagation and preservation, and it is generally *not* free. A central goal of privacy research is to determine and, when possible, attain the minimum utility loss necessary to meet a desired privacy guarantee. Unfortunately, in the classic regime of *Input-Independent Indistinguishability* (III), as the foundation to protect *intractable a priori* secret, strong impossibility results are known that a privacy solution cannot be even validated in general. However, in the regime of randomly-generated secret with entropy, the problem, either from possibility and impossibility side, largely remains open. In this paper, we provide an affirmative answer in a more restrictive *black-box* setting: we assume no structural knowledge of the leakage-generation procedure, and the user is granted only limited access via end-to-end simulation. Stemmed from *PAC Privacy* and *$\alpha$-Mutual Information*, we establish a universal framework that *automatically* and *asymptotically* approaches the optimal randomization strategy for an arbitrary *black-box* leakage function, given a Bayesian posterior budget against an *arbitrary* adversarial inference procedure.

We further generalize our results to the composition scenario, in which a random secret is reused across a sequence of potentially adversarially selected leakage applications. Still in a *black-box* setting, with respect to both adversarial strategy and leakage functions, we present a *locally optimal* randomization strategy and an asymptotically tight characterization of the cumulative privacy risk. We also provide a mechanism-agnostic variant by establishing a general composition upper bound using only marginal Bayesian risk measurements. Our results have strong practical potential for mitigating side-channel leakage, especially from physical channels that lack tractable analytical models. We include experiments on mitigating leakage from power consumption during Advanced Encryption Standard (AES) secret-key generation and from timing measurements in RSA encryption. The code and the full version with appendix can be found in the following anonymous Github link `https://anonymous.4open.science/r/Optimal_Blackbox_Privacy-812F/`.

**Keywords:** Black-Box Privacy Analysis, Adversarial Adaptive Composition, Secret Entropy, PAC Privacy, $\alpha$-Mutual Information

## 1 Introduction

Any processing of a secret can potentially leak information. As one of the most extensively-studied security problems, privacy essentially asks the following: *given the information revealed through leakage, how much of the underlying*

*secret can an adversary recover?* Despite its intuitive interpretation, rigorous formalization and provable protection are highly non-trivial for many practical applications. At a high level, modern privacy research can be organized along three main directions: ❶ *definition-level*—how to characterize privacy risk and privacy guarantees, both *semantically* and *mathematically*; ❷ *measurement/accounting-level*—how to *tightly* quantify the privacy risk induced by a given leakage; and ❸ *operation-level*—how to construct *optimal* mitigations for a given leakage to ensure a required privacy guarantee.

### 1.1 Privacy Definition

On ❶, existing privacy definitions can be largely categorized by two types of secrets to protect: **(A)** entropic secrets sampled from an *accessible distribution*[1], and **(B)** secrets for which such entropy is not tractable. A canonical example of a Type-(A) secret is a uniformly-generated secret key in a cryptographic protocol. For Type-(A) secrets, a natural notion of privacy is Bayesian: the adversary's optimal posterior success probability for recovering the secret [44]. In particular, when the adversary's objective is *identification* (i.e., the success criterion is exact recovery of the secret), this risk coincides with the classic notion of (conditional) *min-entropy* studied in [7,34,22]. Related notions include *guessing entropy* [26], which captures the expected number of guesses required for successful identification.

A key prerequisite of the Bayesian viewpoint above is the ability to specify a prior distribution over the secret $\mathsf{x}$, which is often inappropriate for Type-(B) secrets where we may not know, or may not have access to, the secret-generation mechanism[2]. To obtain a rigorous notion in a *prior-free* setting, Shannon [36], in 1949, initiated the concept of *input-independent indistinguishability* (III), which essentially considers the worst case over all priors. Rather than directly measuring the adversary's posterior knowledge, III evaluates the worst-case *posterior advantage*: the maximal change from prior to posterior when distinguishing between two secret hypotheses under an arbitrary prior distribution. Goldwasser and Micali [15] further introduced a computational form of III against computationally-bounded adversaries, laying the foundation of modern cryptography. They also showed that negligible III for binary hypotheses in the worst case implies negligible advantage for *any* efficient inference task. As a prior-free worst-case notion, III admits a clear semantic interpretation: regardless of the adversary's (subjective) prior belief about the secret, observing the leakage should not significantly change their posterior belief. Beyond cryptography, many information-theoretic privacy definitions for Type-(B) secrets, including *Differential Privacy* (DP) [10,9], *Pufferfish Privacy* [21], and *Maximal Leakage* [18], are also rooted in the III philosophy.

[1] An *accessible distribution* refers to a secret-generation procedure from which we can repeatedly sample instances of the secret.

[2] Even if one could posit that all possible data is generated by some complicated stochastic process, in many real scenarios the observation is effectively unique and cannot be resampled at the relevant moment.

### 1.2 Universal Privacy Risk Measurement and Leakage Mitigation

The ❷ measurement problem and the ❸ mitigation problem require a deeper study of the leakage itself. Broadly speaking, leakage can be any quantity statistically correlated with a secret $\mathsf{x}$ that contains sensitive information, and it can be generally represented as the output of a function $\mathcal{F}(\mathsf{x})$. In practice, leakage $\mathcal{F}(\mathsf{x})$ may arise from the release of any algorithmic processing of sensitive data $\mathsf{x}$, for instance, a machine learning model trained on health data [2]. Leakage can also arise as a physical signal related to the processing of $\mathsf{x}$: a large body of work has identified many side channels, such as timing [6,48], power [33], memory [3] consumption, or network traffic patterns [41], that may enable successful attacks.

To mitigate leakage (❸), most existing approaches introduce additional randomization into $\mathcal{F}(\mathsf{x})$. A widely used strategy is perturbation, releasing a noisy version $\mathcal{F}(\mathsf{x})+\mathsf{e}$. When $\mathcal{F}(\mathsf{x})$ is a statistic (e.g., a mean salary estimate or neural-network weights), $\mathsf{e}$ may be independent Gaussian [9] or Laplace [10] noise, as is standard in the DP literature. For physical signals $\mathcal{F}(\mathsf{x})$, $\mathsf{e}$ may also be introduced physically, e.g., by sending dummy queries in anonymous communication systems [41]. In any case, privacy mechanisms typically trade utility for protection. The appropriate utility loss measure depends on the application: in anonymous communication, for example, the loss may be the delay introduced by dummy messages, with the relevant metric determined by latency sensitivity or protocol constraints.

Given the complexity of practical leakage functions $\mathcal{F}(\cdot)$ and the diversity of utility metrics, it is desirable to have a *universal* framework that addresses ❷ (tight measurement) and ❸ (optimal mitigation) simultaneously. Unfortunately, in the Type-(B) regime, neither tight measurement nor optimal mitigation can be efficiently determined under III-style definitions in general. Indeed, Xiao and Tao [46] show that computing sensitivity, the maximal outcome difference

$$\sup_{\bar{\mathsf{x}},\bar{\mathsf{x}}'} \|\mathcal{F}(\bar{\mathsf{x}}) - \mathcal{F}(\bar{\mathsf{x}}')\|$$

over arbitrary secret hypotheses $(\bar{\mathsf{x}},\bar{\mathsf{x}}')$, is NP-hard in the worst case. This suggests that even *verifying* whether a leakage function $\mathcal{F}(\cdot)$ satisfies a given III-style guarantee can be computationally intractable.

As a consequence, known III-style privacy analysis for Type-(B) secrets typically apply only in specific *white-box* settings, and one of most representative methodologies is *decompose-then-compose*. In particular, over the last two decades the DP literature has developed a rich collection of such tools, including *subsample-then-aggregate* [29], the *Sparse Vector Technique* [16,25], *Differentially Private SGD* (DP-SGD) [1], PATE [30], and *private evolution* [23]. In these frameworks, a complex processing pipeline is decomposed into simpler components[3], each of which is separately privatized and then composed. However,

[3] Ideally, each component in a *decompose-then-compose* framework ensures bounded sensitivity $\sup_{\bar{\mathsf{x}},\bar{\mathsf{x}}'} \|\mathcal{F}(\bar{\mathsf{x}}) - \mathcal{F}(\bar{\mathsf{x}}')\| \leq c$ for some threshold $c$ under an appropriate

this methodology can lead to overly conservative (loose) noise requirements, especially for high-dimensional tasks with intricate dependencies among sensitive inputs, such as clustering [14] and outcomes from machine learning models [32]. These limitations can be even more severe for settings without tractable algorithmic structure, including many of the physical side channels [6,33,3] mentioned earlier.

In contrast, for Type-(A) entropic secrets, the landscape can be starkly different for both ❷ and ❸. While tightness and optimality remained open, recent work PAC Privacy [44] demonstrates the possibility of universal privacy analysis even in a more restricted *black-box* setting. Here, *black-box* means that we assume no particular algorithmic structure for $\mathcal{F}(\cdot)$; instead, the user only has oracle access to end-to-end evaluations of $\mathcal{F}(\cdot)$. In the context of ❷ and ❸, this means that both risk measurement and mechanism design must be derived from a finite number $m$ of evaluations of $\mathcal{F}(\cdot)$ on selected inputs $\mathsf{X} = \{\bar{\mathsf{x}}_i : i \in [m]\}$. Not surprisingly, the follow-up works [45,39] show that the provable Bayesian upper bounds derived from PAC Privacy can be significantly sharper than the prior-free worst case in many applications, after the secret's inherent entropy is exploited which makes inference intrinsically harder.

Fundamentally, this difference in addressing ❷ and ❸ with respect to Type-(A) and Type-(B) secrets arises because an III-style guarantee in Type-(B) scenario demands divergence control between leakages from *arbitrary* pairs of secret hypotheses.[4] This strict "arbitrary" requirement is the root cause of the computational impossibility [46]. For Type-(A) secrets, however, it suffices to only analyze Bayesian posteriors with respect to the *ground-truth* secret distribution $\mathcal{D}$ [20], since any mis-specified or biased prior always weakens the adversary's inference performance relative to the optimal Bayesian adversary. Thus, even if extreme worst cases exist and are hard to identify, they may be negligible under $\mathcal{D}$, enabling near-optimal Bayesian analysis. Moreover, higher entropy in $\mathsf{x}$ (more uncertainty) should allow substantially cheaper privacy mechanisms than conservative prior-free worst-case approaches. We formalize ❷ and ❸ for Type-(A) entropic secrets in the black-box setting as follows.

**Problem 1 (Tight black-box privacy analysis with entropy).** Given a prior distribution $\mathcal{D}$ of an entropic secret $\mathsf{x}$ and an arbitrary leakage function $\mathcal{F}(\cdot) : \mathcal{X} \to \mathcal{O}$, we want to design a (possibly randomized) perturbation-selection mechanism $\mathsf{Alg} : (\mathcal{X} \times \mathcal{O})^m \to \mathbb{P}$ that outputs a distribution $\mathcal{D}_e$ of perturbation $\mathsf{e}$ with *minimal overhead* (measured by some metric $\kappa(\cdot)$), such that for an arbitrary inference task captured by a criterion $\rho$ and target posterior success rate $\delta_\rho$, the following experiment is *impossible* for an informed adversary:

---

metric $\|\cdot\|$. In the DP context, a common structure is *aggregation*: each data point can be clipped to a bounded range and contributes *independently* to the released output [1].

[4] This is necessary if one wants to upper bound the posterior advantage under *arbitrary* priors [15].

1. The user selects $m$ strings $\mathsf{X} = \{\bar{\mathsf{x}}_i \in \mathcal{X} : i \in [m]\}$, evaluates $\mathcal{F}(\bar{\mathsf{x}}_i)$, and applies $\mathsf{Alg}$ to the simulation data $\mathsf{S} = \{(\bar{\mathsf{x}}_i, \mathcal{F}(\bar{\mathsf{x}}_i)) : i \in [m]\}$. The mechanism outputs a perturbation distribution $\mathsf{Alg}(\mathsf{S}) = \mathcal{D}_e$.
2. The user samples an $\mathsf{x} \sim \mathcal{D}$, samples perturbation $\mathsf{e} \sim \mathcal{D}_e$, and sends the noisy leakage $\mathcal{F}(\mathsf{x}) + \mathsf{e}$ to the adversary.
3. The adversary, who has full knowledge of $\mathcal{F}(\cdot)$ and the user's mechanism $\mathsf{Alg}$, outputs $\hat{\mathsf{x}}$ such that $\rho(\hat{\mathsf{x}}, \mathsf{x}) = 1$ with probability greater than $\delta_\rho$.

In Problem 1, we follow PAC Privacy [44] and model an arbitrary adversarial inference task via a criterion function $\rho(\cdot, \cdot)$, measuring privacy risk by the Bayesian posterior success probability $\delta_\rho$. The criterion $\rho$ can be chosen so that the preimage of $\rho(\cdot, \mathsf{x}) = 1$ captures all *unacceptable* reconstructions of $\mathsf{x}$. For instance, given an $l$-bit secreting string $\mathsf{x}$, $\rho(\hat{\mathsf{x}}, \mathsf{x}) = 1$ iff $\hat{\mathsf{x}} = \mathsf{x}$ captures exact identification; $\rho(\hat{\mathsf{x}}, \mathsf{x}) = 1$ iff the Hamming distance between $\hat{\mathsf{x}}$ and $\mathsf{x}$ is at most $c$ captures approximate reconstruction with at most $c$-bit errors. As for the utility-loss metric $\kappa$, a common choice is the variance of the injected noise $\mathsf{e}$; in general, one may determine a score for each output instance in $\mathcal{O}$ and selects $\kappa$ as the expected quality of the outcome from the perturbed mechanism $\mathcal{M}(\mathsf{x}) = \mathcal{F}(\mathsf{x}) + \mathsf{e}$.

On the other hand, it is worthwhile noting that the protection strategy $\mathcal{D}_e$, the distribution of the perturbation/modification $\mathsf{e}$ imposed to leakage $\mathcal{F}(\mathsf{x})$ can be dependent on the secret input $\mathsf{x}$, which, strictly speaking, should be written as $\mathcal{D}_e(\mathsf{x})$, although in most existing works [9,1], $\mathsf{e}$ is simply an *independent* (e.g., Gaussian) noise. For notation simplicity we still use $\mathcal{D}_e$ in the following. We also emphasize the adversarial model in the following remark.

*Remark 1 (White-Box Adversarial View).* Throughout the paper, the *black-box* assumption is from the user's perspective: the user lacks structural knowledge of how leakage is incurred. The adversary, in contrast, is always assumed to be fully informed of both the leakage function $\mathcal{F}$ and the user's privacy-preserving strategy $\mathsf{Alg}$. The only uncertainty from the adversary's perspective is the randomness in secret generation (e.g., $\mathsf{x} \sim \mathcal{D}$) and in the modified leakage process (e.g., $\mathsf{e} \sim \mathcal{D}_e$), over which the posterior success probability $\delta_\rho$ is defined.

### 1.3 Privacy Composition

So far, we have focused on a static leakage function $\mathcal{F}(\cdot)$. In practice, leakage may arise in a potentially adversarial interactive environment, and when the secret is reused, we must track its cumulative leakage. To model leakage in such a dynamic setting, we consider a more general function $\mathcal{F}(\mathsf{x}, \mathsf{v})$, where an environment parameter $\mathsf{v}$ can be adaptively selected by an adversary to maximize information gained through sequential leakages. For example, given a secret key $\mathsf{x}$, a user may be asked to encrypt multiple messages $\{\mathsf{v}_1, \mathsf{v}_2, \ldots, \mathsf{v}_T\}$, which can be adversarially chosen; each encryption produces leakage $\mathcal{F}(\mathsf{x}, \mathsf{v}_t)$. Intuitively, an adversary can adaptively choose the next query $\mathsf{v}_t$ based on accumulated information about $\mathsf{x}$, potentially increasing their posterior success probability.

Existing composition results primarily address Type-(B) secrets under III-style guarantees, and have largely been developed in the DP literature [11,19,5,27,8]. Tight composition accounting has also motivated new privacy definitions, evolving from pure $\epsilon$-DP to approximate $(\epsilon, \delta)$-DP, to concentrated DP [5] and Rényi DP [27], and most recently to $f$-DP [8]. Typically, these results rely on two key prerequisites: (i) a *global* worst-case III guarantee for all possible leakage functions $\mathcal{F}(\cdot, \mathsf{v})$ (independent of auxiliary information and adversarial adaptivity), and (ii) independence of the randomness used at each step of the composed mechanism: the global worst-case guarantee (i) bounds the posterior advantage increment regardless of the adversary's query, while independence in (ii) enables aggregation via concentration techniques.

For Type-(A) entropic secrets in the *black-box* regime, however, the situation changes dramatically and existing III-style composition analyses do *not* directly apply. First, in a black-box setting, we cannot generally obtain non-trivial input-independent guarantees from finitely many end-to-end simulations. Second, and more importantly, the entropic secret $\mathsf{x}$ is generated once and reused: even if fresh randomization is applied at each step, the resulting leakages remain correlated through the shared secret. We formalize the composition version of Problem 1 as follows.

**Problem 2 (Noise mechanism for tight composition).** Given a prior secret distribution $\mathcal{D}$, a privacy budget in terms of posterior success rate $\delta_\rho$ for an inference task captured by $\rho$, and a family of (black-box) leakage functions $\mathcal{F}(\cdot,\cdot) : \mathcal{X} \times \mathcal{A} \to \mathcal{O}$, we want to design a joint noise mechanism $\mathsf{Alg}$ such that there does *not* exist an adversarial query algorithm $\mathcal{Q}_{\text{adv}}$ for which the following experiment between a user (the secret holder) and the adversary is possible:

1. The user samples a secret $\mathsf{x} \sim \mathcal{D}$.
2. At iteration $t \in [T]$, from the adversary side, possibly based on all previously observed leakages $\{\mathsf{o}_j = \mathcal{F}(\mathsf{x}, \mathsf{v}_j) + \mathsf{e}_j : j \in [t-1]\}$ and past queries $\{\mathsf{v}_j : j \in [t-1]\}$, the adversary chooses a new query $\mathsf{v}_t \leftarrow \mathcal{Q}_{\text{adv}}(\{\mathsf{o}_j, \mathsf{v}_j\}_{j<t})$ and sends $\mathsf{v}_t$ to the user.
3. At iteration $t \in [T]$, from the user side, after receiving the adversarial query $\mathsf{v}_t$, the user performs end-to-end simulations of the leakage functions $\{\mathcal{F}(\cdot, \mathsf{v}_j) : j \in [t]\}$ on a selected input set $\{\bar{\mathsf{x}}_i^{(t)} : i \in [m]\}$, and applies $\mathsf{Alg}$ to the resulting evaluations

$$\mathsf{S}^{(t)} = \{(\bar{\mathsf{x}}_i^{(t)}, \mathcal{F}(\bar{\mathsf{x}}_i^{(t)}, \mathsf{v}_j)) : i \in [m],\, j \in [t]\},$$

   together with the previously sampled noises $\mathbf{e}^{(t-1)} = (\mathsf{e}_1, \mathsf{e}_2, \ldots, \mathsf{e}_{t-1})$, to determine a noise distribution $\mathcal{D}_e^{(t)}$. The user then samples $\mathsf{e}_t \sim \mathcal{D}_e^{(t)}$ and returns the noisy leakage $\mathsf{o}_t = \mathcal{F}(\mathsf{x}, \mathsf{v}_t) + \mathsf{e}_t$ to the adversary.
4. After $T$ iterations, the adversary outputs an estimate $\hat{\mathsf{x}}$ based on $\{(\mathsf{o}_t, \mathsf{v}_t)\}_{t\in[T]}$ such that $\rho(\hat{\mathsf{x}}, \mathsf{x}) = 1$ with probability greater than $\delta_\rho$.

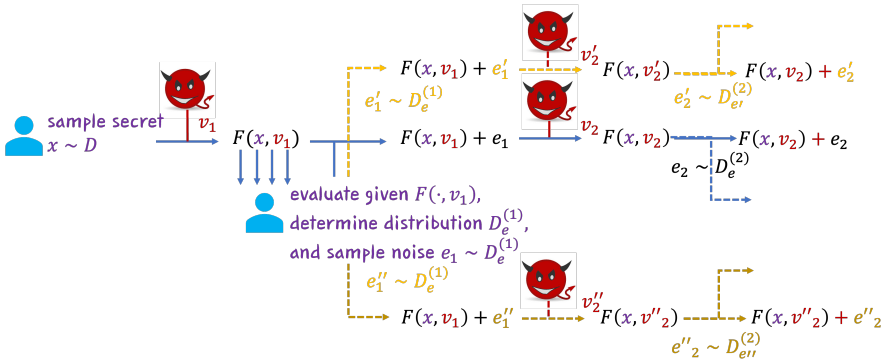


**Fig. 1.** Illustration of the *Parallel Universe Challenge* in adversarially adaptive composition for entropic secrets. The solid blue line corresponds to the specific (local) realization observed by the user, while dashed lines represent alternative (unobserved) realizations induced by randomness in both the secret and the mechanism.

Figure 1 illustrates the technical challenge behind Problem 2, which we refer to as the *parallel universe* restriction. With an adaptive adversary, the joint distribution of the composed leakages $\{\mathcal{F}(\mathsf{x}, \mathsf{v}_t)\}_{t=1}^{T}$ is jointly determined by the (unknown) adversarial strategy $\mathcal{Q}_{\text{adv}}$ and the user's actions (secret sampling and noise injection). Under the black-box model, the user cannot infer $\mathcal{Q}_{\text{adv}}$ or the induced global leakage distribution required for a direct Bayesian analysis. The user's only visibility into $\mathcal{Q}_{\text{adv}}$ is through the adversary's reaction *conditional* on the particular randomness realized in the user's own execution. Concretely, even conditioning on a first query $\mathsf{v}_1$ and a first-round noise distribution $\mathcal{D}_e^{(1)}$, the user observes only the adversary's next query $\mathsf{v}_2$ corresponding to the specific sampled noise instance $\mathsf{e}_1$. Different noise realizations $\mathsf{e}'_1 \sim \mathcal{D}_e^{(1)}$ could induce different adversarial branches—analogous to "parallel universes" inaccessible to the user. Problem 2 thus asks the user to infer and mitigate leakage from the *local* universe they observe, while ensuring that aggregating over all (unobserved) branches yields the desired global posterior guarantee.

Stemmed from Problem 2, a further generalization is to develop composition theory that is agnostic to the specific privacy-preserving mechanism used at each step. A natural application setting is one where multiple parties share a secret but are processing the secret separately, and each party lacks knowledge of the leakage functions and defenses used by others. That is, given $T$ black-box leakage functions $\{\mathcal{F}(\mathsf{x}, \mathsf{v}_t) : t \in [T]\}$ that share the same entropic secret $\mathsf{x}$, and given only that each marginal leakage satisfies certain posterior inference hardness with respect to $\mathsf{x}$, how can we upper bound the cumulative Bayesian privacy risk? This motivates the following problem.

**Problem 3 (Mechanism-agnostic composition).** Given a random secret $\mathsf{x} \sim \mathcal{D}$ and an arbitrary sequence of $T$ leakage functions $\{\mathcal{F}(\cdot, \mathsf{v}_t) : t \in [T]\}$, suppose that, conditioned on the shared secret input $\mathsf{x}$, all additional random-

ness in each $\mathcal{F}(\cdot, \mathsf{v}_t)$ is independent across $t \in [T]$. Under what types of *marginal* measurements for each $\mathcal{F}(\cdot, \mathsf{v}_t)$ can we derive a global bound on the cumulative Bayesian privacy risk?

### 1.4 Techniques Overview

In this section, we provide a roadmap of techniques to address Problem 1 - 3. We begin with a special case of Problem 1 where the inference task $\rho$ is fixed to identification (exact reconstruction). In this case, the adversary's optimal inference rule $\mathcal{A}(o)$ given an observation $o = \mathcal{F}(\mathsf{x})$ is the classical *Maximum A Posteriori* (MAP) estimator [47]:

$$\mathcal{A}(o) \in \arg\max_{\hat{\mathsf{x}}} \Pr_{\mathsf{x}\sim\mathcal{D}}\big(\mathsf{x} = \hat{\mathsf{x}} \mid \mathcal{F}(\mathsf{x}) = o\big), \tag{1}$$

i.e., the adversary outputs the hypothesis with maximal posterior probability.

Accordingly, the optimal identification success probability, denoted by $\delta_I$, can be written as an aggregate of weighted 0–1 losses:

$$\begin{aligned}
\delta_I &= \Pr_{\mathsf{x}\sim\mathcal{D}}\left[\mathcal{A}(\mathcal{F}(\mathsf{x})) = \mathsf{x}\right] \\
&= \sum_{\hat{\mathsf{x}}} \Pr_{\mathsf{x}\sim\mathcal{D}}(\mathsf{x} = \hat{\mathsf{x}}) \cdot \int_o \mathbf{1}\Big[\hat{\mathsf{x}} \in \arg\max_z \Pr_{\mathsf{x}\sim\mathcal{D}}\big(\mathsf{x} = z \mid \mathcal{F}(\mathsf{x}) = o\big)\Big] \cdot \Pr(\mathcal{F}(\hat{\mathsf{x}}) = o)\,\mathrm{d}o \\
&= \mathbb{E}_{\hat{\mathsf{x}}\sim\mathcal{D}}\left[\int_o \mathbf{1}\Big[\hat{\mathsf{x}} \in \arg\max_z \Pr_{\mathsf{x}\sim\mathcal{D}}\big(\mathsf{x} = z \mid \mathcal{F}(\mathsf{x}) = o\big)\Big] \cdot \Pr(\mathcal{F}(\hat{\mathsf{x}}) = o)\,\mathrm{d}o\right].
\end{aligned} \tag{2}$$

Here, $\mathbf{1}(C) = 1$ iff condition $C$ holds. Equation (2) shows that $\delta_I$ is the probability mass of observation events for which the *true* secret happens to be MAP-optimal given the observed leakage.

Unfortunately, in the black-box setting, directly computing (2) is typically infeasible because the secret space $\mathcal{X}$ may be exponentially large. However, the expectation form in (2) suggests a route toward ❷: if there exists a *bounded* function $\mathcal{G}(\cdot)$ such that

$$\delta_I = \mathbb{E}_{\bar{\mathsf{x}}\sim\mathcal{D}}\left[\mathcal{G}\big(\bar{\mathsf{x}}, \mathcal{F}(\bar{\mathsf{x}})\big)\right], \tag{3}$$

then the empirical average $\frac{1}{m}\sum_{i=1}^{m} \mathcal{G}\big(\bar{\mathsf{x}}_i, \mathcal{F}(\bar{\mathsf{x}}_i)\big)$ over samples $\mathsf{X} = \{\bar{\mathsf{x}}_i\}_{i=1}^m$ is an unbiased estimator of $\delta_I$, and concentration inequalities (e.g., Hoeffding's) yield non-asymptotic error bounds.

To incorporate optimal mitigation (❸), we can represent a privacy mechanism by a perturbation term $\mathsf{e} \sim \mathcal{D}_e$ and formulate a constrained optimization problem: under a utility loss measure $\kappa$ and a privacy requirement $\delta_I \le \epsilon$,

$$\min_{\mathcal{D}_e} \ \kappa(\mathcal{D}_e) \quad \text{s.t.} \quad \mathbb{E}_{\bar{\mathsf{x}}\sim\mathcal{D}}\left[\mathcal{G}\big(\bar{\mathsf{x}}, \mathcal{F}(\bar{\mathsf{x}})\big)\right] \le \epsilon. \tag{4}$$

As in statistical learning, because the population expectation constraint in (4) is unknown, one may replace it with an empirical proxy:

$$\min_{\mathcal{D}_e} \ \kappa(\mathcal{D}_e) \quad \text{s.t.} \quad \frac{1}{n}\sum_{i=1}^{m} \mathcal{G}\big(\bar{\mathsf{x}}_i, \mathcal{F}(\bar{\mathsf{x}}_i)\big) \le \epsilon'. \tag{5}$$

To ensure that solutions satisfying (5) also satisfy (4) with high probability (i.e., to control generalization error), a standard approach is to choose a stricter threshold $\epsilon' < \epsilon$ and bound the gap $\epsilon - \epsilon'$ via uniform convergence or complexity measures (e.g., Rademacher complexity) [43]. In black-box settings, one may alternatively adopt a *propose–verify* strategy that gradually increases noise $\mathsf{e}$ until an independent test (using fresh samples) is passed.

All of the above depends on constructing such a function $\mathcal{G}$, which is generally challenging. Building on PAC Privacy [44] and $\alpha$-mutual information [37,42,18], we establish a new family of estimators $\{\tilde{\mathcal{G}}_\alpha : \alpha \in \mathbb{R}^+\}$ (termed *reference-prior-weighted $\alpha$-information*), which yields an unbiased representation of $\delta_I$ in a *doubly asymptotic* sense:

$$\delta_I = \mathbb{E}_{\bar{\mathsf{x}}\sim\mathcal{D}}\Big[\lim_{\alpha\to\infty} \tilde{\mathcal{G}}_\alpha\big(\bar{\mathsf{x}}, \mathcal{F}(\bar{\mathsf{x}})\big)\Big]. \tag{6}$$

More importantly, for any fixed $\alpha$, we have the upper bound

$$\delta_I \le \mathbb{E}_{\bar{\mathsf{x}}\sim\mathcal{D}}\Big[\tilde{\mathcal{G}}_\alpha\big(\bar{\mathsf{x}}, \mathcal{F}(\bar{\mathsf{x}})\big)\Big],$$

so any non-asymptotic analysis based on a specific $\tilde{\mathcal{G}}_\alpha$ provides a provable high-probability guarantee, with larger $\alpha$ yielding tighter bounds. Building on the identification case, in Section 3.2 we further show that analyzing an arbitrary inference criterion $\rho$ can be reduced to an identification problem, which completes the solution to Problem 1.

For composition, as illustrated in Fig. 1, the central challenge is that intermediate randomness and adversarial adaptivity induce a global leakage distribution—a mixture over potentially infinitely many conditional branches ("parallel universes"); but the user observes only a single realized path (the "local universe"). Rather than performing a direct Bayesian analysis of this global mixture, we propose a randomization mechanism constrained to the local universe, ensuring that the cumulative risk for the realized leakage instance *matches* the privacy budget exactly. Intuitively, while each instance operates locally, if all alternate instances across parallel universes satisfies their respective local guarantees, the collection of local controls yields a tight global composition bound.

To capture this, we consider decomposing the overall privacy risk accounting into a sequence of smaller subproblems, each solvable within its appropriate "universe". Once composed, these local controls provide a global privacy guarantee. We consider two composition setups, distinguished by whether the per-round mechanism may depend on the realized transcript of previous rounds.

In the *history-accessible* setting, the user is allowed to condition each round's mechanism on the full interaction history, including previously selected mechanisms, realized outputs, and adversarial queries. This allows one to track a running "leakage score" for each candidate secret along the current transcript. Consequently, this enables a *branch-dependent* decomposition: per-round subproblems may depend on the current history, allowing the user to adaptively calibrate the randomization at each round.

Conversely, in the *mechanism-agnostic* setting, the decomposition must be determined *independently* of the realized history. In this scenario, the adversary may concentrate on the most vulnerable region of the secret space to maximize inference success, and the user is therefore compelled to enforce a worst-case (uniform) guarantee over all possible, unseen histories with carefully-selected separate measurement of each marginal leakage function.

## 2 Preliminaries

### 2.1 Statistical Divergence

We first introduce a family of useful distributional-distance measurements.

**Definition 1 ((Generalized) f-Divergence).** *Let $f : (0, +\infty) \to \mathbb{R}$ be a convex function. Let $P$ and $Q$ be two probability distributions defined on the same measurable space, with $P$ absolutely continuous with respect to $Q$. The $f$-divergence between $P$ and $Q$ is defined as*

$$D_f(P \,\|\, Q) \;:=\; \mathbb{E}_{X\sim Q}\left[f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(X)\right)\right]. \tag{7}$$

*Here, $\mathrm{d}P$ (resp., $\mathrm{d}Q$) denotes the probability density function in the continuous case or the probability mass function in the discrete case.*

Classic definition of f-divergence [35] further requires $f(1) = 0$ to ensure the divergence between two identical distributions equals 0, which is not necessary for the analysis in this paper. It is also noted that many commonly-used statistical divergences are special cases of f-divergence. For example, *KL-divergence* and *total variation* correspond to the selection of $f(z) = z\log(z)$ and $f(z) = |z-1|$ in Definition 1, respectively. In particular, one important variant we will constantly use in the paper is the $\alpha$-divergence, $D_\alpha$, by selecting $f(z) = z^\alpha$.

**Definition 2 ($\alpha$-Divergence).** *Let $P$ and $Q$ be two probability distributions defined on the same measurable space. For any $\alpha > 1$,*

$$D_\alpha(P \,\|\, Q) := \mathbb{E}_{\mathsf{x}\sim Q}\left[\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\mathsf{x})\right)^\alpha\right].$$

As a side note, the well-known Rényi-divergence [27] can be viewed as a logarithmic transformation of the $\alpha$-divergence, $\frac{1}{\alpha-1}\log D_\alpha(P \,\|\, Q)$. An important property of f-divergence is their joint convexity, as shown below.

**Lemma 1 (Joint convexity of f-Divergence).** *Let $f : (0, +\infty) \to \mathbb{R}$ be a convex function. Then the function*

$$(z, v) \;\mapsto\; v\, f\left(\frac{z}{v}\right)$$

*is jointly convex in* $(z, v)$ *over* $z \geq 0$ *and* $v > 0$*. Consequently, f-divergence is jointly convex with respect to the pair of mixture distributions* $\big(c \cdot P_1 + (1-c)P_2, cQ_1 + (1-c)Q_2\big)$ *for any* $c \in [0,1]$*, such that*

$$D_f(c \cdot P_1 + (1-c) \cdot P_2 \,\|\, c \cdot Q_1 + (1-c) \cdot Q_2) \leq cD_f(P_1 \,\|\, Q_1) + (1-c)D_f(P_2 \,\|\, Q_2).$$

### 2.2 Privacy Definitions

In the following, we cover the background of privacy definitions. As discussed earlier, depending on the type of secret to protect, there are two main philosophies: *Input-Independent Indistinguishability* (III) for Type-(B) secret and Bayesian for Type-(A) secret.

**Definition 3 ($\epsilon$ input-independent indistinguishability [36,15,10,21]).** *Given a leakage function* $\mathcal{F}(\cdot) : \mathcal{X} \to \mathcal{O}$ *and for an arbitrary pair of* adjacent *secret hypotheses* $\{\bar{\mathsf{x}}, \bar{\mathsf{x}}'\}$ *such that* $\bar{\mathsf{x}}$ *and* $\bar{\mathsf{x}}'$ *only differ in the sensitive information that desires protection, then* $\mathcal{F}(\cdot)$ *satisfies* $\epsilon$*-III if the following experiment is impossible:*

*A user randomly selects* $\mathsf{x}$ *from* $\{\bar{\mathsf{x}}, \bar{\mathsf{x}}'\}$ *and sends* $\mathcal{F}(\mathsf{x})$ *to the adversary; the (computationally bounded or unbounded) adversary can return a guessing* $\hat{\mathsf{x}}$ *such that with probability at least* $\epsilon$*, taken over the randomness of* $\mathsf{x}$ *and* $\mathcal{F}$*,* $\hat{\mathsf{x}} = \mathsf{x}$*.*

As for Type-(A) entropic secret, the main focus of this paper, a general and formal definition of Bayesian risk is *PAC Privacy* [44].

**Definition 4 ($(\mathcal{D}, \rho, \delta_\rho)$ PAC Privacy [44]).** *For a leakage function* $\mathcal{F} : \mathcal{X} \to \mathcal{O}$*, let* $\mathcal{D}$ *be the prior distribution of secret* $\mathsf{x}$ *over* $\mathcal{X}$*, and* $\rho : \mathcal{X} \times \mathcal{X} \to \{0, 1\}$ *be an inference criterion:* $\rho(\hat{\mathsf{x}}, \mathsf{x}) = 1$ *iff an adversarial guess* $\hat{\mathsf{x}}$ *is successful. Then* $\mathcal{F}$ *satisfies* $(\mathcal{D}, \rho, \delta_\rho)$ *PAC Privacy if the following experiment is impossible:*

*A user samples* $\mathsf{x} \sim \mathcal{D}$ *and releases* $\mathcal{F}(\mathsf{x})$ *to an informed adversary who knows both* $\mathcal{D}$*,* $\mathcal{F}$ *and* $\rho$*. The adversary returns an estimation* $\hat{\mathsf{x}}$ *of* $\mathsf{x}$ *such that with probability greater than* $\delta_\rho$*,* $\rho(\hat{\mathsf{x}}, \mathsf{x}) = 1$*.*

In the following, when there is no contextual ambiguity, we will simply call the $(\mathcal{D}, \rho, \delta_\rho)$ PAC Privacy as $\delta_\rho$ privacy.

### 2.3 Bayesian Analysis Tools

In this section, we introduce some information-theoretical tools for the objective Bayesian analysis. One operationally-efficient result, which we term *f-divergence Fano Inequality*, as a slightly-generalized version of the main theorem of PAC Privacy (Theorem 1 in [44]), presents an upper bound of target $\delta_\rho$ in an expectation form.

**Lemma 2 (f-divergence Fano Inequality [44]).** *For any selected $f$-divergence $D_f$, an arbitrary leakage function $\mathcal{F}: \mathcal{X} \to \mathcal{O}$ satisfies $(\delta_\rho, \rho, \mathcal{D})$* PAC Privacy *where the adversarial posterior success rate $\delta_\rho$ can be selected as the minimal number satisfying either one of the following inequalities:*

$$\begin{aligned} D_f\big(\mathbf{1}_{\delta_\rho} \,\big\|\, \mathbf{1}_{\delta_{o,\rho}}\big) \;&\geq\; \inf_{\mathbb{P}_W} D_f\big(\mathbb{P}_{\mathsf{x},\mathcal{F}(\mathsf{x})} \,\big\|\, \mathbb{P}_{\mathsf{x}} \otimes \mathbb{P}_W\big) = \inf_{\mathbb{P}_W}\; \mathbb{E}_{\mathsf{x}\sim\mathcal{D}}\big[\, D_f\big(\mathbb{P}_{\mathcal{F}(\mathsf{x})} \,\big\|\, \mathbb{P}_W\big)\big]\,; \\ D_f\big(\mathbf{1}_{\delta_{o,\rho}} \,\big\|\, \mathbf{1}_{\delta_\rho}\big) \;&\geq\; \inf_{\mathbb{P}_W} D_f\big(\mathbb{P}_{\mathsf{x}} \otimes \mathbb{P}_W \,\big\|\, \mathbb{P}_{\mathsf{x},\mathcal{F}(\mathsf{x})}\big) = \inf_{\mathbb{P}_W}\; \mathbb{E}_{\mathsf{x}\sim\mathcal{D}}\big[\, D_f\big(\mathbb{P}_W \,\big\|\, \mathbb{P}_{\mathcal{F}(\mathsf{x})}\big)\big]\,. \end{aligned} \tag{8}$$

*Here, $\delta_{o,\rho}$ is the optimal prior success rate of recovering $\mathsf{x}$ before observing the leakage; $\mathbf{1}_{\delta_\rho}$ and $\mathbf{1}_{\delta_{o,\rho}}$ are two Bernoulli distributions of parameters $\delta_\rho$ and $\delta_{o,\rho}$, respectively; $\mathbb{P}_{X,\mathcal{F}(X)}$ and $\mathbb{P}_X$ are the joint distribution and the marginal distribution of $(\mathsf{x}, \mathcal{F}(\mathsf{x}))$ and $\mathsf{x}$, respectively; and $\mathbb{P}_W$ represents the distribution of an arbitrary random variable $W$ defined over $\mathcal{O}$.*

It has been proved (Lemma 1 in [44]) that the left hand sides of (8), $D_f\big(\mathbf{1}_{\delta_\rho} \,\big\|\, \mathbf{1}_{\delta_{o,\rho}}\big)$ and $D_f\big(\mathbf{1}_{\delta_{o,\rho}} \,\big\|\, \mathbf{1}_{\delta_\rho}\big)$, are both monotone to $\delta_\rho$. Thus, Lemma 2 suggests that through optimizing a *reference* marginal distribution $\mathbb{P}_W$ over the leakage domain, the expectation of an arbitrary f-divergence between leakage likelihood $\mathbb{P}_{\mathcal{F}(\mathsf{x})}$ conditional on specific secret $\mathsf{x}$ and the $\mathbb{P}_W$ can form an upper bound to target $\delta_\rho$. In particular, when the f-divergence $D_f$ is selected to be the KL-divergence, (8) reduces to the well-known Fano's inequality [13], which connects the Shannon entropy and the Bayesian. Another concept related to (8) is $\alpha$-mutual information [37] by selecting the f-divergence $D_f$ to be $\alpha$-divergence (Definition 2).

**Definition 5 (Sibson's $\alpha$-mutual information [37]).** *Let $\mathbb{P}_{XY}$ be some joint distribution of random variables $(X, Y)$ whose marginal distribution with respect to $X$ is $\mathbb{P}_X$. For any $\alpha > 1$,* Sibson's $\alpha$-mutual information *is defined by the Rényi information radius for arbitrary marginal distribution $Q$ of $Y$,*

$$I_\alpha^{sib}(X;Y) \;:=\; \inf_{Q}\; \frac{1}{\alpha - 1} \cdot \log D_\alpha\big(\mathbb{P}_{XY} \,\big\|\, \mathbb{P}_X \otimes Q\big). \tag{9}$$

When $\alpha \to \infty$, $I_\infty^{sib}(X;Y)$ forms the foundation of *maximal leakage* [18]. In a *prior-free* setup, taken over the worst case of prior distribution $\mathbb{P}_X$, $\sup_{\mathbb{P}_X} I_\infty^{sib}(X;Y)$ captures the logarithm of the maximal multiplicative posterior gain [18].

## 3 Towards Tight Privatization of Static Leakage Function

In this section, we systemically address Problem 1 for static leakage function. Provided the simulatable nature of the expectation form in (8), a natural question stemmed from Lemma 2 is that whether the equality of (8) is achievable. That is to say, through optimizing the selection of f-divergence and reference distribution $\mathbb{P}_W$, whether the optimal posterior rate $\delta_\rho$ can be tightly characterized from (8). In Section 3.1, we start to identify the possibility of a special

case: the uniform-prior identification setting, for which we develop progressively tighter upper bounds on $\delta_I$ through $\alpha$-divergence and translate them into concrete, asymptotically optimal randomization rule $\mathcal{D}_e$ (Algorithm 1). However, for general priors and inference settings, the application of Lemma 2 is *not* tight anymore and thus in Section 3.2, we establish a generalized structure, termed *Prior-(Reference)-Weighted* $\alpha$-Information, and show how to tightly determine and control general $\delta_\rho$.

### 3.1 Tight $\alpha$-Information to Identification under Uniform Priors

We start by revisiting f-divergence Fano Inequality (Lemma 2) in a setting when $\mathcal{D}$ is a uniform distribution and studying the conditions that can lead to the equality in (8). As discussed in Section 1.4, after observing a leakage $\mathsf{o} = \mathcal{F}(\mathsf{x})$, the adversary's optimal decision rule $\mathcal{A}$ is MAP decoding:

$$\mathcal{A}(\mathsf{o}) = \arg\max_{\hat{\mathsf{x}}} \Pr_{\mathsf{x}\sim\mathcal{D}}\big(\mathsf{x} = \hat{\mathsf{x}} \mid \mathcal{F}(\mathsf{x}) = \mathsf{o}\big) = \arg\max_{\hat{\mathsf{x}}} \Pr_{\mathsf{x}\sim\mathcal{D}}(\mathsf{x} = \hat{\mathsf{x}}) \Pr\big(\mathcal{F}(\mathsf{x}) = \mathsf{o}\big).$$

Under a uniform prior, $\Pr(\mathsf{x} = \hat{\mathsf{x}})$ is constant, so the success probability is driven by the largest likelihood $\max_{\hat{\mathsf{x}}} \Pr(\mathcal{F}(\hat{\mathsf{x}}) = \mathsf{o})$. This leads to the 0-1-loss expression of $\delta_I$ in (2), where only these most likely hypotheses determine the optimal posterior rate. This intuitively contrasts with the expectation structure in the right-hand side of (8)

$$\mathbb{E}_{\mathsf{x}\sim\mathcal{D}}\, D_f\big(\mathbb{P}_{\mathcal{F}(\mathsf{x})} \,\big\|\, \mathbb{P}_W\big) = \int_{\mathcal{O}} \Pr(W = \mathsf{o})\, \mathbb{E}_{\mathsf{x}\sim\mathcal{D}}\left[f\left(\frac{\Pr(\mathcal{F}(\mathsf{x}) = \mathsf{o})}{\Pr(W = \mathsf{o})}\right)\right] d\mathsf{o}, \quad (10)$$

which averages contributions from all $\mathsf{x}$. This suggests that the choice of $f$ matters. To approach equality, we expect the $f$ in (10) grows steeply for large arguments $\Pr(\mathcal{F}(\mathsf{x}) = \mathsf{o})$ so that the expectation can be dominated by the largest ratio terms, and (10) behaves more like a worst-case control rather than an average-case one, yielding a tighter bound.

Motivated by this, we focus on the family $f(z) = z^\alpha$, which induces the $\alpha$-divergence (Definition 2). Compared to the KL choice $f(z) = z \log z$ adopted in PAC Privacy [44], which grows only slightly super linearly, the power function $t^\alpha$ penalizes large likelihood ratios much more aggressively as $\alpha$ increases. Indeed, it can be proved that as $\alpha \to \infty$, (10) provides tight characterization of $\delta_I$.

**Proposition 1 (Asymptotic tightness of $\alpha$-divergence approximation).** *Let $\mathsf{x} \sim \mathcal{D}$ be a secret input and let $\mathcal{F}(\cdot) : \mathcal{X} \to \mathcal{O}$ be an arbitrary leakage function. Define a reference distribution $\mathbb{P}_W$ on the $\mathcal{O}$ by*

$$\Pr(W = \mathsf{o}) = \frac{\big(\mathbb{E}_{\mathsf{x}\sim\mathcal{D}}[\Pr(\mathcal{F}(\mathsf{x}) = \mathsf{o})^\alpha]\big)^{1/\alpha}}{C},$$

*where the normalizing constant $C$ is given by*

$$C = \int_{\mathcal{O}} \big(\mathbb{E}_{\mathsf{x}\sim\mathcal{D}}[\Pr(\mathcal{F}(\mathsf{x}) = \mathsf{o})^\alpha]\big)^{1/\alpha}\, d\mathsf{o}.$$

*Then, as $\alpha \to \infty$, the upper bound of posterior success rate obtained from the Lemma 2 with respect to $W$ is exactly the constant $C$, which converges to the optimal posterior success rate $\delta_I$.*

Proposition 1 shows that, by choosing an appropriate reference distribution $\mathbb{P}_W$, the $\alpha$-divergence bound becomes asymptotically tight as $\alpha \to \infty$. It also tells us that the optimal reference $W$ in (8) is proportional to the $\alpha$-norm aggregation of the output likelihoods, namely

$$\Pr(W = \mathsf{o}) \;\propto\; \Big(\mathbb{E}_{\mathsf{x}\sim\mathcal{D}}\big[\Pr(\mathcal{F}(\mathsf{x}) = \mathsf{o})^{\alpha}\big]\Big)^{1/\alpha}, \tag{11}$$

which is exactly the choice that makes Hölder's inequality tight in the proof of Proposition 1.

In practice, exact computation of $\mathbb{P}_W$ and the expectation (10) can be computationally infeasible, given potentially exponentially-large input domain $\mathcal{X}$. Fortunately, both $\mathbb{P}_W$ and (10) can be asymptotically approached from finite samples. Given a i.i.d. sample set $\mathsf{X} = \{\bar{x}_i, i \in [m]\}$, $\bar{\mathsf{x}} \sim \mathcal{D}$, $W$ can be instead selected as

$$\Pr(W = \mathsf{o}) \;\propto\; \Big(\sum_{i=1}^{m} \frac{1}{m} \cdot \big[\Pr(\mathcal{F}(\bar{x}_i) = \mathsf{o})^{\alpha}\big]\Big)^{1/\alpha}, \tag{12}$$

and notably from Lemma 2, any (even if non-optimal) selection of $W$ always leads to a safe upper bound of target $\delta_I$.

As for the estimation of expectation (10), concentration inequalities allow us to make high-probability statement on the estimation error from sampled instance $\mathsf{X}$. However, we need to be careful here since the failure rate of a small estimation error *cannot* be straightforwardly merged into the upper bound of $\delta_I$: by Lemma 2, instead we need a global, *deterministic* upper bound on (10), the right hand side of Lemma 2. To address this, our solution is to treat the randomness of $\mathsf{X}$ as part of the entire privacy-preserving mechanism's randomness, and view our final randomization strategy $\mathcal{D}_e$ as a function of the sampled instance $\mathsf{X}$ as well. To be formal, let $\mathcal{M}(\cdot) = \mathcal{F}(\cdot) + \mathsf{e}$, $\mathsf{e} \sim \mathcal{D}_e$, where $\mathcal{D}_e$ is determined by some mechanism $\mathsf{Alg}(\mathsf{X})$, which takes a randomly-sampled set $\mathsf{X}$ as input. Thus, the output distribution of $\mathcal{M}(\cdot)$ is a mixture over the sample-conditioned mechanisms $\{\mathcal{M}_{\mathsf{X}}\}$ induced by drawing $\mathsf{X}$. To ensure the overall budget (10) is at most $r$, we set a stricter target budget $r' < r$ that holds on "good" sample set $\mathsf{X}$ with small enough estimation error; the fraction of "good" sample set $\mathsf{X}$ can be lower bounded by some threshold $(1-\gamma)$ from standard concentration inequality. For the remaining, up to $\gamma$-fraction (exponentially decaying in the number of samplings $m$), "bad" sample set $\mathsf{X}$, we may switch to some conservative worst-case bound $R$. By joint convexity (Lemma 1) of $f$-divergences, we can show (10) is deterministically upper bounded by $(1-\gamma)r' + \gamma R$. Therefore it suffices to choose parameters $(r', \gamma, R)$ such that $(1-\gamma)r' + \gamma R \le r$. We put the above ideas together to establish Algorithm 1, a general framework to optimize privacy-preserving randomization $\mathcal{D}_e$ in the black-box setting with formal grantees shown in Theorem 1.

**Algorithm 1** Randomization $\mathcal{D}_e$ Solver

1: **Input:** Leakage function $\mathcal{F}$; privacy-loss function $\mathcal{G}(\mathcal{D}_e, \cdot)$; input distribution $D$; utility loss metric $\kappa$, privacy budget $r$; sample size $m$; global bound $B(\mathcal{D}_e)$.
2: Sample $\mathsf{X} = \{\bar{\mathsf{x}}_1, \ldots, \bar{\mathsf{x}}_m\}$ and $\mathsf{X}' = \{\bar{\mathsf{x}}'_1, \ldots, \bar{\mathsf{x}}'_m\}$, all i.i.d. from $D$.
3: Choose a target budget $r'$, a failure probability $\gamma \in (0, 1)$, and a fallback, universal bound $R$ such that

$$(1 - \gamma)\, r' + \gamma\, R \leq r.$$

4: Solve the constrained optimization problem

$$\min_{\mathcal{D}_e} \kappa(\mathcal{D}_e), \quad \text{such that} \quad \frac{1}{m} \sum_{i=1}^{m} \mathcal{G}(\mathcal{D}_e, \bar{\mathsf{x}}_i) \leq r' \tag{13}$$

to obtain an initial noise distribution $\mathcal{D}_e$.
5: Based on $\mathsf{X}'$, compute

$$L \;\leftarrow\; \frac{1}{m} \sum_{i=1}^{m} g(\mathcal{D}_e, \bar{\mathsf{x}}'_i), \qquad \beta \;\leftarrow\; B(\mathcal{D}_e) \sqrt{\frac{\log(1/\gamma)}{2m}}.$$

6: **while** $L > r' - \beta$ **or** $B(\mathcal{D}_e) > R$ **do**
7: Update $\mathcal{D}_e \leftarrow \mathcal{D}_e + \mathcal{N}(0, c^2 \cdot \boldsymbol{I})$, by adding an independent Gaussian noise $z \sim \mathcal{N}(0, \sigma^2 \cdot \boldsymbol{I})$ to the current perturbation solution $\mathsf{e} \sim \mathcal{D}_e$.
8: Recompute $L$ and $\beta$ as in step 5 for the updated $\mathcal{D}_e$.
9: **end while**
10: **Output:** Mechanism $\mathcal{M}_{\mathcal{D}_e}$.

**Theorem 1 (Provable guarantees of Algorithm 1).** *Let $\mathcal{F} : \mathcal{X} \to \mathcal{O}$ be a given leakage function, and $\mathcal{M}_{\mathcal{D}_e}(\cdot) = \mathcal{F}(\cdot) + \mathsf{e}$ denote the privatized mechanism of $\mathcal{F}$ based on strategy $\mathsf{e} \sim \mathcal{D}_e$. For any given $\mathcal{D}_e$, define the privacy-loss function*

$$\mathcal{G}(\mathcal{D}_e, \mathsf{x}) \;:=\; \mathcal{D}_\alpha\big(\mathbb{P}_{\mathcal{M}_{\mathcal{D}_e}(\mathsf{x})} \,\|\, \mathbb{P}_{W_{\mathcal{D}_e}}\big),$$

*where $\mathsf{x} \in \mathcal{X}$ is the secret input and $\mathbb{P}_{W_{\mathcal{D}_e}}$ is some reference distribution on $\mathcal{O}$, potentially depending on $\mathcal{D}_e$. Suppose that for every $\mathsf{x} \in \mathcal{X}$ and every admissible $\mathcal{D}_e$, we have $\mathcal{G}(\mathcal{D}_e, \hat{\mathsf{x}}) \leq B(\mathcal{D}_e)$ for some finite bound $B(\mathcal{D}_e)$. Let $\mathcal{D}_e$ be the mechanism returned by Algorithm 1, then we have*

$$\mathbb{E}_{\mathsf{x}\sim D}\Big[\mathcal{G}(\mathcal{D}_e, \mathsf{x})\Big] \;=\; \mathbb{E}_{\mathsf{x}\sim D}\Big[\mathcal{D}_\alpha\big(\mathbb{P}_{\mathcal{M}_{\mathcal{D}_e}(\mathsf{x})} \,\|\, \mathbb{P}_{W_{\mathcal{D}_e}}\big)\Big] \;\leq\; r.$$

Algorithm 1 uses two independent sample sets $\mathsf{X}$ and $\mathsf{X}'$. The first set $\mathsf{X}$ is used to initialize a candidate $\mathcal{D}_e$ (step 4), which with high chance can be asymptotically close to the optimal strategy $\mathcal{D}_e^*$ when we replace the empirical constraint $\frac{1}{m}\sum_{i=1}^{m} \mathcal{G}(\mathcal{D}_e, \bar{\mathsf{x}}_i)$ in (13) by the true population constraint $\mathbb{E}_{\mathsf{x}\sim\mathcal{D}}\mathcal{G}(\mathcal{D}_e, \bar{\mathsf{x}}_i)$. The second set $\mathsf{X}'$ is used to validate whether this candidate satisfies the desired privacy budget $r$, which has a bijection relationship to the posterior $\delta_I$ from an equality in (8). If the proposed solution $\mathcal{D}_e$ does not pass the test, we will continuously add additional independent noise. This separate application on disjoint

$\mathsf{X}$ and $\mathsf{X}'$ is essential: once $\mathcal{D}_e$ has been chosen based on $\mathsf{X}$, the random variables $\{\mathcal{G}(\mathcal{D}_e, \bar{\mathsf{x}}_i)\}_{i=1}^m$ are no longer independent of the choice of $\mathcal{D}_e$, so applying concentration inequality (e.g., Hoeffding's) to the same samples would be invalid. Using an independent set $\mathsf{X}'$ restores the required independence and enables a standard concentration bound, thereby establishing the relationship between the confidence level $\gamma$ and the sample size $m$.

### 3.2 Prior–Reference-Weighted (PRW) $\alpha$-Information and General Inference under Arbitrary Prior

The previous section analyzed identification attacks in the special case where the secret $\mathsf{x}$ is drawn uniformly. We now turn to the general case for arbitrary $\rho$ and prior $\mathcal{D}$. However, when $\alpha$-divergence is used as the underlying $f$-divergence, the bound in Lemma 2 is no longer tight in general. With a closer look at the proof of Proposition 1, we see that when the prior distribution $\mathcal{D}$ is not uniform, the factor $\Pr_{\mathsf{x}\sim\mathcal{D}}(\mathsf{x} = \hat{\mathsf{x}})$ does not appear in the definition of the optimal reference distribution $\mathbb{P}_W$, which creates a strict gap between the optimal posterior success rate $\delta_\rho$ and the guarantee provided by Lemma 2. It is unclear whether there exists another choice of $f$ for which the resulting bound converges to the ground truth for arbitrary priors $\mathcal{D}$; we leave this as an interesting direction for future work.

Although the $\alpha$-divergence version of Lemma 2 is not tight, the MAP rule still holds for an arbitrary inference criteria $\rho$, where the adversary's optimal strategy to maximize success rate is still to return the guess $\hat{\mathsf{x}}$ with the highest posterior probability, formalized in the following lemma.

**Lemma 3 (Maximal success rate under a general inference criterion and prior distribution).** *Let $\mathsf{x}$ be the secret with prior $\Pr_{\mathsf{x}\sim\mathcal{D}}(\mathsf{x} = \hat{\mathsf{x}})$, and let $\mathcal{F}$ be a randomized mechanism with output $\mathsf{o} = \mathcal{F}(\mathsf{x})$. Fix an inference criterion $\rho : \mathcal{X} \times \mathcal{X} \to \{0, 1\}$, where $\rho(\mathsf{x}, \hat{\mathsf{x}}) = 1$ means that the guessing $\hat{\mathsf{x}}$ is counted as success when the true secret is $\mathsf{x}$. Define the joint weights*

$$H(\mathsf{x}, \mathsf{o}) \;:=\; \sum_{\hat{\mathsf{x}}' \,:\, \rho(\mathsf{x},\hat{\mathsf{x}}')=1} \Pr(\mathsf{x} = \hat{\mathsf{x}}')\, \Pr\big(\mathcal{F}(\mathsf{x}) = \mathsf{o} \,\big|\, \mathsf{x} = \hat{\mathsf{x}}'\big).$$

*Then, the maximal success probability over all decision rules is*

$$\delta_\rho = \int_{\mathcal{O}} \max_{\mathsf{x}\in\mathcal{X}} H(\mathsf{x}, \mathsf{o})\, \mathrm{d}\mathsf{o}, \tag{14}$$

To tightly approach (14), similarly we seek some function taking $\Pr(\mathcal{F}(\mathsf{x}) = \mathsf{o})$ as a building block that converges to $\Pr\big(\mathsf{x} = \hat{\mathsf{x}},\ \mathcal{F}(\mathsf{x}) = \mathsf{o}\big)$ in the appropriate limit. It is observed that the max over $x$ can be viewed as an $\ell_\infty$-norm, so, still, it is natural to approximate it by an $\ell_\alpha$-norm as $\alpha \to \infty$. The subtlety is that we need an *upper* bound on $\max_{\mathsf{x}\in\mathcal{X}} H(\mathsf{x}, \mathsf{o})$, whereas the standard normalized $\ell_\alpha$ expression $\big(\frac{1}{|\mathcal{X}|} \sum_{\mathsf{x}\in\mathcal{X}} H(\mathsf{x}, \mathsf{o})^\alpha\big)^{1/\alpha}$ is a *lower* bound on the $\max_{\mathsf{x}\in\mathcal{X}} H(\mathsf{x}, \mathsf{o})$.

To obtain an upper bound, we instead use

$$\left(\sum_{\mathsf{x}\in\mathcal{X}} H(\mathsf{x},\mathsf{o})^{\alpha}\right)^{1/\alpha},$$

which clearly upper-bounds $\max_{\mathsf{x}\in\mathcal{X}} H(\mathsf{x},\mathsf{o})$. Notice that

$$\left(\sum_{\mathsf{x}\in\mathcal{X}} H(\mathsf{x},\mathsf{o})^{\alpha}\right)^{1/\alpha} \;\geq\; \left(\max_{\mathsf{x}\in\mathcal{X}} H(\mathsf{x},\mathsf{o})^{\alpha}\right)^{1/\alpha} \;=\; \max_{\mathsf{x}\in\mathcal{X}} H(\mathsf{x},\mathsf{o}), \tag{15}$$

since the sum includes the largest term $H(\mathsf{x}^*,\mathsf{o})^{\alpha}$ for $\mathsf{x}^* \in \arg\max_{\mathsf{x}} H(\mathsf{x},\mathsf{o})$. Moreover, as $\alpha \to \infty$, the expression $\left(\sum_{\mathsf{x}} H(\mathsf{x},\mathsf{o})^{\alpha}\right)^{1/\alpha}$ converges to $\max_x H(\mathsf{x},\mathsf{o})$, as summarized in the following lemma.

**Lemma 4 ($\ell_\alpha$-relaxation of the max posterior over $x$).** *For any $\alpha > 1$, we have*

$$\int_{\mathcal{O}} \max_{\mathsf{x}\in\mathcal{X}} H(\mathsf{x},\mathsf{o})\,\mathrm{d}\mathsf{o} \;\leq\; \int_{\mathcal{O}} \left(\sum_{\mathsf{x}\in\mathcal{X}} H(\mathsf{x},\mathsf{o})^{\alpha}\right)^{1/\alpha} \mathrm{d}\mathsf{o}. \tag{16}$$

*Moreover, for every fixed $o \in \mathcal{O}$,*

$$\left(\sum_{\mathsf{x}\in\mathcal{X}} H(\mathsf{x},\mathsf{o})^{\alpha}\right)^{1/\alpha} \;\geq\; \max_{\mathsf{x}\in\mathcal{X}} H(\mathsf{x},\mathsf{o}),$$

*and the left-hand side converges to $\max_{\mathsf{x}\in\mathcal{X}} H(\mathsf{x},\mathsf{o})$ as $\alpha \to \infty$. Hence the right-hand side above is an $\ell_\alpha$-norm-based upper bound that becomes asymptotically tight as $\alpha$ grows.*

Moreover, for operational efficiency, we further seek an upper bound of (16) above that pulls the summation over $x$ outside the integral, so that the summation can be normalized by $\Pr(\mathsf{x} = \hat{\mathsf{x}})$ and recognized as an expectation over $\mathsf{x}$. This will allow us to approximate the resulting quantity using a finite sample drawn from $\mathcal{D}$. Let $\mathbb{P}$ be an arbitrary reference distribution on $\mathcal{O}$. By Hölder's inequality, we have

**Lemma 5 (Prior–reference-weighted $\alpha$-norm upper bound).** *For any reference distribution $\mathbb{P}_W$ on $\mathcal{O}$ and any $\alpha > 1$,*

$$\int_{\mathcal{O}} \left(\sum_{\mathsf{x}\in\mathcal{X}} H(\mathsf{x},\mathsf{o})^{\alpha}\right)^{1/\alpha} \mathrm{d}\mathsf{o} \leq \left(\int_{\mathcal{O}} \frac{\sum_{x\in\mathcal{X}} H(\mathsf{x},\mathsf{o})^{\alpha}}{\Pr(W=\mathsf{o})^{\alpha-1}} \mathrm{d}\mathsf{o}\right)^{1/\alpha} \tag{17}$$

$$= \left(\mathbb{E}_{\mathsf{x}\sim D} \int_{\mathcal{O}} \frac{H(\mathsf{x},\mathsf{o})^{\alpha}/\Pr(\mathsf{x})}{\Pr(W=\mathsf{o})^{\alpha-1}} \mathrm{d}\mathsf{o}\right)^{1/\alpha}. \tag{18}$$

*Moreover, the equality of (17) holds if and only if the behind Hölder's inequality is tight, i.e., there exists a constant $\lambda > 0$ such that*

$$\sum_{\mathsf{x}\in\mathcal{X}} H(\mathsf{x},\mathsf{o})^{\alpha} \;=\; \lambda \Pr(W=\mathsf{o})^{\alpha} \quad \textit{for almost every } \mathsf{o}.$$

*Equivalently, the bound is tight when the reference distribution $\mathbb{P}_W$ is chosen whose probability density is proportional to the $\alpha$-norm of $H(\cdot, o)$,*

$$\Pr(W=\mathsf{o}) \;\propto\; \left(\sum_{x\in\mathcal{X}} H(\mathsf{x},\mathsf{o})^{\alpha}\right)^{1/\alpha}.$$

Lemma 4 and Lemma 5 provide two constructions, both of which yield an asymptotically-tight upper bound of the optimal posterior success rate $\delta_\rho$ for an arbitrary inference criterion $\rho$ and a prior secret distribution $\mathcal{D}$. With respect to (16) and (17), we formalize the two constructions as follows, respectively.

**Definition 6 (Prior–reference-weighted $\alpha$-information).** *Let $\mathsf{x}$ be a random variable taking values in $\mathcal{X}$, $\mathcal{M}$ be a randomized mechanism with output in $\mathcal{O}$, and $\rho$ be an inference criterion. For each $\mathsf{x}\in\mathcal{X}$ and $\mathsf{o}\in\mathcal{O}$, define*

$$H(\mathsf{x},\mathsf{o}) \;:=\; \sum_{\hat{\mathsf{x}}'\in\mathcal{X}:\,\rho(\mathsf{x},\hat{\mathsf{x}}')=1} \Pr(\mathsf{x}=\hat{\mathsf{x}}')\,\Pr\big(\mathcal{F}(\mathsf{x})=\mathsf{o}\mid \mathsf{x}=\hat{\mathsf{x}}'\big),$$

*and let $W$ be any reference distribution on $\mathcal{O}$ that is independent of $\mathsf{x}$. For $\alpha>1$, the* prior–reference-weighted $\alpha$-information with reference $W$ *is defined by*

$$\mathsf{I}^{W}_{\alpha,\rho}(\mathsf{x};\mathcal{F}) \;:=\; \left(\sum_{\mathsf{x}\in\mathcal{X}}\int_{\mathcal{O}} \frac{H(\mathsf{x},\mathsf{o})^{\alpha}}{\Pr(W=\mathsf{o})^{\alpha-1}}\,\mathrm{d}\mathsf{o}\right)^{\frac{1}{\alpha}}.$$

*Equivalently, in an expectation form,*

$$\mathsf{I}^{W}_{\alpha,\rho}(\mathsf{x};\mathcal{F}) = \left(\mathbb{E}_{\mathsf{x}}\left[\frac{1}{\Pr(\mathsf{x})}\int_{\mathcal{O}} \frac{H(\mathsf{x},\mathsf{o})^{\alpha}}{\Pr(W=\mathsf{o})^{\alpha-1}}\,\mathrm{d}\mathsf{o}\right]\right)^{\frac{1}{\alpha}}.$$

**Definition 7 (Prior-weighted $\alpha$-information).** *We further define the* prior-weighted $\alpha$-information *by optimizing over all reference distributions $W$ that are independent of $\mathsf{x}$:*

$$\mathsf{I}_{\alpha,\rho}(\mathsf{x};Y) \;:=\; \inf_{W\perp\mathsf{x}} \mathsf{I}^{W}_{\alpha,\rho}(\mathsf{x};Y).$$

As a corollary from Lemma 4 and Lemma 5, we have the following theorems:

**Theorem 2 (General upper bound on posterior success).** *Let $\mathsf{x}$ be a secret, $\mathcal{F}(\mathsf{x})$ an output mechanism, and $\rho$ an inference criterion as in Definition 4. Then, for any $\alpha>1$ and any reference $W$,*

$$\delta_\rho \;\le\; \mathsf{I}^{W}_{\alpha,\rho}\big(\mathsf{x};\mathcal{F}(\mathsf{x})\big),$$

*and by taking the infimum of $W$ on the both sides of the above inequality, we obtain*

$$\delta_\rho \;\le\; \mathsf{I}_{\alpha,\rho}\big(\mathsf{x};\mathcal{F}(\mathsf{x})\big).$$

Since the choice of $W$ is arbitrary, we aim to identify the optimal reference distribution to derive the tight characterization of $\delta_\rho$. Looking back to Lemma 5, we know that the optimal $W$ should satisfy the condition under which the Hölder's inequality behind (17) is tight. This leads to the following result.

**Theorem 3 (Optimal reference selection).** *Let $\mathsf{x}$ be the secret, a leakage function $\mathcal{F}$ with output space $\mathcal{O}$, and $H(\mathsf{x},\mathsf{o})$ as in Definition 6. For $\alpha > 1$, consider $\mathsf{I}^{W}_{\alpha,\rho}(\mathsf{x};\mathcal{F})$ as $\mathbb{P}_W$ ranges over all reference distributions on $\mathcal{O}$ independent of $\mathsf{x}$. Then the infimum over such $W$ is attained (by Hölder's inequality) at*

$$\Pr(W=\mathsf{o}) \;\propto\; \left(\sum_{\mathsf{x}\in\mathcal{X}} H(\mathsf{x},\mathsf{o})^\alpha\right)^{1/\alpha},$$

*i.e.,*

$$\Pr(W=\mathsf{o}) \;=\; \frac{\left(\sum_{\mathsf{x}\in\mathcal{X}} H(\mathsf{x},\mathsf{o})^\alpha\right)^{1/\alpha}}{\displaystyle\int_{\mathcal{O}}\left(\sum_{\mathsf{x}\in\mathcal{X}} H(\mathsf{x},\mathsf{o}')^\alpha\right)^{1/\alpha}\mathrm{d}\mathsf{o}'}.$$

*For this optimal choice $W^\star$, the prior–reference-weighted $\alpha$-information satisfies*

$$\mathsf{I}_{\alpha,\rho}(\mathsf{x};\mathcal{F}) \;=\; \int_{\mathcal{O}}\left(\sum_{\mathsf{x}\in\mathcal{X}} H(\mathsf{x},\mathsf{o})^\alpha\right)^{1/\alpha}\mathrm{d}\mathsf{o},$$

*where $\mathsf{I}_{\alpha,\rho}(\mathsf{x};\mathcal{F})$ denotes the infimum of $\mathsf{I}^{W}_{\alpha,\rho}(\mathsf{x};\mathcal{F})$ over all such reference distributions $W$.*

Finally, we prove the tightness of prior-weighted $\alpha$-information, which asymptotically approaches the target $\delta_\rho$.

**Theorem 4 (Asymptotic tightness of weighted $\alpha$ information).**

$$\lim_{\alpha\to\infty} \mathsf{I}_{\alpha,\rho}\big(\mathsf{x};\mathcal{F}(\mathsf{x})\big) = \int_{\mathcal{O}} \max_{\mathsf{x}\in\mathcal{X}} H\big(\mathcal{F}(\mathsf{x})=\mathsf{o}\big)\,\mathrm{d}\mathsf{o} = \;\delta_\rho$$

Finally, it is observed that

$$\mathsf{I}^{W}_{\alpha,\rho}(\mathsf{x};\mathcal{F})^\alpha = \left(\mathbb{E}_{\mathsf{x}}\left[\frac{1}{\Pr(\mathsf{x})}\int_{\mathcal{O}}\frac{H(\mathsf{x},\mathsf{o})^\alpha}{\Pr(W=\mathsf{o})^{\alpha-1}}\,\mathrm{d}\mathsf{o}\right]\right)$$

is still an expectation with respect to the secret $\mathsf{x}\sim\mathcal{D}$. Consequently, it can be approximated by Monte Carlo estimation using finitely many i.i.d. samples from $\mathcal{D}$. This leads to a concrete sample-based procedure for enforcing the target posterior success rate, which we formalize in the following theorem.

**Theorem 5 (Sample based tight privacy guarantee).** *Given a function $\mathcal{F} : \mathcal{X} \to \mathcal{O}$, let $\mathcal{M}_{\mathcal{D}_e}$ denote the mechanism that applies $\mathcal{F}$ and then adds noise drawn from a distribution $\mathcal{D}_e$ on $\mathcal{O}$. For each noise distribution $\mathcal{D}_e$ and each $\mathsf{x} \in \mathcal{X}$, define the measure*

$$H_{\mathcal{D}_e}\big(\mathcal{M}_{\mathcal{D}_e}(\mathsf{x}) = \mathsf{o}\big) \;:=\; \sum_{\rho(\mathsf{x},\hat{\mathsf{x}}')=1} \Pr(\mathsf{x} = \hat{\mathsf{x}}')\, \Pr\big(\mathcal{M}_{\mathcal{D}_e}(\hat{\mathsf{x}}') = \mathsf{o}\big), \tag{19}$$

*and let $W_{\mathcal{D}_e}$ be a reference distribution on $\mathcal{O}$. Define the privacy-loss function*

$$\mathcal{G}(\mathcal{D}_e, \hat{\mathsf{x}}) \;:=\; \frac{1}{\Pr(\mathsf{x} = \hat{\mathsf{x}})}\, \mathcal{D}_\alpha\big(P_{H_{\mathcal{D}_e}(\hat{\mathsf{x}})} \,\|\, P_{W_{\mathcal{D}_e}}\big),$$

*Suppose that for every $\mathsf{x} \in \mathcal{X}$ and every admissible $\mathcal{D}_e$ we have $\mathcal{G}(\mathcal{D}_e, \hat{\mathsf{x}}) \le B(\mathcal{D}_e)$ for some finite bound $B(\mathcal{D}_e)$, and an optimizer searching over $\mathcal{D}_e$ to minimize some metric $\kappa(\mathcal{D}_e)$ subject to the constraint*

$$\frac{1}{m}\sum_{i=1}^{m} \mathcal{G}(\mathcal{D}_e, \bar{\mathsf{x}}_i) \;\le\; r,$$

*for samples $\{\bar{\mathsf{x}}_i\}_{i=1}^m \sim D^m$. To make the mechanism satisfy $\delta_\rho$ privacy, set the privacy budget to*

$$r \;=\; (\delta_\rho)^\alpha.$$

*Then the privatization mechanism $\mathcal{M}_{\mathcal{D}_e}$ returned by Algorithm 1 guarantees that the true posterior success rate under the inference criterion $\rho$ is at most $\delta_\rho$.*

In the limit where the sample size $m \to \infty$, the order $\alpha \to \infty$, and $W_{\mathcal{D}_e}$ is chosen optimally, the budget choice $r = \delta_\rho^\alpha$ becomes asymptotically tight: the bound converges to the true posterior success rate under $\rho$, and $\mathcal{D}_e$ is the optimal noise distribution which minimizes $\kappa(\mathcal{D}_e)$ under the privacy budget constraint.

Before concluding this section, we point out a difference between Theorem 5 and Lemma 2. For an identification problem with a uniformly generated secret $\mathsf{x}$, both bounds are asymptotically tight, but for finite $\alpha < \infty$ the bound given by Lemma 2 is slightly better than proposed Theorem 5. Notice that

$$D_\alpha\big(\mathbf{1}_{\delta_\rho} \,\big\|\, \mathbf{1}_{\delta_{o,\rho}}\big) = (1-\delta_{o,\rho})\left(\frac{1-\delta_\rho}{1-\delta_{o,\rho}}\right)^\alpha + \delta_{o,\rho}\left(\frac{\delta_\rho}{\delta_{o,\rho}}\right)^\alpha > \delta_{o,\rho}\left(\frac{\delta_\rho}{\delta_{o,\rho}}\right)^\alpha$$

When $\mathsf{x}$ is uniform, we have $\delta_{o,\rho} = \Pr(\mathsf{x} = \hat{\mathsf{x}}) = 1/|\mathcal{X}|$, so the inequality (20) above directly implies

$$\delta_\rho \;\le\; \left(\mathbb{E}_{\mathsf{x}\sim D} \int_{\mathcal{O}} \frac{\Pr(\mathsf{x} = \hat{\mathsf{x}})^{\alpha-1}\, \Pr(\mathcal{M}(\mathsf{x}) = \mathsf{o})^\alpha}{\Pr(W = \mathsf{o})^{\alpha-1}}\, \mathrm{d}\mathsf{o}\right)^{\frac{1}{\alpha}}, \tag{20}$$

which is exactly the bound we obtain in Theorem 5. In the high-entropy regime with large $\alpha$, the second term $\delta_{o,\rho}\left(\frac{\delta_\rho}{\delta_{o,\rho}}\right)^\alpha$ generally dominates and thus non-asymptomatically the difference between Lemma 2 and Theorem 5 is minor. However, our result (Theorem 5) applies to *arbitrary* priors and general inference criteria $\rho$, while remaining asymptotically tight.

## 4 History-Dependent Adaptive Mechanisms

In this section, we address Problem 2 to control cumulative privacy risk in the presence of an *adaptive* adversary. We assume the user has access to the full interaction history, including all previous outputs, the adversary's queries, and the mechanisms realized in earlier rounds. We focus on the identification problem and our results are generalizable to general inference $\rho$ using a similar reasoning as shown in Section 3.2.

At round $t$, the adversary chooses the next query based on the entire interaction history, i.e., its past queries and the observed outputs. Such adaptivity can cause significantly more privacy leakage than in the static (non-adaptive) model. Instead of committing to all queries in advance, the adversary can repeatedly probe the most *vulnerable* aspects of the secret revealed by earlier outputs, and thereby tries to extract more information from the interaction.

As discussed previously, the black-box nature of both the leakage function and the adversarial strategy $\mathcal{Q}_{\text{adv}}$ presents a "parallel universe" challenge (Fig. 1). Because the user observes only a single realized transcript among infinite possibilities, we cannot simply simulate the combination of leakage function $\mathcal{F}(\cdot, \mathsf{v})$ and $\mathcal{Q}_{\text{adv}}$ and apply the results from Section 3. However, one key insight is that *before* the interactive query-response process begins, once a privacy-preserving mechanism $\mathsf{Alg}$ is determined by the user, although it is infeasible for the user to either predict or simulate *all* possible future transcripts, a same rule determined by $\mathsf{Alg}$ will be consistently applied to privatize the leakage in every branch of the interaction tree. As a result, it suffices to prove a *history-dependent* guarantee that holds at each step conditioned on the past; once such a guarantee is established, it automatically lifts to a global guarantee over the entire adaptive interaction. Before proceeding, we first formalize the adaptive interaction model introduced in Problem 2.

**Definition 8 (Adaptive Adversarial Composition).**

***Spaces and transcripts.*** *Given a total iteration number $T$ and three spaces: $\mathcal{X}$ is the* secret space*, $\mathcal{V}$ is the* query space *from which the adversary selects the query at each iteration, and $\mathcal{O}$ is the* leakage space.*A secret $\mathsf{x} \sim \mathcal{D}$ is sampled once at the beginning and reused throughout all $T$ rounds. For each round $t$, define the pre-round-$t$ transcript space as*

$$\mathcal{T}_t := \mathcal{V}^{t-1} \times \mathcal{O}^{t-1}.$$

*Accordingly, we write the realized pre-round-$t$ transcript as*

$$\tau_t := (\mathsf{v}_{<t}, \mathsf{o}_{<t}) \in \mathcal{T}_t, \quad \textit{here} \quad \mathsf{v}_{\le t} = (\mathsf{v}_1, \ldots, \mathsf{v}_{t-1}), \;\; \mathsf{o}_{<t} = (\mathsf{o}_1, \ldots, \mathsf{o}_{t-1}),$$

*to record the adversary's past queries and the user's past outputs. Let $\rho : \mathcal{X} \times \mathcal{X} \to \{0, 1\}$ be an inference criterion, where $\rho(\hat{\mathsf{x}}, \mathsf{x}) = 1$ indicates that the adversary's estimate $\widehat{\mathsf{x}}$ is deemed successful for the true secret $\mathsf{x}$.*

***Leakage and randomization.*** *Given $T$ processing functions*

$$\mathcal{F}_t : \mathcal{X} \times \mathcal{V} \to \mathcal{O}, \qquad t \in [T],$$

*and $T$ transcript-dependent randomization rules (corresponding to each step of* Alg *in Problem 1 and 2)*

$$\Sigma_t : \mathcal{V} \times \mathcal{T}_t \to \mathcal{O}, \qquad t \in [T],$$

*At round $t$, the user combines $\mathcal{F}_t$ and $\Sigma_t$ to form a randomized mechanism*

$$\mathcal{M}_t = \mathcal{F}_t + \Sigma_t.$$

***Adaptive interaction experiment.*** *An adversary's strategy $\mathcal{Q}_{adv}$ is a sequence of algorithm which can be either deterministic or randomized.*

$$\mathcal{A}dv_t : \mathcal{T}^t \to \mathcal{V}, \qquad t \in [T].$$

*The interaction proceeds for $t = 1, 2, \ldots, T$ as follows:*

1. *Given $\mathcal{T}_t$, the adversary outputs $\mathsf{v}_t \leftarrow \mathcal{A}dv_t$.*
2. *Given the secret $\hat{\mathsf{x}}$, the current query $\mathsf{v}_t$, and the transcript $\tau_t$, the user and returns $\mathsf{o}_t \leftarrow \mathcal{M}_t$ to the adversary.*

*At the end of the interaction, the adversary outputs an estimate $\hat{\mathsf{x}} \in \mathcal{X}$ of the secret $\mathsf{x}$, and we use $\mathcal{M}^{Adv}$ to represent the aggregation of $T$ adaptively-dependent leakage functions.*

In Definition 8, the per-round mechanism $\mathcal{M}_t$ may look complicated, but it admits a simple interpretation. Conditioned on the current query $\mathsf{v}_t$ and the current transcript $\tau_t$, the user's response is obtained by applying $\mathcal{F}_t$ and then adding noise chosen as a function of $(\mathsf{v}_t, \tau_t)$. In particular, if the randomization rule $\Sigma_t(\mathsf{v}_t, \tau_t)$ outputs a Gaussian distribution, then *conditionally on* $(\mathsf{v}_t, \tau_t)$ the mapping $\mathsf{x} \mapsto \mathcal{M}_t(\mathsf{x}, \mathsf{v}_t, \tau_t)$ is simply a Gaussian mechanism.

Our next goal is to design transcript-dependent randomization rules $\{\Sigma_t\}_{t=1}^T$ so that, for every adversary, the resulting composed interaction guarantees that the posterior success rate is bounded by the target level. To do so, we need a bridge between the induced output distributions and the adversary's posterior success rate. Lemma 2 provides one, but as discussed earlier, its $\alpha$-divergence specialization is not tight under non-uniform priors. Therefore, we instead work with Theorem 2.

At first glance, Theorem 2 appears to apply only to a *non-interactive* mechanism, since it does not explicitly model adversarially chosen queries. The key observation is that this is without loss of generality: we may equivalently regard the adversary's strategy $\mathcal{A}dv$ as part of the overall randomized mechanism. Concretely, imagine that the adversary hands the user a "strategy machine" implementing $\{\mathcal{A}dv_t\}_{t=1}^T$ as a black box. At round $t$, the user first runs this machine on the current transcript to generate the next query $\mathsf{v}_t$, and then runs $\mathcal{M}_t$ on

$(\mathsf{x}, \mathsf{v}_t, \tau_t)$ to generate the response $o_t$. At the end, the adversary receives the full transcript $(\mathsf{v}_{1:T}, \mathsf{o}_{1:T})$ and outputs an estimate $\hat{\mathsf{x}}$.

This reformulation is equivalent to the original interaction: the joint distribution of the transcript $(\mathsf{v}_{1:T}, \mathsf{o}_{1:T})$ is unchanged, and the adversary has access to exactly the same information when forming $\hat{\mathsf{x}}$. Consequently, any posterior-success bound that holds for the resulting randomized mapping from $\mathsf{x}$ to the final transcript applies equally to the original adaptive interaction.

For the reference distribution associated with the adversary's strategy $\{\mathcal{A}dv_t\}$, we may simply take an independent copy of the same strategy, since $\mathcal{A}dv_t$ is independent of the secret $\mathsf{x}$, which is the only requirement we impose on the reference. This leads to the theorem below.

**Theorem 6 (Posterior success bound for adaptive adversary).** *Under the setting of Definition 8, for any $\alpha > 1$, we have*

$$(\delta_\rho)^\alpha \le \mathbb{E}_{\mathsf{x}\sim D} \int_{\mathcal{O}^T \times \mathcal{V}^T} \Pr_{\mathcal{D}}(\mathsf{x})^{\alpha-1} \cdot \prod_{t=1}^{T} \frac{\big(\Pr\big(\mathcal{M}_t(\mathsf{x}, \mathsf{v}_t, \tau_t) = \mathsf{o}_t\big)\big)^{\alpha}}{\big(\Pr\big(W_t(\mathsf{v}_t, \tau_t) = \mathsf{o}_t\big)\big)^{\alpha-1}} \cdot \prod_{t=1}^{T} \Pr\big(\mathrm{Adv}_t(\tau_t) = \mathsf{v}_t\big)\, \mathrm{d}\mathsf{o}_{[1:T]}\, \mathrm{d}\mathsf{v}_{[1:T]}$$

*Here, we may take the reference distribution $\mathbb{P}_W$ for the adversary's queries to be the true law of $\mathrm{Adv}_t(\tau_t)$ itself, since it is independent of $\mathsf{x}$ conditionally on $\tau_t$.*

There is still a gap between Theorem 6 and an implementable procedure. As discussed above, the user observes only a single realized interaction (one "universe") induced by the adversary's adaptivity and thus lacks the information of $\Pr\big(\mathrm{Adv}_t(\tau_t) = \mathsf{v}_t\big)$. So the right-hand side of Theorem 6 is still not computable. We now return to the key insight mentioned at the beginning of this section. We consider further decomposing the global control problem in Theorem 6 into a collection of *local* subproblems, each of which depends only on the current branch (i.e., only on the currently observed $(\mathsf{v}_t, \tau_t)$ and the local mechanism at round $t$), then each "self" in each branch can solve its own local subproblem using only the information available in that branch. Since the same local rule is executed in all branches, combining these locally enforced constraints yields a global control on the full adaptive interaction. The next Theorem 7 shows a way to decompose and since the local control can lead to a global privacy guarantee.

**Theorem 7 (Adaptive Composition of $\alpha$-divergence).** *Under the setting of Definition 8 with a single secret $\mathsf{x} \sim \mathcal{D}$ reused across all $T$ rounds. For any possible query sequence, consider an* auxiliary reference process

$$W_{1:T}(\mathsf{v}_{1:T}, \tau_{1:T}) := (W_1(\mathsf{v}_1, \tau_1), \ldots, W_T(\mathsf{v}_T, \tau_T)) \in \mathcal{O}^T$$

*which is independent of $\mathsf{x}$. For each round $t \in \{1, \ldots, T\}$ and transcript $\tau_t = (\mathsf{v}_{<t}, \mathsf{o}_{<t})$, define the* prefix coefficient

$$A_{t-1}(\hat{\mathsf{x}}) := \Pr(\mathsf{x} = \hat{\mathsf{x}})^{\alpha-1} \left( \prod_{s=1}^{t-1} \frac{\Pr\big(\mathcal{M}_s(\hat{\mathsf{x}}, \mathsf{v}_s, \tau_s) = \mathsf{o}_s\big)}{\Pr\big(W_s(\mathsf{v}_s, \tau_s) = \mathsf{o}_s\big)} \right)^{\alpha}$$

*and*

$$D_t(\hat{\mathsf{x}}) := \mathcal{D}_\alpha(\mathcal{M}_t(\hat{\mathsf{x}}, \mathsf{v}_t, \tau_t) \,\|\, W_t(\mathsf{v}_t, \tau_t)).$$

*If for every possible transcript $\tau_t$ there exists such a choice of reference process $W_{1:T}(\mathsf{v}_{1:T})$ such that for all $t = 1, \ldots, T$,*

$$\mathbb{E}_{\mathsf{x}\sim D}\big[A_{t-1}(\mathsf{x})\, D_t(\mathsf{x})\big] \;\le\; r_t \cdot \mathbb{E}_{\mathsf{x}\sim D}\big[A_{t-1}(\mathsf{x})\big], \tag{21}$$

*then the final interactive mechanism $\mathcal{M}^{\mathcal{A}dv}$ satisfies for any $\mathcal{Q}_{adv}$ adversarial strategy, the optimal posterior success rate $\delta_\rho$ satisfies*

$$(\delta_\rho)^\alpha \;\le\; \prod_{t=1}^{T} r_t. \tag{22}$$

Theorem 7 makes adaptive randomization implementable in practice. Once the adversary's past queries and the user's past outputs have been realized, the transcript $\tau_t$ is fixed. The user therefore only needs to enforce the per-round condition (21) along the single interaction path that actually occurs, rather than over exponentially many counterfactual transcripts.

Note that the left-hand side of (21) is a *weighted* average of per-secret $\alpha$-divergences. Intuitively, $D_t(\hat{\mathsf{x}})$ quantifies how much information about the particular secret $\hat{\mathsf{x}}$ could be revealed at round $t$, while the prefix coefficient $A_{t-1}(\hat{\mathsf{x}})$ reweighs secrets according to how plausible they remain under the realized transcript. Consequently, secrets with larger $A_{t-1}(\hat{\mathsf{x}})$ contribute more to the cumulative leakage and must be protected more aggressively.

Finally, if $\mathcal{D}_e$ is simply some unrestricted noise distribution, for example, a Gaussian, the existence of a randomization solution is also guaranteed. As the variance of $\mathsf{e}$ grows, the output distribution of $\mathcal{M}_t(\cdot, \mathsf{v}_t, \tau_t)$ becomes increasingly insensitive to the secret, and hence $D_t(\hat{\mathsf{x}})$ decreases toward its minimum value. In the limit of infinite variance, the left-hand side of (21) approaches the right-hand side, so the inequality can always be met by choosing sufficiently large variance whenever $r_t > 1$.

*Remark 2 (Multiplicative vs. additive budgets for $\alpha$-divergence).* In our setting we work exclusively with $\alpha$-divergence, and the per-round budget *can be expressed* either multiplicatively or additively. When the $\alpha$-divergence bound is unrolled across $T$ rounds, the integrand contains a product of per-round likelihood ratios raised to the power $\alpha$. For this reason, it is often convenient to allocate per-round factors $\{r_t\}_{t=1}^T$ so that the overall budget takes the multiplicative form

$$r_{\text{tot}} \;=\; \prod_{t=1}^{T} r_t.$$

At the same time, an additive accounting is equally valid: defining $\varepsilon_t := \log r_t$ yields

$$\log r_{\text{tot}} \;=\; \sum_{t=1}^{T} \log r_t \;=\; \sum_{t=1}^{T} \varepsilon_t,$$

This log-additive view mirrors the familiar additive composition rule for Rényi-type privacy losses.

In the following, we further show the *tightness* of the privatization protocol derived by Theorem 7 for adversarially-adaptive composition: with proper selection of the reference process $W_{1:T}$, the equality of (22) is achievable. That is to say, given the iterative budget $\{r_1, r_2, \cdots, r_T\}$, our privacy solution *exactly* reaches the global budget $\prod_{t=1}^{T} r_t$.

Since the user can only choose the reference distribution $W_t$ *locally* at each round $t$ based on the realized transcript, given a per-round budget allocation $r_{1:T}$, a natural construction of the reference process $W_{1:T}$ is through a greedy design: $W_t$ is optimized to tighten the bound for the current step. Apparently, this greedy solution may not be jointly optimal for the *entire* interactive transcript but somewhat surprisingly, such locally-optimal choice indeed realizes the globally optimal reference process, as formalized in the following theorem.

**Theorem 8 (Local selection realizes the optimal reference process $W$).** *With the same notation of Theorem 7, define the per-round* locally optimal *reference $W_t$ (for each fixed realized $(\mathsf{v}_t, \tau_t)$) by*

$$\Pr\big(W_t(\mathsf{v}_t, \tau_t) = \mathsf{o}_t\big) \;\propto\; \left(\mathbb{E}_{\mathsf{x}\sim D}\left[\Pr_D(\mathsf{x})^{\alpha-1}\left(A_{t-1}(\mathsf{x}) \cdot \Pr\big(\mathcal{M}_t(\mathsf{x}, \mathsf{v}_t, \tau_t) = \mathsf{o}_t\big)\right)^{\alpha}\right]\right)^{1/\alpha}. \tag{23}$$

*Assume that the cumulative condition* (21) *holds with equality at every round. Then the resulting composed reference process $W_{1:T}$ is optimal which means it minimizes the upper bound in Lemma 2 among all admissible reference processes independent of $\mathsf{x}$. Consequently, when $\alpha \to \infty$, the resulting posterior-success bound is asymptotically tight.*

Finally, to further make Theorem 7 implementable from finite sampling, we establish Algorithm 2 by iteratively running Algorithm 1 based on locally-constructed privacy loss $\mathcal{G}_t$ defined in (95). We provide the provable guarantees of Algorithm 2 in Theorem 9.

**Theorem 9 (Provable guarantees of Algorithm 2).** *Define the* prefix coefficient

$$A_{t-1}(\hat{\mathsf{x}}) := \Pr(\mathsf{x} = \hat{\mathsf{x}})^{\alpha-1}\left(\prod_{s=1}^{t-1}\frac{\Pr\big(\mathcal{M}_s(\hat{\mathsf{x}}, \mathsf{v}_s, \tau_s) = \mathsf{o}_s\big)}{\Pr\big(W_s(\mathsf{v}_s, \tau_s) = \mathsf{o}_s\big)}\right)^{\alpha},$$

$$D_t(\hat{\mathsf{x}}) := \mathcal{D}_\alpha(\mathcal{M}_t(\hat{\mathsf{x}}, \mathsf{v}_t, \tau_t) \,\|\, W_t(\mathsf{v}_t, \tau_t)),$$

*and the privacy-loss function*

$$\mathcal{G}_t(\mathcal{D}_e, \hat{\mathsf{x}}) \;:= (A_{t-1}(\hat{\mathsf{x}})\,(D_t(\hat{\mathsf{x}}) - r_t) \tag{24}$$

**Algorithm 2** Randomization Optimizer under Adversarial Composition

**Input:** number of rounds $T$; leakage functions $\{\mathcal{F}_t : \mathcal{X} \times \mathcal{V} \to \mathcal{O}\}_{t=1}^{T}$; input distribution $D$; per-round $\alpha$-divergence budgets $\{r_t\}_{t=1}^{T}$; sample size $m$; global bound $B(\mathcal{D}_e)$.
Sample a secret input $\hat{\mathsf{x}} \sim D$.
Choose per-round targets $\{r'_t\}_{t=1}^{T}$, failure probabilities $\{\gamma_t\}_{t=1}^{T}$, and a fallback bound $R$ such that for every $t \in [T]$,

$$(1-\gamma_t)\, r'_t + \gamma_t\, R \le r_t.$$

**for** $t = 1, 2, \ldots, T$ **do**
  Receive adversarial parameter $v_t$.
  Define the *prefix coefficient*

$$A_{t-1}(\hat{\mathsf{x}}) := \Pr(\mathsf{x} = \hat{\mathsf{x}})^{\alpha-1} \left( \prod_{s=1}^{t-1} \frac{\Pr\big(\mathcal{M}_s(\hat{\mathsf{x}}, \mathsf{v}_s, \tau_s) = \mathsf{o}_s\big)}{\Pr\big(W_s(\mathsf{v}_s, \tau_s) = \mathsf{o}_s\big)} \right)^{\alpha}$$

$$D_t(\hat{\mathsf{x}}) := \mathcal{D}_\alpha(\mathcal{M}_t(\hat{\mathsf{x}}, \mathsf{v}_t, \tau_t) \,\|\, W_t(\mathsf{v}_t, \tau_t)),$$

  and

$$\mathcal{G}_t(\mathcal{D}_e, \hat{\mathsf{x}}) \;:= A_{t-1}(\hat{\mathsf{x}})\, D_t(\hat{\mathsf{x}})$$

  Take $\mathcal{F}(\cdot, \mathsf{v}_t), g_t(\mathcal{D}_e, \hat{\mathsf{x}}), D, r_t, B(\mathcal{D}_e)$ as the input for Algorithm 1
  Sample noise $e_t$ from the calibrated distribution.
  Output $\mathsf{o}_t \leftarrow \mathcal{F}_t(\hat{\mathsf{x}}, \mathsf{v}_t) + e_t$.
**end for**

*Suppose that for every $\hat{\mathsf{x}} \in \mathcal{X}$ and every admissible $\mathcal{D}_e$ we have $\mathcal{G}_t(\mathcal{D}_e, \hat{\mathsf{x}}) \le B(\mathcal{D}_e)$ for some finite bound $B(\mathcal{D}_e)$ given per-round budget $0$(It's because the the second part of the privacy loss function minus $r_t$), Let $\mathcal{M}_{[1:T]}$ be the interactive mechanism produced by Algorithm 2. Then for any $\mathcal{A}dv$, the optimal posterior success rate $\delta_\rho$ satisfies*

$$\delta_\rho^{\alpha} \le \prod_{t=1}^{T} r_t.$$

Finally, we give an example of how $B(\mathcal{D}_e)$ can be computed. Assume that $\|\mathcal{F}\|_2 \le b$ for all $\hat{\mathsf{x}} \in \mathcal{X}$ and $\mathsf{v} \in \mathcal{V}$. Consider adding Gaussian noise, and define the (empirical) local reference

$$W \;\propto\; \left( \frac{1}{m} \sum_{i=1}^{m} \Pr(\mathsf{x} = \bar{\mathsf{x}}_i)^{\alpha-1} \left( \prod_{s=1}^{t-1} \frac{\Pr\big(\mathcal{M}_s(\bar{\mathsf{x}}_i, \mathsf{v}_s, \tau_s) = \mathsf{o}_s\big)}{\Pr\big(W_s(\mathsf{v}_s, \tau_s) = \mathsf{o}_s\big)} \cdot \Pr\big(\mathcal{M}_t(\bar{\mathsf{x}}_i, \mathsf{v}_t, \tau_t) = \mathsf{o}_t\big) \right)^{\alpha} \right)^{\frac{1}{\alpha}}.$$

The quantity $A_{t-1}$ can be bounded along the realized transcript: the denominator terms $\Pr(W_s(\mathsf{v}_s, \tau_s) = \mathsf{o}_s)$ are fixed once the transcript is fixed, and the numerator terms admit an upper bound under the assumption $\|\mathcal{F}\|_2 \le b$.

Therefore the normalizing factor is bounded. In particular,

$$\begin{aligned}
&\int_{\mathcal{O}} \left( \frac{1}{m} \sum_{i=1}^{m} \Pr(\mathsf{x} = \bar{\mathsf{x}}_i)^{\alpha-1} \left( \prod_{s=1}^{t-1} \frac{\Pr\big(\mathcal{M}_s(\bar{\mathsf{x}}_i, \mathsf{v}_s, \tau_s) = \mathsf{o}_s\big)}{\Pr\big(W_s(\mathsf{v}_s, \tau_s) = \mathsf{o}_s\big)} \cdot \Pr\big(\mathcal{M}_t(\bar{\mathsf{x}}_i, \mathsf{v}_t, \tau_t) = \mathsf{o}\big) \right)^{\alpha} \right)^{\frac{1}{\alpha}} \mathsf{do} \\
&\quad \leq \frac{\max_{\hat{\mathsf{x}} \in \mathcal{X}} A_{t-1}(\hat{\mathsf{x}})}{\min_{\hat{\mathsf{x}} \in \mathcal{X}} \Pr(\mathsf{x} = \hat{\mathsf{x}})} \int_{\mathcal{O}} \max_{\hat{\mathsf{x}} \in \mathcal{X}} \Pr(\mathsf{x} = \hat{\mathsf{x}}) \ \Pr\big(\mathcal{M}_t(\hat{\mathsf{x}}, \mathsf{v}_t, \tau_t) = \mathsf{o}\big) \, \mathsf{do} \\
&\quad \leq \frac{\max_{\hat{\mathsf{x}} \in \mathcal{X}} A_{t-1}(\hat{\mathsf{x}})}{\min_{\hat{\mathsf{x}} \in \mathcal{X}} \Pr(\mathsf{x} = \hat{\mathsf{x}})}. \qquad (25)
\end{aligned}$$

Hence the normalizing constant is bounded. Next, by Hölder's inequality we have

$$\frac{1}{\left(\frac{1}{m} \sum_{i=1}^{m} a_i^{\alpha}\right)^{\frac{\alpha-1}{\alpha}}} \ \leq \ \sum_{i=1}^{m} \frac{1}{m \, a_i^{\alpha-1}}. \qquad (26)$$

Omitting $(\mathsf{v}_t, \tau_t)$ since they are fixed, then for any $\hat{\mathsf{x}} \in \mathcal{X}$, we obtain

$$\int_{\mathcal{O}} \Pr\big(\mathcal{M}_t(\hat{\mathsf{x}}) = \mathsf{o}\big)^{\alpha} \ \Pr\big(W_t = \mathsf{o}\big)^{1-\alpha} \, \mathsf{do} \ \leq \ C^{\alpha} \cdot \max_{\hat{\mathsf{x}}' \in \mathcal{X}} \mathcal{D}_{\alpha}\big(\mathcal{M}_t(\hat{\mathsf{x}}) \,\|\, \mathcal{M}_t(\hat{\mathsf{x}}')\big), \qquad (27)$$

where $C$ is a normalization constant, and therefore the privacy-loss function is bounded.

## 5 Mechanism Agnosticism Adaptive Composition

In this section, we further study the composition in the *mechanism-agnostic* setup. In this setting, across each round $t \in [T]$, a mechanism $\mathcal{M}_t$ must determine its internal randomization using only the information restricted to the privacy budget allocated to that round and the specifically observed leakage function $\mathcal{F}(\cdot, \mathsf{v}_t)$. Formally speaking, the difference between the model in this section and Definition 8 is that $\tau_t$ is not allowed to be an input to the mechanism.

A key question we want to address in this section is that under what kind of separate measurement of marginal leakage, they can be finally aggregated form an upper bound of the cumulative privacy risk, independent of any potential adversarial strategy that controls these leakage functions behind. In the following we present two kinds of measurements established by KL-divergence and $\alpha$-divergence that enables the *mechanism-agnostic* composition, as characterized in the following two theorems.

**Theorem 10 (KL agnostic composition).** *For each $t \in [T]$, any $\mathsf{v}_t$, and an arbitrary $W_t(\mathsf{v}_t)$ as reference, if both the following worse-case KL divergence and the average-case KL divergence are bounded:*

$$\mathcal{D}_{KL}(P_{\mathcal{M}_t(\hat{\mathsf{x}}, \mathsf{v}_t)} \| P_{W_t(\mathsf{v}_t)}) \leq R_t, \quad \forall \hat{\mathsf{x}} \in \mathcal{X}, \mathsf{v}_t \in \mathcal{V}_t \qquad (28)$$

$$\mathbb{E}_{\mathsf{x} \sim D} \mathcal{D}_{KL}(P_{\mathcal{M}_t(\mathsf{x}, \mathsf{v}_t)} \| P_{W_t(\mathsf{v}_t)}) \leq r_t, \quad \forall \mathsf{v}_t \in \mathcal{V}_t \qquad (29)$$

*then the optimal posterior success rate $\delta_\rho$ of the final $T$-aggregated interactive mechanism $\mathcal{M}^{\mathcal{A}dv}$ for any $\mathcal{A}dv$ satisfies*

$$\mathcal{D}_{KL}(\mathbf{1}_{\delta_\rho} \| \mathbf{1}_{\delta_{o,\rho}}) \leq B_T,$$

*where $B_t$ is defined in a recursive form*

$$B_1 = r_1, \tag{30}$$

$$B_t = B_{t-1} + \min\{\sqrt{2B_{t-1}}R_t + r_t, R_T\}, t \geq 2. \tag{31}$$

In Theorem 10, $R_t$ behaves as a universal, input-independent upper bound. From (31), the increment between $B(t)$ and $B(t-1)$ in two consecutive iterations is upper bounded by either $\sqrt{2B(t-1)}R_t + r_t$, that is, the average increment $r_t$ plus a worst-case increment $R_t$ scaled by a factor $\sqrt{2B(t-1)}$ depending on the current accumulated risk, or simply by the worst-case increment $R_t$. Asymptotically, $B_T$ increases quadratically in $T$.

We also show the result through $\alpha$-divergence in the following theorem:

**Theorem 11 ($\alpha$ agnostic composition).** *For each $t \in [T]$, any $\mathsf{v}_t$, and an arbitrary $W_t(\mathsf{v}_t)$ as reference, if the $\alpha$-divergence between $\mathbb{P}_{\mathcal{M}_t(\cdot,\mathsf{v}_t)}$ and $\mathbb{P}_{W_t(\mathsf{v}_t)}$ is bounded as*

$$\mathbb{E}_{\mathsf{x}\sim D} \Pr_D(\mathsf{x})^{p_t-1} D_{p_t}(\mathbb{P}_{\mathcal{M}_t(\mathsf{x},\mathsf{v}_t)} \| \mathbb{P}_{W_t(\mathsf{v}_t)}) \leq r_t, \quad \forall \mathsf{v}_t \in \mathcal{V}_t, \tag{32}$$

*then the the optimal posterior success rate $\delta_\rho$ of the final $T$-aggregated interactive mechanism $\mathcal{M}^{\mathcal{A}dv}$ for any $\mathcal{A}dv$ satisfies*

$$\delta_\rho^\alpha \leq \prod_{t=1}^{T} r_t^{\frac{1}{p_t}}.$$

*Here $p_{[1:t]}$ are positive real numbers satisfying the following conditions:*

$$\sum_{i=1}^{T} \frac{1}{p_t} = 1 \tag{33}$$

Compared with KL-agnostic composition in Theorem 10 which requires a global worst-case guarantee $R_t$, in a Hölder's inequality style $\alpha$-agnostic composition in Theorem 11 does not explicitly require such a worst-case guarantee but only normal marginal $\alpha$-divergence. However, when the iteration number $T$ is large, the corresponding $\alpha$-divergence control likewise becomes effectively dominated by the $\max_t r_t$. This is because $\sum_{i=1}^{T} \frac{1}{p_t} = 1$, which implies that there must exist some $p_t \geq T \to \infty$. As expected, compared to the history-dependent composition in Section 4, the mechanism-agnostic one forces the user to add larger noise per-iteration, since the adversary can concentrate its attack on the most vulnerable region of the secret space, while the user cannot adapt the mechanism to counteract that vulnerability along the realized branch.

## 6 Experiments and Additional Related Work

To demonstrate the practical utility of our theory and algorithms, we evaluate the protection of secret keys in two realistic side-channel scenarios: power consumption leakage in an AES key generator and timing leakage in RSA encryption. We consider both static and composition settings. In our experiments, we model the privatization mechanism $\mathsf{e}$ as independent Gaussian noise. We analyze how the optimal Gaussian distribution $\mathcal{D}_e$ (characterized by minimal variance) is influenced by entropy (varying secret bit-lengths), prior success rate (across different inference tasks, including identification and approximate reconstruction), and the composition number $T$. Detailed experimental results are provided in Appendix P.

In Appendix Q, we derive analogous results based on Kullback-Leibler (KL) divergence, serving as variants to our $\alpha$-divergence and prior-reference-weighted $\alpha$-information bounds. While KL divergence is a standard metric in information privacy, we show that these variants, despite their simpler forms, yield bounds on the posterior rate $\delta_\rho$ that can be exponentially looser. This gap is illustrated in Figures 2–5 in Appendix P and is particularly pronounced in high-entropy regimes.

Finally, we discuss additional related work.

**Looseness of KL-divergence and Mutual Information.** Mutual Information (MI) is often the de facto metric for quantifying the leakage of entropic secrets. However, independent of our findings regarding the looseness of posterior inference bounds, existing literature also identifies other limitations of MI. For instance, MI generally captures average leakage rather than a high-probability sense [38] or the worst case [4]. Furthermore, it may not accurately reflect the empirical success rate of practical attackers [40].

**Leakage-Resilient Cryptography.** In contrast to our approach, which applies a universal randomization mechanism to any (potentially black-box) leakage, Leakage-Resilient Cryptography [31] focuses on redesigning cryptographic primitives to mitigate leakage directly. These methods typically rely on specific assumptions about the leakage channel, such as the $t$-probing model (where an adversary observes at most $t$ wires) [17] or the bounded-leakage model [12]. Common techniques include masking and periodic key updates. Hardware- and software-level countermeasures like constant-time encryption [24] also exist, they are often computationally expensive or platform-specific.

## 7 Conclusion and Future Prospects

In this paper, we focus on entropic secrets and establish an asymptotically-tight Bayesian analysis framework to optimize privacy-preserving solutions for both static and adversarially-adaptive composition settings given black-box leakage functions. In contrast to the strong impossibility results for universal privatization in the regime of *Input-Independent Indistinguishability* (III) for the prior-free worst case, we demonstrate that secret entropy can be tightly exploited to

amplify privacy, enhancing adversarial inference hardness even in a black-box manner with probable guarantees.

A straightforward and promising future direction for our results is software/hardware co-design for side-channel leakage mitigation. Given a specific leakage channel (e.g., power or timing consumption from a hardware environment) and an objective task (e.g., encryption), tight black-box privatization enables the free exploration of different processing functions (e.g., cryptographic protocols) to identify the sharpest utility-privacy tradeoff.

Several technical questions also remain open within our framework. For example, it is unclear whether a more powerful $f$-function exists, improving upon the polynomial $f(z) = z^\alpha$, that can yield faster and more stable convergence to $\delta_\rho$. Additionally, from an operational perspective, we have not systematically explored the construction but the just assumed the existence of numerical solvers for the constrained optimization problem (13) in Algorithm 1.

Finally, regarding mechanism-agnostic composition, we note that computing tight composition for III-style worst-case guarantees is known to be #P-hard [28]. A natural open problem is whether this complexity result also holds for entropic secrets. Furthermore, beyond the KL- and $\alpha$-divergence-based marginal privacy risk measurements ( Section 5), exploring other constructions that enable such mechanism-agnostic composition remains an interesting avenue for future work.

## References


1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. pp. 308–318 (2016)
2. Abouelmehdi, K., Beni-Hessane, A., Khaloufi, H.: Big healthcare data: preserving security and privacy. Journal of big data **5**(1), 1–18 (2018)
3. Aga, S., Narayanasamy, S.: Invisimem: Smart memory defenses for memory bus side channel. ACM SIGARCH Computer Architecture News **45**(2), 94–106 (2017)
4. Alvim, M.S., Andrés, M.E., Chatzikokolakis, K., Degano, P., Palamidessi, C.: Differential privacy: on the trade-off between utility and information leakage. In: International workshop on formal aspects in security and trust. pp. 39–54. Springer (2011)
5. Bun, M., Steinke, T.: Concentrated differential privacy: Simplifications, extensions, and lower bounds. In: Theory of cryptography conference. pp. 635–658. Springer (2016)
6. Coppens, B., Verbauwhede, I., De Bosschere, K., De Sutter, B.: Practical mitigations for timing-based side-channel attacks on modern x86 processors. In: 2009 30th IEEE symposium on security and privacy. pp. 45–60. IEEE (2009)
7. Dodis, Y., Smith, A.: Entropic security and the encryption of high entropy messages. In: Theory of Cryptography Conference. pp. 556–577. Springer (2005)
8. Dong, J., Roth, A., Su, W.J.: Gaussian differential privacy. Journal of the Royal Statistical Society Series B: Statistical Methodology **84**(1), 3–37 (2022)
9. Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., Naor, M.: Our data, ourselves: Privacy via distributed noise generation. In: Advances in Cryptology-EUROCRYPT 2006: 24th Annual International Conference on the Theory and

Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28-June 1, 2006. Proceedings 25. pp. 486–503. Springer (2006)
10. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3. pp. 265–284. Springer (2006)
11. Dwork, C., Rothblum, G.N., Vadhan, S.: Boosting and differential privacy. In: 2010 IEEE 51st annual symposium on foundations of computer science. pp. 51–60. IEEE (2010)
12. Dziembowski, S., Pietrzak, K.: Leakage-resilient cryptography. In: 2008 49th Annual IEEE Symposium on Foundations of Computer Science. pp. 293–302. IEEE (2008)
13. Fano, R.M.: Fano inequality. Scholarpedia **3**(10), 6648 (2008)
14. Ghazi, B., Kumar, R., Manurangsi, P.: Differentially private clustering: Tight approximation ratios. Advances in Neural Information Processing Systems **33**, 4040–4054 (2020)
15. Goldwasser, S., Micali, S.: Probabilistic encryption. Journal of computer and system sciences **28**(2), 270–299 (1984)
16. Hardt, M., Rothblum, G.N.: A multiplicative weights mechanism for privacy-preserving data analysis. In: 2010 IEEE 51st annual symposium on foundations of computer science. pp. 61–70. IEEE (2010)
17. Ishai, Y., Sahai, A., Wagner, D.: Private circuits: Securing hardware against probing attacks. In: Annual International Cryptology Conference. pp. 463–481. Springer (2003)
18. Issa, I., Wagner, A.B., Kamath, S.: An operational approach to information leakage. IEEE Transactions on Information Theory **66**(3), 1625–1657 (2019)
19. Kairouz, P., Oh, S., Viswanath, P.: The composition theorem for differential privacy. In: International conference on machine learning. pp. 1376–1385. PMLR (2015)
20. Kasiviswanathan, S.P., Smith, A.: On the'semantics' of differential privacy: A bayesian formulation. Journal of Privacy and Confidentiality **6**(1) (2014)
21. Kifer, D., Machanavajjhala, A.: Pufferfish: A framework for mathematical privacy definitions. ACM Transactions on Database Systems (TODS) **39**(1), 1–36 (2014)
22. Köpf, B., Basin, D.: Automatically deriving information-theoretic bounds for adaptive side-channel attacks. Journal of Computer Security **19**(1), 1–31 (2011)
23. Lin, Z., Gopi, S., Kulkarni, J., Nori, H., Yekhanin, S.: Differentially private synthetic data via foundation model APIs 1: Images. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=YEhQs8POIo
24. Liu, J., Jager, T., Kakvi, S.A., Warinschi, B.: How to build time-lock encryption. Designs, Codes and Cryptography **86**(11), 2549–2586 (2018)
25. Lyu, M., Su, D., Li, N.: Understanding the sparse vector technique for differential privacy. Proceedings of the VLDB Endowment **10**(6), 637–648 (2017)
26. Massey, J.L.: Guessing and entropy. In: Proceedings of 1994 IEEE International Symposium on Information Theory. p. 204. IEEE (1994)
27. Mironov, I.: Rényi differential privacy. In: 2017 IEEE 30th computer security foundations symposium (CSF). pp. 263–275. IEEE (2017)
28. Murtagh, J., Vadhan, S.: The complexity of computing the optimal composition of differential privacy. In: Theory of Cryptography Conference. pp. 157–175. Springer (2015)

29. Nissim, K., Raskhodnikova, S., Smith, A.: Smooth sensitivity and sampling in private data analysis. In: Proceedings of the thirty-ninth annual ACM symposium on Theory of computing. pp. 75–84 (2007)
30. Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., Erlingsson, U.: Scalable private learning with PATE. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=rkZB1XbRZ
31. Pietrzak, K.: A leakage-resilient mode of operation. In: Annual International Conference on the Theory and Applications of Cryptographic Techniques. pp. 462–482. Springer (2009)
32. Ponomareva, N., Hazimeh, H., Kurakin, A., Xu, Z., Denison, C., McMahan, H.B., Vassilvitskii, S., Chien, S., Thakurta, A.G.: How to dp-fy ml: A practical guide to machine learning with differential privacy. Journal of Artificial Intelligence Research **77**, 1113–1201 (2023)
33. Randolph, M., Diehl, W.: Power side-channel attack analysis: A review of 20 years of study for the layman. Cryptography **4**(2), 15 (2020)
34. Russell, A., Wang, H.: How to fool an unbounded adversary with a short key. IEEE Transactions on Information Theory **52**(3), 1130–1140 (2006)
35. Sason, I., Verdú, S.: $f$-divergence inequalities. IEEE Transactions on Information Theory **62**(11), 5973–6006 (2016)
36. Shannon, C.E.: Communication theory of secrecy systems. The Bell system technical journal **28**(4), 656–715 (1949)
37. Sibson, R.: Information radius. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete **14**(2), 149–160 (1969)
38. Smith, G.: On the foundations of quantitative information flow. In: International Conference on Foundations of Software Science and Computational Structures. pp. 288–302. Springer (2009)
39. Sridhar, M., Xiao, H., Devadas, S.: Pac-private algorithms. In: 2025 IEEE Symposium on Security and Privacy (SP). pp. 3839–3857. IEEE (2025)
40. Standaert, F.X., Malkin, T.G., Yung, M.: A unified framework for the analysis of side-channel key recovery attacks. In: Annual international conference on the theory and applications of cryptographic techniques. pp. 443–461. Springer (2009)
41. Van Den Hooff, J., Lazar, D., Zaharia, M., Zeldovich, N.: Vuvuzela: Scalable private messaging resistant to traffic analysis. In: Proceedings of the 25th Symposium on Operating Systems Principles. pp. 137–152 (2015)
42. Verdú, S.: $\alpha$-mutual information. In: 2015 Information Theory and Applications Workshop (ITA). pp. 1–6. IEEE (2015)
43. Wainwright, M.J.: High-dimensional statistics: A non-asymptotic viewpoint, vol. 48. Cambridge university press (2019)
44. Xiao, H., Devadas, S.: Pac privacy: Automatic privacy measurement and control of data processing. In: Annual International Cryptology Conference. pp. 611–644. Springer (2023)
45. Xiao, H., Suh, G.E., Devadas, S.: Formal privacy proof of data encoding: The possibility and impossibility of learnable encryption. In: Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security. pp. 1834–1848 (2024)
46. Xiao, X., Tao, Y.: Output perturbation with query relaxation. Proceedings of the VLDB Endowment **1**(1), 857–869 (2008)
47. Young, G.A., Smith, R.L., Smith, R.L.: Essentials of statistical inference, vol. 16. Cambridge University Press (2005)
48. Zhang, J., Chen, C., Cui, J., Li, K.: Timing side-channel attacks and countermeasures in cpu microarchitectures. ACM Computing Surveys **56**(7), 1–40 (2024)

## A Proof of Proposition 1

Taking the $\frac{1}{\alpha}$ power of both sides from Lemma 2 and let $\alpha \to \infty$, then the left-hand-side converges to $\frac{\delta_\rho}{\delta_{o,\rho}}$ and the right-hand-side converges to $\exp\{I_\infty^{sib}(\mathsf{x}; \mathcal{F}(\mathsf{x})\}$. Here $I_\infty^{sib}(\mathsf{x}; \mathcal{F}(\mathsf{x}))$ is Sibson's mutual information of order infinity (Definition 5). Notice that for any $\hat{\mathsf{x}} \in \mathcal{X}$, $\Pr_{\mathsf{x}\sim\mathcal{D}}(\mathsf{x} = \hat{\mathsf{x}}) = \frac{1}{|\mathcal{X}|}$

Therefore,

$$\delta_\rho = \delta_{o,\rho} \cdot \int_{\mathcal{O}} \max_{\hat{\mathsf{x}}\in\mathcal{X}} \Pr\big(\mathcal{F}(\hat{\mathsf{x}}) = \mathsf{o}\big)\, \mathrm{d}\mathsf{o} \tag{34}$$

$$= \int_{\mathcal{O}} \max_{\hat{\mathsf{x}}\in\mathcal{X}} \Pr(\mathsf{x} = \hat{\mathsf{x}})\ \Pr\big(\mathcal{F}(\hat{\mathsf{x}}) = \mathsf{o}\big)\, \mathrm{d}\mathsf{o} \tag{35}$$

$$= \int_{\mathcal{O}} \max_{\hat{\mathsf{x}}\in\mathcal{X}} \Pr\big(\mathsf{x} = \hat{\mathsf{x}},\ \mathcal{F}(\hat{\mathsf{x}}) = \mathsf{o}\big)\, \mathrm{d}o. \tag{36}$$

## B Proof of Lemma 3

Let $g : \mathcal{O} \to \mathcal{X}$ be any deterministic decision rule, where $g(\mathsf{o})$ is the guess when the observation is $\mathsf{o}$. The success event is $\{\rho(\mathsf{x}, g(\mathsf{o})) = 1\}$, so

$$\begin{aligned}\Pr\big(\rho(\mathsf{x}, g(\mathsf{o})) = 1\big) &= \int_{\mathcal{O}} \sum_{\bar{\mathsf{x}}\in\mathcal{X}} \mathbf{1}_{\{\rho(\bar{\mathsf{x}}, g(\mathsf{o}))=1\}}\ \Pr(\mathsf{x} = \bar{\mathsf{x}}, \mathsf{o})\, \mathrm{d}\mathsf{o} \\ &= \int_{\mathcal{O}} \sum_{\bar{\mathsf{x}}\in\mathcal{X}} \mathbf{1}_{\{\rho(\bar{\mathsf{x}}, g(\mathsf{o}))=1\}}\ \Pr(\mathsf{x} = \bar{\mathsf{x}}) \Pr(\mathsf{o} \mid \mathsf{x} = \bar{\mathsf{x}})\, \mathrm{d}\mathsf{o}.\end{aligned}$$

For each fixed $\mathsf{o}$, the inner sum is exactly $H(g(\mathsf{o}), \mathsf{o})$ by the definition of $H$, hence

$$\Pr\big(\rho(\mathsf{x}, g(\mathsf{o})) = 1\big) = \int_{\mathcal{O}} H(g(\mathsf{o}), \mathsf{o})\, \mathrm{d}\mathsf{o}\ \le\ \int_{\mathcal{O}} \max_{\hat{\mathsf{x}}\in\mathcal{X}} H(\hat{\mathsf{x}}, \mathsf{o})\, \mathrm{d}\mathsf{o},$$

since $H(g(\mathsf{o}), \mathsf{o}) \le \max_{\hat{\mathsf{x}}} H(\hat{\mathsf{x}}, \mathsf{o})$ for every $\mathsf{o}$.

Conversely, for each $\mathsf{o} \in \mathcal{O}$ we can choose $g^\star(\mathsf{o}) \in \arg\max_{\hat{\mathsf{x}}\in\mathcal{X}} H(\hat{\mathsf{x}}, \mathsf{o})$ (choosing any maximizer when there are ties). For this $g^\star$ we have $H(g^\star(\mathsf{o}), \mathsf{o}) = \max_{\hat{\mathsf{x}}} H(\hat{\mathsf{x}}, \mathsf{o})$ for all $\mathsf{o}$, and thus

$$\Pr\big(\rho(\mathsf{x}, g^\star(\mathsf{o})) = 1\big) = \int_{\mathcal{O}} \max_{\hat{\mathsf{x}}\in\mathcal{X}} H(\hat{\mathsf{x}}, \mathsf{o})\, \mathrm{d}\mathsf{o}.$$

Combining the upper bound with the equality achieved by $g^\star$ proves the claim.

## C Proof of Lemma 4

Notice that

$$\left(\sum_{\mathsf{x}\in\mathcal{X}} H(\mathsf{x}, \mathsf{o})^\alpha\right)^{1/\alpha}\ \ge\ \left(\max_{\mathsf{x}\in\mathcal{X}} H(\mathsf{x}, \mathsf{o})^\alpha\right)^{1/\alpha}\ =\ \max_{\mathsf{x}\in\mathcal{X}} H(\mathsf{x}, \mathsf{o}), \tag{37}$$

take integral over $\mathcal{O}$ we directly get

$$\int_{\mathcal{O}} \left( \sum_{\mathsf{x} \in \mathcal{X}} H(\mathsf{x}, \mathsf{o})^{\alpha} \right)^{1/\alpha} \mathrm{d}\mathsf{o} > \int_{\mathcal{O}} \max_{\mathsf{x} \in \mathcal{X}} H(\mathsf{x}, \mathsf{o}) \, \mathrm{d}\mathsf{o}$$

## D Proof of Lemma 5

Fix $\alpha > 1$ and define, for each $\mathsf{o} \in \mathcal{O}$,

$$G(\mathsf{o}) \;:=\; \left( \sum_{\mathsf{x} \in \mathcal{X}} H(\mathsf{x}, \mathsf{o})^{\alpha} \right)^{1/\alpha}.$$

Then the left-hand side of the inequality is simply $\int_{\mathcal{O}} G(\mathsf{o}) \, \mathrm{d}\mathsf{o}$. We insert the reference density $W$ and apply Hölder's inequality with exponents $\alpha$ and $\frac{\alpha}{\alpha-1}$:

$$\begin{aligned}
\int_{\mathcal{O}} G(\mathsf{o}) \, \mathrm{d}\mathsf{o} &= \int_{\mathcal{O}} G(\mathsf{o}) \, \Pr(W = \mathsf{o})^{\frac{\alpha-1}{\alpha}} \, \Pr(W = \mathsf{o})^{-\frac{\alpha-1}{\alpha}} \, \mathrm{d}\mathsf{o} \\
&\leq \left( \int_{\mathcal{O}} \left( G(\mathsf{o}) \, \Pr(W = \mathsf{o})^{\frac{\alpha-1}{\alpha}} \right)^{\alpha} \mathrm{d}\mathsf{o} \right)^{1/\alpha} \left( \int_{\mathcal{O}} \left( \Pr(W = \mathsf{o})^{-\frac{\alpha-1}{\alpha}} \right)^{\frac{\alpha}{\alpha-1}} \mathrm{d}\mathsf{o} \right)^{\frac{\alpha-1}{\alpha}} \\
&= \left( \int_{\mathcal{O}} \frac{G(\mathsf{o})^{\alpha}}{\Pr(W = \mathsf{o})^{\alpha-1}} \mathrm{d}\mathsf{o} \right)^{1/\alpha} \left( \int_{\mathcal{O}} \Pr(W = \mathsf{o}) \, \mathrm{d}\mathsf{o} \right)^{\frac{\alpha-1}{\alpha}}.
\end{aligned}$$

Since $W$ is a probability distribution, $\int_{\mathcal{O}} \Pr(W = \mathsf{o}) \, \mathrm{d}\mathsf{o} = 1$, so we obtain

$$\int_{\mathcal{O}} G(\mathsf{o}) \, \mathrm{d}\mathsf{o} \;\leq\; \left( \int_{\mathcal{O}} \frac{G(\mathsf{o})^{\alpha}}{\Pr(W = \mathsf{o})^{\alpha-1}} \mathrm{d}\mathsf{o} \right)^{1/\alpha} = \left( \int_{\mathcal{O}} \frac{\sum_{\mathsf{x} \in \mathcal{X}} H(\mathsf{x}, \mathsf{o})^{\alpha}}{\Pr(W = \mathsf{o})^{\alpha-1}} \mathrm{d}\mathsf{o} \right)^{1/\alpha},$$

which is the first inequality.

For the expectation form, note that

$$\sum_{\mathsf{x} \in \mathcal{X}} H(\mathsf{x}, \mathsf{o})^{\alpha} = \sum_{\mathsf{x} \in \mathcal{X}} \Pr(\mathsf{x}) \, \frac{H(\mathsf{x}, \mathsf{o})^{\alpha}}{\Pr(\mathsf{x})},$$

so

$$\int_{\mathcal{O}} \frac{\sum_{\mathsf{x} \in \mathcal{X}} H(\mathsf{x}, \mathsf{o})^{\alpha}}{\Pr(W = \mathsf{o})^{\alpha-1}} \, \mathrm{d}\mathsf{o} = \mathbb{E}_{\mathsf{x} \sim D} \int_{\mathcal{O}} \frac{H(\mathsf{x}, \mathsf{o})^{\alpha} / \Pr(\mathsf{x})}{\Pr(W = \mathsf{o})^{\alpha-1}} \, \mathrm{d}\mathsf{o},$$

which yields the second equality.

Finally, Hölder's inequality is tight if and only if there exists a constant $\lambda > 0$ such that

$$G(\mathsf{o})^{\alpha} \;=\; \lambda \, \Pr(W = \mathsf{o})^{\alpha} \quad \text{for almost every } \mathsf{o},$$

which is equivalent to

$$\sum_{\mathsf{x} \in \mathcal{X}} H(\mathsf{x}, \mathsf{o})^{\alpha} \;=\; \lambda \, \Pr(W = \mathsf{o})^{\alpha} \quad \text{a.e. } \mathsf{o}.$$

Rewriting this gives the stated condition that the bound is tight when $W$ is chosen proportional to the $\alpha$-norm of $H(\cdot, \mathsf{o})$:

$$\Pr(W = \mathsf{o}) \;\propto\; \left(\sum_{\mathsf{x}\in\mathcal{X}} H(\mathsf{x}, \mathsf{o})^{\alpha}\right)^{1/\alpha}.$$

## E Proof of Theorem 1

For any admissible noise distribution $\mathcal{D}_e$, write

$$\mathcal{G}(\mathcal{D}_e, \mathsf{x}) = \mathcal{D}_\alpha\big(P_{\mathcal{M}_{\mathcal{D}_e}(\mathsf{x})} \,\|\, P_{W_{\mathcal{D}_e}}\big), \qquad \mu(\mathcal{D}_e) := \mathbb{E}_{\mathsf{x}\sim D}\big[\mathcal{G}(\mathcal{D}_e, \mathsf{x})\big].$$

Algorithm 1 draws two i.i.d. samples $\mathsf{X} = \{\bar{\mathsf{x}}_1, \ldots, \bar{\mathsf{x}}_m\}$ and $\mathsf{X}' = \{\bar{\mathsf{x}}'_1, \ldots, \bar{\mathsf{x}}'_m\}$ from $D$, and then:

1. uses only $\mathsf{X}$ to solve the empirical constraint (13) and obtain an initial candidate $\mathcal{D}_e$; 2. uses only the independent sample $\mathsf{X}'$ to test the proposed distribution $\mathcal{D}_e$. Let

$$L(\mathcal{D}_e) = \frac{1}{m}\sum_{i=1}^{m} \mathcal{G}(\mathcal{D}_e, \bar{\mathsf{x}}'_i), \qquad \beta(\mathcal{D}_e) = B(\mathcal{D}_e)\sqrt{\frac{\log(1/\gamma)}{2m}},$$

and runs the while-loop until both $L(\mathcal{D}_e) \le r' - \beta(\mathcal{D}_e)$ and $B(\mathcal{D}_e) \le R$ hold. Let $\widehat{\mathcal{D}}_e$ denote the final noise distribution returned by the algorithm.

We shall show that $\mu(\widehat{\mathcal{D}}_e) \le r$, which is exactly

$$\mathbb{E}_{\mathsf{x}\sim D}\Big[\mathcal{G}(\widehat{\mathcal{D}}_e, \mathsf{x})\Big] = \mathbb{E}_{\mathsf{x}\sim D}\Big[\mathcal{D}_\alpha\big(P_{\mathcal{M}_{\widehat{\mathcal{D}}_e}(\mathsf{x})} \,\|\, P_{W_{\widehat{\mathcal{D}}_e}}\big)\Big] \le r.$$

*Step 1: Condition on the training sample and define the ideal parameter.* Fix the first sample set $\mathsf{X}$ and condition on it. Under this conditioning, the candidate produced by the empirical optimization (13) depends only on $\mathsf{X}$. We further assume that there exists $\sigma^\star \ge 0$ and a corresponding noise distribution $\mathcal{D}_e^\star$, given by the law of $\mathsf{e} + \sigma^\star Z$ with $\mathsf{e} \sim \mathcal{D}_e$ and $Z \sim \mathcal{N}(0, \mathbf{I})$ independent, that *saturates* the constraint, i.e.

$$\mu(\mathcal{D}_e^\star) = r'.$$

If no such $\sigma^\star$ exists (that is, $\mu(\mathcal{D}_e) < r'$ for every admissible $\mathcal{D}_e$), then in particular the final output $\widehat{\mathcal{D}}_e$ of Algorithm 1 satisfies $\mu(\widehat{\mathcal{D}}_e) \le r' \le r$, and the desired guarantee already holds. Hence, without loss of generality, we may assume that such a $\sigma^\star$ (and $\mathcal{D}_e^\star$) exists. This $\mathcal{D}_e^\star$ is a function of the fixed training sample $\mathsf{X}$, but crucially it is *independent* of the second, validation sample $\mathsf{X}'$.

*Step 2: Hoeffding bound on the independent validation sample.* Now condition on $\mathsf{X}$ and on the choice of $\mathcal{D}_e^\star$. Under this conditioning, the random variables

$$Y_i := \mathcal{G}(\mathcal{D}_e^\star, \bar{\mathsf{x}}'_i), \qquad i = 1, \ldots, m,$$

are i.i.d. and bounded: $0 \le Y_i \le B(\mathcal{D}_e^\star)$, and

$$\mathbb{E}[Y_i] = \mu(\mathcal{D}_e^\star) = r'.$$

Their empirical average is

$$L^\star := L(\mathcal{D}_e^\star) = \frac{1}{m}\sum_{i=1}^{m} Y_i.$$

By Hoeffding's inequality, for any $\beta > 0$,

$$\Pr_{\mathsf{X}'}\Big(L^\star \le r' - \beta \,\Big|\, \mathsf{X}\Big) \;\le\; \exp\Big(-\frac{2m\beta^2}{B(\mathcal{D}_e^\star)^2}\Big).$$

Set

$$\beta^\star := B(\mathcal{D}_e^\star)\sqrt{\frac{\log(1/\gamma)}{2m}},$$

then the right-hand side is exactly $\gamma$, and we obtain

$$\Pr_{\mathsf{X}'}\Big(L^\star \le r' - \beta^\star \,\Big|\, \mathsf{X}\Big) \;\le\; \gamma. \tag{38}$$

Equivalently,

$$\Pr_{\mathsf{X}'}\Big(L^\star > r' - \beta^\star \,\Big|\, \mathsf{X}\Big) \;\ge\; 1-\gamma.$$

Define the "good" event

$$\mathsf{E} := \big\{L^\star > r' - \beta^\star\big\}.$$

On $\mathsf{E}$ we have

$$L^\star > r' - \beta^\star,$$

so $L^\star$ does *not* satisfy the validation condition $L(\mathcal{D}_e) \le r' - \beta(\mathcal{D}_e)$, and hence the ideal benchmark $\mathcal{D}_e^\star$ fails the validation test on $\mathsf{E}$.

*Step 3: Monotonicity of the update step.* Whenever the validation test fails, Algorithm 1 updates the current noise distribution by adding an independent Gaussian:

$$\mathsf{e}_{\text{new}} \;=\; \mathsf{e}_{\text{old}} + z, \qquad z \sim \mathcal{N}(0, \sigma^2\mathbf{I}),$$

so that the new privatized mechanism is a post-processing of the old one. By the data-processing inequality for $f$-divergences (in particular, for $\alpha$-divergence), such a common post-processing cannot increase the privacy loss:

$$\mathcal{G}(\mathcal{D}_e^{\text{new}}, \mathsf{x}) \;\le\; \mathcal{G}(\mathcal{D}_e^{\text{old}}, \mathsf{x}), \qquad \forall \mathsf{x}.$$

Consequently, the population risk is nonincreasing along the updates:

$$\mu(\mathcal{D}_e^{\text{new}}) \;:=\; \mathbb{E}_{\mathsf{x}\sim D}[\mathcal{G}(\mathcal{D}_e^{\text{new}}, \mathsf{x})] \;\le\; \mu(\mathcal{D}_e^{\text{old}}).$$

Moreover, our global bound $B(\mathcal{D}_e)$ also decreases under this update: we have

$$B(\mathcal{D}_e^{\text{new}}) \;\le\; B(\mathcal{D}_e^{\text{old}}),$$

and hence the Hoeffding deviation term

$$\beta(\mathcal{D}_e) := B(\mathcal{D}_e)\sqrt{\frac{\log(1/\gamma)}{2m}}$$

is also nonincreasing along the updates:

$$\beta(\mathcal{D}_e^{\text{new}}) \;\le\; \beta(\mathcal{D}_e^{\text{old}}).$$

Equivalently, the validation threshold $r' - \beta(\mathcal{D}_e)$ is *nondecreasing* as the algorithm adds more noise:

$$r' - \beta(\mathcal{D}_e^{\text{new}}) \;\ge\; r' - \beta(\mathcal{D}_e^{\text{old}}).$$

Thus, for a fixed training set $\mathsf{X}$, as the algorithm iteratively updates $\mathcal{D}_e$ using only $\mathsf{X}'$, we obtain a sequence of mechanisms whose population risks $\mu(\mathcal{D}_e)$ form a nonincreasing sequence, while the validation threshold $r' - \beta(\mathcal{D}_e)$ is nondecreasing. On the event $\mathsf{E}$ from Step 2, the benchmark $\mathcal{D}_e^\star$ satisfies

$$L(\mathcal{D}_e^\star) > r' - \beta(\mathcal{D}_e^\star),$$

so it fails the validation test and the algorithm cannot stop at $\mathcal{D}_e^\star$. Every subsequent update strictly increases the noise level and hence yields a mechanism $\mathcal{D}_e^{\text{new}}$ with

$$\mu(\mathcal{D}_e^{\text{new}}) \;\le\; \mu(\mathcal{D}_e^\star) = r',$$

while the corresponding threshold $r' - \beta(\mathcal{D}_e^{\text{new}})$ is no smaller than $r' - \beta(\mathcal{D}_e^\star)$. When the algorithm eventually stops at some $\widehat{\mathcal{D}}_e$ that passes the test $L(\widehat{\mathcal{D}}_e) \le r' - \beta(\widehat{\mathcal{D}}_e)$, we necessarily have

$$\mu(\widehat{\mathcal{D}}_e) \;\le\; \mu(\mathcal{D}_e^\star) = r' \qquad \text{on the event } \mathsf{E}, \tag{39}$$

because $\widehat{\mathcal{D}}_e$ arises from $\mathcal{D}_e^\star$ by a sequence of such a post-processing updates.

*Step 4: Handling the complement event and using the fallback bound.* On the complement event $\mathsf{E}^c$, we do not control the deviation of $L(\mathcal{D}_e^\star)$ from $r'$. However, by the stopping rule of Algorithm 1, the final output $\widehat{\mathcal{D}}_e$ always satisfies

$$B(\widehat{\mathcal{D}}_e) \;\le\; R,$$

and hence, using $\mathcal{G}(\mathcal{D}_e, \mathsf{x}) \le B(\mathcal{D}_e)$ for all $\mathsf{x}$, we get

$$\mu(\widehat{\mathcal{D}}_e) = \mathbb{E}_{\mathsf{x}\sim D}\big[\mathcal{G}(\widehat{\mathcal{D}}_e, \mathsf{x})\big] \;\le\; B(\widehat{\mathcal{D}}_e) \;\le\; R \qquad \text{on } \mathsf{E}^c. \tag{40}$$

*Step 5: Final bound.* Now take expectation over both $\mathsf{X}$ and $\mathsf{X}'$. Combining (39) and (40) and using $\Pr(\mathsf{E}) \geq 1-\gamma$ and $\Pr(\mathsf{E}^c) \leq \gamma$, we obtain

$$\mathbb{E}\big[\mu(\widehat{\mathcal{D}}_e)\big] \;\leq\; (1-\gamma)\, r' + \gamma\, R \;\leq\; r,$$

where the last inequality is exactly the requirement imposed in step 3 of Algorithm 1.

Finally, by the definition of $\mu(\widehat{\mathcal{D}}_e)$, this is equivalent to

$$\mathbb{E}_{\mathsf{x}\sim D}\Big[\mathcal{G}(\widehat{\mathcal{D}}_e, \mathsf{x})\Big] = \mathbb{E}_{\mathsf{x}\sim D}\Big[\mathcal{D}_\alpha\big(P_{\mathcal{M}_{\widehat{\mathcal{D}}_e}(\mathsf{x})} \,\|\, P_{W_{\widehat{\mathcal{D}}_e}}\big)\Big] \;\leq\; r.$$

This proves the theorem.

## F Proof of Theorem 2

Apply Lemma 3, Lemma 4 and Lemma 5, we know

$$\begin{aligned}
\delta_\rho &= \int_{\mathcal{O}} \max_{\mathsf{x}\in\mathcal{X}} H(\mathsf{x},\mathsf{o})\,\mathrm{d}\mathsf{o} && (41)\\
&\leq \int_{\mathcal{O}} \left(\sum_{\mathsf{x}\in\mathcal{X}} H(\mathsf{x},\mathsf{o})^\alpha\right)^{1/\alpha} \mathrm{d}\mathsf{o} && (42)\\
&\leq \left(\int_{\mathcal{O}} \frac{\sum_{\mathsf{x}\in\mathcal{X}} H(\mathsf{x},\mathsf{o})^\alpha}{\Pr(W=\mathsf{o})^{\alpha-1}}\mathrm{d}\mathsf{o}\right)^{1/\alpha} && (43)\\
&= \mathsf{I}^W_{\alpha,\rho}\big(\mathsf{x};\mathcal{F}(\mathsf{x})\big) && (44)
\end{aligned}$$

since this hold for arbitrary $W$, the next inequality also holds:

$$\delta_\rho \;\leq\; \mathsf{I}_{\alpha,\rho}\big(\mathsf{x};\mathcal{F}(\mathsf{x})\big).$$

## G Proof of Theorem 3

Applying Lemma 5, we see that choosing

$$\Pr(W=\mathsf{o}) \;\propto\; \left(\sum_{\mathsf{x}\in\mathcal{X}} H(\mathsf{x},\mathsf{o})^\alpha\right)^{1/\alpha}$$

is exactly the condition under which Hölder's inequality becomes tight.

## H Proof of Theorem 4

$$\lim_{\alpha\to\infty} \mathsf{I}_{\alpha,\rho}\big(\mathsf{x};\mathcal{F}(\mathsf{x})\big) \tag{45}$$

$$= \int_{\mathcal{O}} \left(\sum_{\mathsf{x}\in\mathcal{X}} H(\mathsf{x},\mathsf{o})^{\alpha}\right)^{1/\alpha} \mathrm{d}\mathsf{o}, \tag{46}$$

$$= \int_{\mathcal{O}} \max_{\mathsf{x}\in\mathcal{X}} H\big(\mathcal{F}(\mathsf{x})=\mathsf{o}\big)\,\mathrm{d}\mathsf{o} \tag{47}$$

$$= \delta_\rho \tag{48}$$

## I Proof of Theorem 5

First we need to make it clear that $\delta_\rho$ is a parameter in this theorem we use another notation $\delta_\rho^\star$ to denote the optimal posterior success rate. By applying Theorem 1, we have

$$\mathbb{E}_{\mathsf{x}\sim D}\Big[\mathcal{G}(\mathcal{D}_e,\mathsf{x})\Big] \le r.$$

Use $\mathcal{G}(\mathcal{D}_e,\hat{\mathsf{x}}) := \frac{1}{\Pr(\mathsf{x}=\hat{\mathsf{x}})}\mathcal{D}_\alpha\big(P_{H_{\mathcal{D}_e}(\hat{\mathsf{x}})} \,\|\, P_{W_{\mathcal{D}_e}}\big)$ into the above equation we have

$$\mathbb{E}_{\mathsf{x}\sim D}\ \frac{1}{\Pr(\mathsf{x}=\hat{\mathsf{x}})}\mathcal{D}_\alpha\big(P_{H_{\mathcal{D}_e}(\hat{\mathsf{x}})} \,\|\, P_{W_{\mathcal{D}_e}}\big) \le r$$

Notice that

$$\mathbb{E}_{\mathsf{x}\sim D}\ \frac{1}{\Pr(\mathsf{x}=\hat{\mathsf{x}})}\mathcal{D}_\alpha\big(P_{H_{\mathcal{D}_e}(\hat{\mathsf{x}})} \,\|\, P_{W_{\mathcal{D}_e}}\big) = \mathsf{I}^W_{\alpha,\rho}(\mathsf{x};\mathcal{M})^\alpha$$

and

$$r = (\delta_\rho)^\alpha.$$

Apply Theorem 2 we know

$$\delta_\rho^\star \le \mathsf{I}^W_{\alpha,\rho}(\mathsf{x};\mathcal{M}) \le \delta_\rho$$

So the optimal poster success rate $\delta_\rho^\star$ is no more than $\delta_\rho$.

## J Proof of Theorem 6

We use $\mathcal{M}^{\mathrm{A}dv}:\mathcal{X}\to\mathcal{O}^t\times V^t$ to denote the composed mechanism including the query output from A$dv$. For example, when $T=2$

$$\begin{aligned}&\Pr(\mathcal{M}^{\mathrm{A}dv}=(\mathsf{v}_1,\mathsf{o}_1,\mathsf{v}_2,\mathsf{o}_2))\\ =&\Pr(\mathrm{A}dv_1(\tau_1)=\mathsf{v}_1)\Pr(\mathcal{M}_1(\mathsf{x},\mathsf{v}_1,\tau_1)=\mathsf{o}_1)\Pr(\mathrm{A}dv_2(\tau_2)=\mathsf{v}_2)\Pr(\mathcal{M}_2(\mathsf{x},\mathsf{v}_2,\tau_2)=\mathsf{o}_2)\end{aligned}$$

Apply Theorem 2, we have

$$\delta_\rho \le \mathsf{I}^W_{\alpha,\rho}(\mathsf{x};\mathcal{M}^{\mathrm{A}dv})$$

For the reference distribution we define for a family of distribution:

$$W_t : \mathcal{V} \times \mathcal{T}_t \to \mathcal{O}$$

and

$$\mathrm{A}dv'_t : \mathcal{T}_t \to \mathcal{O}$$

Apply Theorem 2 then we have the optimal posterior success rate $\delta_\rho$ satisfies:

$$\delta_\rho^\alpha \le \mathsf{I}^W_{\alpha,\rho}(\mathsf{x}; \mathcal{M}^{\mathrm{A}dv}) \tag{49}$$

$$= \mathbb{E}_{\mathsf{x}\sim D} \int_{\mathcal{O}^T \times \mathcal{V}^T} \Pr_{\mathcal{D}}(\mathsf{x})^{\alpha-1} \cdot \prod_{t=1}^{T} \frac{\big(\Pr\big(\mathcal{M}_t(\mathsf{x}, \mathsf{v}_t, \tau_t) = \mathsf{o}_t\big)\big)^\alpha}{\big(\Pr\big(W_t(\mathsf{v}_t, \tau_t) = \mathsf{o}_t\big)\big)^{\alpha-1}} \cdot \prod_{t=1}^{T} \frac{\big(\Pr\big(\mathrm{Adv}_t(\tau_t) = \mathsf{v}_t\big)\big)^\alpha}{\big(\Pr\big(\mathrm{Adv}'_t(\tau_t) = \mathsf{v}_t\big)\big)^{\alpha-1}} \,\mathrm{d}\mathsf{o}_{[1:T]}\,\mathrm{d}\mathsf{v}_{[1:T]} \tag{50}$$

We choose $\mathrm{A}dv'$ has the exact same distribution with $\mathrm{A}dv$, then

$$\prod_{t=1}^{T} \frac{\big(\Pr\big(\mathrm{Adv}_t(\tau_t) = \mathsf{v}_t\big)\big)^\alpha}{\big(\Pr\big(\mathrm{Adv}'_t(\tau_t) = \mathsf{v}_t\big)\big)^{\alpha-1}} = \prod_{t=1}^{T} \Pr\big(\mathrm{Adv}_t(\tau_t) = \mathsf{v}_t\big)$$

As a result

$$\delta_\rho^\alpha \le (50) = \mathbb{E}_{\mathsf{x}\sim D} \int_{\mathcal{O}^T \times \mathcal{V}^T} \Pr_{\mathcal{D}}(\mathsf{x})^{\alpha-1} \cdot \prod_{t=1}^{T} \frac{\big(\Pr\big(\mathcal{M}_t(\mathsf{x}, \mathsf{v}_t, \tau_t) = \mathsf{o}_t\big)\big)^\alpha}{\big(\Pr\big(W_t(\mathsf{v}_t, \tau_t) = \mathsf{o}_t\big)\big)^{\alpha-1}} \tag{51}$$

$$\cdot \prod_{t=1}^{T} \Pr\big(\mathrm{Adv}_t(\tau_t) = \mathsf{v}_t\big)\,\mathrm{d}\mathsf{o}_{[1:T]}\,\mathrm{d}\mathsf{v}_{[1:T]} \tag{52}$$

## K Proof of Theorem 7

Apply Theorem 5, we have

$$(\delta_\rho)^\alpha \le \mathbb{E}_{\mathsf{x}\sim D} \int_{\mathcal{O}^T \times \mathcal{V}^T} \Pr_{\mathcal{D}}(\mathsf{x})^{\alpha-1} \cdot \prod_{t=1}^{T} \frac{\big(\Pr\big(\mathcal{M}_t(\mathsf{x}, \mathsf{v}_t, \tau_t) = \mathsf{o}_t\big)\big)^\alpha}{\big(\Pr\big(W_t(\mathsf{v}_t, \tau_t) = \mathsf{o}_t\big)\big)^{\alpha-1}} \cdot \prod_{t=1}^{T} \Pr\big(\mathrm{Adv}_t(\tau_t) = \mathsf{v}_t\big)\,\mathrm{d}\mathsf{o}_{[1:T]}\,\mathrm{d}\mathsf{v}_{[1:T]} \tag{53}$$

For simplicity but without loss of generality, suppose $T = 2$, notice that inequality (53) can be rewritten as

$$(\delta_\rho)^\alpha \leq \mathbb{E}_{\mathsf{x}\sim D} \int_{\mathcal{O}\times\mathcal{V}} \Pr\big(\mathrm{Adv}_1(\tau_1) = \mathsf{v}_1\big) \Pr\big(W_1(\mathsf{v}_1,\tau_1) = \mathsf{o}_1\big) \int_{\mathcal{V}} \Pr\big(\mathrm{Adv}_1(\tau_2) = \mathsf{v}_2\big)$$
$$\cdot \Pr_{\mathcal{D}}(\mathsf{x})^{\alpha-1} \frac{\big(\Pr\big(\mathcal{M}_1(\mathsf{x},\mathsf{v}_1,\tau_1) = \mathsf{o}_1\big)\big)^\alpha}{\big(\Pr\big(W_1(\mathsf{v}_1,\tau_1) = \mathsf{o}_1\big)\big)^\alpha} \int_{\mathcal{O}} \frac{\big(\Pr\big(\mathcal{M}_t(\mathsf{x},\mathsf{v}_t,\tau_t) = \mathsf{o}_t\big)\big)^\alpha}{\big(\Pr\big(W_t(\mathsf{v}_t,\tau_t) = \mathsf{o}_t\big)\big)^{\alpha-1}} \mathrm{d}\mathsf{o}_2 \mathrm{d}\mathsf{v}_2 \mathrm{d}\mathsf{o}_1 \mathrm{d}\mathsf{v}_1 \tag{54}$$

$$= \int_{\mathcal{O}\times\mathcal{V}} \Pr\big(\mathrm{Adv}_1(\tau_1) = \mathsf{v}_1\big) \Pr\big(W_1(\mathsf{v}_1,\tau_1) = \mathsf{o}_1\big) \int_{\mathcal{V}} \Pr\big(\mathrm{Adv}_1(\tau_2) = \mathsf{v}_2\big)$$
$$\cdot \mathbb{E}_{\mathsf{x}\sim D} \underbrace{\Pr_{\mathcal{D}}(\mathsf{x})^{\alpha-1} \frac{\big(\Pr\big(\mathcal{M}_1(\mathsf{x},\mathsf{v}_1,\tau_1) = \mathsf{o}_1\big)\big)^\alpha}{\big(\Pr\big(W_1(\mathsf{v}_1,\tau_1) = \mathsf{o}_1\big)\big)^\alpha}}_{(A)} \underbrace{\int_{\mathcal{O}} \frac{\big(\Pr\big(\mathcal{M}_2(\mathsf{x},\mathsf{v}_2,\tau_2) = \mathsf{o}_t\big)\big)^\alpha}{\big(\Pr\big(W_t(\mathsf{v}_2,\tau_2) = \mathsf{o}_2\big)\big)^{\alpha-1}} \mathrm{d}\mathsf{o}_2}_{(B)} \mathrm{d}\mathsf{v}_2 \mathrm{d}\mathsf{o}_1 \mathrm{d}\mathsf{v}_1 \tag{55}$$

Notice that term (A) is $A_1(\mathsf{x})$ and term (B) is $D_2(\mathsf{x})$, the condition in the theorem, tells us that

$$\mathbb{E}_{\mathsf{x}\sim D}\big[A_1(\mathsf{x})\, D_2(\mathsf{x})\big] \;\leq\; r_2 \cdot \mathbb{E}_{\mathsf{x}\sim D}\big[A_1(\mathsf{x})\big], \tag{56}$$

so (55)$\leq$

$$\int_{\mathcal{O}\times\mathcal{V}} \Pr\big(\mathrm{Adv}_1(\tau_1) = \mathsf{v}_1\big) \Pr\big(W_1(\mathsf{v}_1,\tau_1) = \mathsf{o}_1\big) \int_{\mathcal{V}} \Pr\big(\mathrm{Adv}_1(\tau_2) = \mathsf{v}_2\big)$$
$$\cdot r_2 \cdot \mathbb{E}_{\mathsf{x}\sim D} \underbrace{\Pr_{\mathcal{D}}(\mathsf{x})^{\alpha-1} \frac{\big(\Pr\big(\mathcal{M}_1(\mathsf{x},\mathsf{v}_1,\tau_1) = \mathsf{o}_1\big)\big)^\alpha}{\big(\Pr\big(W_1(\mathsf{v}_1,\tau_1) = \mathsf{o}_1\big)\big)^\alpha}}_{(A)} \mathrm{d}\mathsf{v}_2 \mathrm{d}\mathsf{o}_1 \mathrm{d}\mathsf{v}_1 \tag{57}$$

$$= r_2 \cdot \int_{\mathcal{V}} \Pr\big(\mathrm{Adv}_1(\tau_1) = \mathsf{v}_1\big)$$
$$\underbrace{\int_{\mathcal{O}} \Pr\big(W_1(\mathsf{v}_1,\tau_1) = \mathsf{o}_1\big) \cdot \mathbb{E}_{\mathsf{x}\sim D} \Pr_{\mathcal{D}}(\mathsf{x})^{\alpha-1} \frac{\big(\Pr\big(\mathcal{M}_1(\mathsf{x},\mathsf{v}_1,\tau_1) = \mathsf{o}_1\big)\big)^\alpha}{\big(\Pr\big(W_1(\mathsf{v}_1,\tau_1) = \mathsf{o}_1\big)\big)^\alpha} \mathrm{d}\mathsf{o}_1}_{(D)}$$
$$\underbrace{\int_{\mathcal{V}} \Pr\big(\mathrm{Adv}_1(\tau_2) = \mathsf{v}_2\big) \mathrm{d}\mathsf{v}_2}_{(C)} \mathrm{d}\mathsf{v}_1 \tag{58}$$

$$\leq r_2 \cdot \int_{\mathcal{V}} \Pr\big(\mathrm{Adv}_1(\tau_1) = \mathsf{v}_1\big) \cdot r_1 \cdot 1 = r_1 r_2 \tag{59}$$

Equality (58) follows by straightforward algebraic rearrangement. Inequality (59) is because term (D) is smaller than $r_1$ and term C equals to 1.

## L Proof of Theorem 8

We focus on this ratio

$$\left(\prod_{t=1}^{T}\frac{\Pr\big(\mathcal{M}_t(\mathsf{x},\mathsf{v}_t,\tau_t)=\mathsf{o}_t\big)}{\Pr\big(W_t(\mathsf{v}_t,\tau_t)=\mathsf{o}_t\big)}\right)^{\alpha} \tag{60}$$

From the equal condition of Holder's inequality as is stated in Theorem 3, it's sufficient and necessary that ratio (60) is a constant for any selection of $\mathsf{o}_{[1:T]},\mathsf{v}_{[1:T]}$. Given our selection of $W$, we have

$$\mathbb{E}_{\mathsf{x}\sim D}\left[\left(\prod_{t=1}^{T}\frac{\Pr\big(\mathcal{M}_t(\mathsf{x},\mathsf{v}_t,\tau_t)=\mathsf{o}_t\big)}{\Pr\big(W_t(\mathsf{v}_t,\tau_t)=\mathsf{o}_t\big)}\right)^{\alpha}\right] \tag{61}$$

$$= r_T\cdot\mathbb{E}_{\mathsf{x}\sim D}[A_{T-1}(\mathsf{x})] \tag{62}$$

$$\tag{63}$$

Use this proposition iteratively

$$\mathbb{E}_{\mathsf{x}\sim D}\left[\left(\prod_{t=1}^{T}\frac{\Pr\big(\mathcal{M}_t(\mathsf{x},\mathsf{v}_t,\tau_t)=\mathsf{o}_t\big)}{\Pr\big(W_t(\mathsf{v}_t,\tau_t)=\mathsf{o}_t\big)}\right)^{\alpha}\right] \tag{64}$$

$$= \mathbb{E}_{\mathsf{x}\sim D}\left[A_{T-1}(\mathsf{x})\left(\frac{\Pr\big(\mathcal{M}_T(\mathsf{x},\mathsf{v}_T,\tau_T)=\mathsf{o}_T\big)}{\Pr\big(W_T(\mathsf{v}_T,\tau_T)=\mathsf{o}_T\big)}\right)^{\alpha}\right] \tag{65}$$

$$= r_T\cdot\mathbb{E}_{\mathsf{x}\sim D}[A_{T-1}(\mathsf{x})] \tag{66}$$

$$= r_T\, r_{T-1}\cdot\mathbb{E}_{\mathsf{x}\sim D}[A_{T-2}(\mathsf{x})] \tag{67}$$

$$\vdots \tag{68}$$

$$= \left(\prod_{t=1}^{T} r_t\right)\cdot\mathbb{E}_{\mathsf{x}\sim D}[A_0(\mathsf{x})] \tag{69}$$

$$= \prod_{t=1}^{T} r_t, \tag{70}$$

So ratio (60) is a constant for any selection of $\mathsf{o}_{[1:T]},\mathsf{v}_{[1:T]}$. We emphasize again this condition is necessary and sufficient for the optimal choice of $W$. Hence, selecting the locally optimal $W_t$ at each round yieldzs a globally optimal reference $W$.

## M Proof of Theorem 9

By applying Theorem 1, we have

$$\mathbb{E}_{\mathsf{x}\sim D}\Big[\mathcal{G}(\mathcal{D}_e,\mathsf{x})\Big]\ \leq\ 0.$$

Use $\mathcal{G}_t(\mathcal{D}_e, \hat{\mathsf{x}}) := (A_{t-1}(\hat{\mathsf{x}})\,(D_t(\hat{\mathsf{x}}) - r_t)$ into the above equation for every t and every selection of we have

$$\mathbb{E}_{\mathsf{x}\sim D}\big[A_{t-1}(\mathsf{x})\,(D_t(\mathsf{x}) - r_t)\big] \;\le\; 0 \tag{71}$$

which is equals to

$$\mathbb{E}_{\mathsf{x}\sim D}\big[A_{t-1}(\mathsf{x})\,D_t(\mathsf{x})\big] \;\le\; r_t \cdot \mathbb{E}_{\mathsf{x}\sim D}\big[A_{t-1}(\mathsf{x})\big], \tag{72}$$

Consequently the condition 21 is held so we apply Theorem 7 we have:

$$\delta_\rho^\alpha \le \prod_{t=1}^{T} r_t. \tag{73}$$

## N Proof of Theorem 10

By Lemma 2, to bound $D_{\mathrm{KL}}(\mathbf{1}_\rho \| \mathbf{1}_{o,\rho})$, it suffices to consider the entire interactive mechanism $\mathcal{M}$, which merges both the adversary's strategy Adv and the user's implementations across $T$ time slots, and

$$D_f(\mathbf{1}_\rho \| \mathbf{1}_{o,\rho}) \;\le\; \mathbb{E}_{\mathsf{x}\sim D} \inf_{P_W} D_f\big(P_{\mathcal{M}(\mathsf{x})} \,\|\, P_W\big)\,. \tag{74}$$

For simplicity but without loss of generality, suppose $T = 2$ and select $f(z) = z\log z$. Since the choice of the reference distribution $P_W$ can be arbitrary, set $W = (W_1, W_2)$ such that the marginal distribution $P_{W_1}$ equals $P_{W^{(1)}}$ and $W_2$ conditional on $W_1 = \mathsf{o}_1$ is the mixture of $\{P_{W_2(\mathsf{v})}\}_{\mathsf{v}\in\mathcal{V}}$, i.e.,

$$\Pr(W_2 = \mathsf{o}_2 \mid W_1 = \mathsf{o}_1) = \sum_{\mathsf{v}\in\mathcal{V}} \Pr(\mathrm{Adv}(\mathsf{o}_1) = \mathsf{v}) \Pr(W_2(\mathsf{v}) = \mathsf{o}_2). \tag{75}$$

Then under this choice we have

$$\mathbb{E}_{\mathsf{x}\sim D}\; D_f\big(P_{\mathcal{M}(\mathsf{x})} \,\|\, P_W\big) \tag{76}$$

$$\begin{aligned}
&= \int_{\mathsf{o}_1\in\mathcal{O}^{(1)}} \sum_{\mathsf{v}\in\mathcal{V}} \Pr(\mathrm{Adv}(\mathsf{o}_1) = \mathsf{v}) \int_{\mathsf{o}_2\in\mathcal{O}^{(2)}} \Pr(\mathcal{M}_1(\mathsf{x}) = \mathsf{o}_1) \Pr(\mathcal{M}_2(\mathsf{x},\mathsf{v}) = \mathsf{o}_2) \\
&\qquad \cdot \log \frac{\Pr(\mathcal{M}_1(\mathsf{x}) = \mathsf{o}_1) \Pr(\mathcal{M}_2(\mathsf{x},\mathsf{v}) = \mathsf{o}_2)}{\Pr(W_1 = \mathsf{o}_1) \Pr(W_2(\mathsf{v}) = \mathsf{o}_2)} \\
&= \underbrace{\int_{\mathsf{o}_1\in\mathcal{O}^{(1)}} \Pr(\mathcal{M}_1(\mathsf{x}) = \mathsf{o}_1) \log \frac{\Pr(\mathcal{M}_1(\mathsf{x}) = \mathsf{o}_1)}{\Pr(W_1 = \mathsf{o}_1)}}_{A} \cdot \underbrace{\sum_{\mathsf{v}\in\mathcal{V}} \Pr(\mathrm{Adv}(\mathsf{o}_1) = \mathsf{v}) \int_{\mathsf{o}_2\in\mathcal{O}^{(2)}} \Pr(\mathcal{M}_2(\mathsf{x},\mathsf{v}) = \mathsf{o}_2)}_{B} \\
&\qquad + \underbrace{\sum_{\mathsf{v}\in\mathcal{V}} \int_{\mathsf{o}_1\in\mathcal{O}^{(1)}} Q(\mathsf{x},\mathsf{o}_1,\mathsf{v}) \int_{\mathsf{o}_2\in\mathcal{O}^{(2)}} \Pr(\mathcal{M}_2(\mathsf{x},\mathsf{v}) = \mathsf{o}_2) \log \frac{\Pr(\mathcal{M}_2(\mathsf{x},\mathsf{v}) = \mathsf{o}_2)}{\Pr(W_2(\mathsf{v}) = \mathsf{o}_2)}}_{C}.
\end{aligned} \tag{77}$$

where

$$Q(\mathsf{x}, \mathsf{o}_1, \mathsf{v}) = \Pr(\mathcal{M}_1(\mathsf{x}) = \mathsf{o}_1) \Pr(\mathrm{Adv}(\mathsf{o}_1) = \mathsf{v}). \tag{78}$$

Term $B$ equals 1. Define

$$D^{(1)}(\mathsf{x}) = \int_{\mathsf{o}_1 \in \mathcal{O}^{(1)}} \Pr(\mathcal{M}_1(\mathsf{x}) = \mathsf{o}_1) \log \frac{\Pr(\mathcal{M}_1(\mathsf{x}) = \mathsf{o}_1)}{\Pr(W_1 = \mathsf{o}_1)}, \tag{79}$$

$$D^{(2)}_{\mathsf{v}}(\mathsf{x}) = \int_{\mathsf{o}_2 \in \mathcal{O}^{(2)}} \Pr(\mathcal{M}_2(\mathsf{x}, \mathsf{v}) = \mathsf{o}_2) \log \frac{\Pr(\mathcal{M}_2(\mathsf{x}, \mathsf{v}) = \mathsf{o}_2)}{\Pr(W_2(\mathsf{v}) = \mathsf{o}_2)}. \tag{80}$$

From (77),

$$\mathbb{E}_{\mathsf{x} \sim D} D_{\mathrm{KL}}(P_{\mathcal{M}(\mathsf{x})} \| P_W) \le \mathbb{E}_{\mathsf{x} \sim D} \left[ D^{(1)}(\mathsf{x}) + \sum_{\mathsf{v} \in \mathcal{V}} \Pr(\mathsf{v} \mid \mathsf{x})\, D^{(2)}_{\mathsf{v}}(\mathsf{x}) \right]. \tag{81}$$

We now bound $\mathbb{E}_{\mathsf{x} \sim D} |\Pr(\mathsf{v} \mid \mathsf{x}) - \Pr(\mathsf{v})|$. Choose

$$P_{W_1} = \sum_{\mathsf{x}'} \Pr(\mathsf{x} = \mathsf{x}')\, P_{\mathcal{M}_1(\mathsf{x}')}. \tag{82}$$

Under the assumption $\mathbb{E}_{\mathsf{x}} \mathrm{TV}(P_{\mathcal{M}_1(\mathsf{x})} \| P_{W_1}) \le r_1$, Pinsker's inequality gives

$$\mathbb{E}_{\mathsf{x}} \mathrm{TV}(P_{\mathcal{M}_1(\mathsf{x})} \| P_{W_1}) \le \sqrt{2 r_1}. \tag{83}$$

Furthermore,

$$\sum_{\mathsf{v} \in \mathcal{V}} \mathbb{E}_{\mathsf{x}} \big| \Pr(\mathsf{v} \mid \mathsf{x}) - \Pr(\mathsf{v}) \big| = \mathbb{E}_{\mathsf{x}} \mathrm{TV}(P_{\mathcal{M}_1(\mathsf{x})} \| P_{W_1}) \le \sqrt{2 r_1}. \tag{84}$$

Since each $D^{(2)}_{\mathsf{v}}(\mathsf{x}) \le R_2$, the second term in (81) is bounded by

$$r_2 + R_2 \sqrt{2 r_1}. \tag{85}$$

Thus,

$$\mathbb{E}_{\mathsf{x} \sim D} D_{\mathrm{KL}}(P_{\mathcal{M}(\mathsf{x})} \| P_W) \le r_1 + r_2 + R_2 \sqrt{2 r_1}. \tag{86}$$

By induction across all $T$ rounds, we recover the general upper bound $B(t)$ shown in (4.8).

## O Proof of Theorem 11

Apply Theorem 6, to bound the optimal posterior success rate $\delta_\rho$, we only need to bound

$$\mathbb{E}_{\mathsf{x}\sim D}\int_{\mathcal{O}^T\times\mathcal{V}^T}\Pr_{\mathcal{D}}(\mathsf{x})^{\alpha-1}\cdot\prod_{t=1}^{T}\frac{\big(\Pr\big(\mathcal{M}_t(\mathsf{x},\mathsf{v}_t,\tau_t)=\mathsf{o}_t\big)\big)^{\alpha}}{\big(\Pr\big(W_t(\mathsf{v}_t,\tau_t)=\mathsf{o}_t\big)\big)^{\alpha-1}}\cdot\prod_{t=1}^{T}\Pr\big(\mathrm{Adv}_t(\tau_t)=\mathsf{v}_t\big)\,\mathrm{d}\mathsf{o}_{[1:T]}\,\mathrm{d}\mathsf{v}_{[1:T]}$$

(87)

$$=\sum_{\mathsf{x}\in\mathcal{X}}\int_{\mathcal{O}^T\times\mathcal{V}^T}\Pr_{\mathcal{D}}(\mathsf{x})^{\alpha}\cdot\prod_{t=1}^{T}\Pr\big(W_t(\mathsf{v}_t,\tau_t)=\mathsf{o}_t\big)\prod_{t=1}^{T}\Pr\big(\mathrm{Adv}_t(\tau_t)=\mathsf{v}_t\big)\cdot \tag{88}$$

$$\prod_{t=1}^{T}\frac{\big(\Pr\big(\mathcal{M}_t(\mathsf{x},\mathsf{v}_t,\tau_t)=\mathsf{o}_t\big)\big)^{\alpha}}{\big(\Pr\big(W_t(\mathsf{v}_t,\tau_t)=\mathsf{o}_t\big)\big)^{\alpha}}\,\mathrm{d}\mathsf{o}_{[1:T]}\,\mathrm{d}\mathsf{v}_{[1:T]} \tag{89}$$

$$\leq\prod_{t=1}^{T}\left(\begin{array}{l}\displaystyle\sum_{\mathsf{x}\in\mathcal{X}}\int_{\mathcal{O}^T\times\mathcal{V}^T}\Pr_{\mathcal{D}}(\mathsf{x})^{\alpha}\cdot\prod_{t=1}^{T}\Pr\big(W_t(\mathsf{v}_t,\tau_t)=\mathsf{o}_t\big)\prod_{t=1}^{T}\Pr\big(\mathrm{Adv}_t(\tau_t)=\mathsf{v}_t\big)\\ \qquad\cdot\dfrac{\big(\Pr\big(\mathcal{M}_t(\mathsf{x},\mathsf{v}_t,\tau_t)=\mathsf{o}_t\big)\big)^{p_t\alpha}}{\big(\Pr\big(W_t(\mathsf{v}_t,\tau_t)=\mathsf{o}_t\big)\big)^{p_t\alpha}}\,\mathrm{d}\mathsf{o}_{[1:T]}\,\mathrm{d}\mathsf{v}_{[1:T]}\end{array}\right)^{\frac{1}{p_t}} \tag{90}$$

The inequality is from Holder's inequality. So we have

$$\delta_\rho^\alpha\ \leq\ \prod_{t=1}^{T}r_t^{\frac{1}{p_t}}. \tag{91}$$

## P Experiment

We run two groups of experiments.

### P.1 AES

*Leakage model.* We consider a single-round leakage experiment based on the AES S-box with Hamming-weight leakage. The secret $\mathsf{x}$ is a binary string drawn from either a 256-bit keyspace or a 128-bit keyspace, always under the uniform prior:

$$\mathsf{x}\sim\mathrm{Unif}\big(\{0,1\}^{256}\big)\quad\text{or}\quad\mathsf{x}\sim\mathrm{Unif}\big(\{0,1\}^{128}\big).$$

We write $\mathsf{x}=(\mathsf{x}^{(1)},\ldots,\mathsf{x}^{(n)})$ as a sequence of bytes, where $n=32$ in the 256-bit case and $n=16$ in the 128-bit case.

Let $S(\cdot)$ denote the AES S-box and $\mathrm{HW}(\cdot)$ the Hamming weight of an 8-bit value. For a given secret $X$, we model the one-round "raw" leakage score as an aggregate of per-byte S-box Hamming weights.

*Inference criteria $\rho$.* Given the noisy observation $Y$, the attacker outputs a guess $\hat{\mathsf{x}}$ for the secret. We evaluate two Identification criteria:

- *Identification:* the attack is counted as successful if $\hat{\mathsf{x}} = \mathsf{x}$.
- *Miss bits less than 16 bit:* the attack is counted as successful if the Hamming distance between the guess and the true secret is at most 16 bits.

Let $m(X)$ be the deterministic function of the secret. For each choice of noise variance $\sigma^2$, we fix the Gaussian mechanism $Y = m(X) + e$ with $e \sim \mathcal{N}(0, \sigma^2)$ and evaluate the identification criterion $\rho$ on the resulting channel. We first compute the *ground-truth* optimal failure level $\delta_\rho$ by explicitly evaluating the posterior over the secret space and taking the best possible adversarial decision rule. We then compare $\delta_\rho$ with three information-theoretic upper bounds on the same quantity: (i) a mutual-information-based bound obtained from Fano's inequality; (ii) a PAC-privacy bound with $\alpha = 20$ using a Gaussian reference distribution $W_{\mathrm{Gauss}}$; and (iii) our prior–reference-weighted $\alpha$-information bound also with $\alpha = 20$ instantiated with the optimal reference distribution $W^\star$. This setup allows us to assess how tightly each analytic bound approximates the true posterior success rate.

In Figures 2–4, the $x$-axis is the standard deviation $\sigma$ of the Gaussian noise, and the $y$-axis is $\log_2$ of the corresponding upper bound on the posterior success rate, computed under each of the three methods above.

Two qualitative trends are evident:

1. The mutual-information-based bound via Fano is the loosest, while the PAC-privacy bound and our prior–reference-weighted $\alpha$-information bound are substantially tighter. The gap is exponential in the success probability, i.e., linear in the $\log_2$-scale of the plots.
2. As the entropy of the secret increases, the gap between the mutual-information bound and the $\alpha$-divergence-based bounds widens. In this regime, the $\alpha$-divergence approach provides a much stronger guarantee than the mutual-information bound.

### P.2 RSA

We call Algorithm 2 to instantiate our mechanism to mitigate leakage from the running time in RSA encryption. In this experiment, the attacker observes variations in runtime caused by modular exponentiation and attempts to reconstruct the secret key from the leakage. Following our implementation, the timing oracle is simulated by measuring the end-to-end runtime of an RSA pipeline: deterministic key generation from a secret seed plus encryptions the messages given by the adversary.

The secret $\mathsf{x}$ is a binary string drawn uniformly at random from $\{0,1\}^n$, where the key size $n$ varies from 62 to 64 bits. For each round $t$, we report the average accumulated noise variance in the form

$$\frac{1}{t}\sum_{i=1}^{t}\sigma_i^2.$$

The inference criterion $\rho$ is identification: the attack is successful if and only if $\hat{\mathsf{x}} = \mathsf{x}$. We fix the target posterior success rate to $\delta_\rho = 2^{-60}$ and compute the required noise level in our algorithm. We evaluate our composition theorem for $\alpha \in \{50, 75, 100\}$ and compute the per-round average noise variance under each setting. All reported noise values are averaged over 100 independent trials and the composition number $T$ is set to 8, as shown in Fig. 6, Fig. 7, and Fig. 8.

We have two key observations:

1. The required noise grows approximately linearly (after averaging across rounds). This contrasts with Theorem 10, where the required noise grows quadratically. This difference suggests that allowing the mechanism to adaptively calibrate noise based on previous outputs and adversarial queries can significantly reduce overall noise accumulation.
2. As $\alpha$ increases, the marginal reduction in required noise becomes progressively smaller. This indicates that the noise level converges to a minimal value as $\alpha$ grows. Empirically, this supports that the bound becomes increasingly tight as $\alpha \to \infty$.

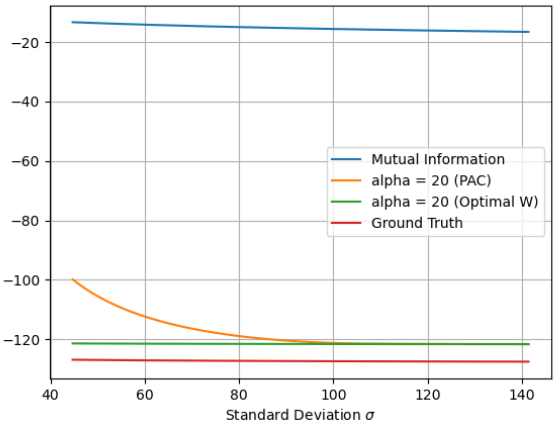


**Fig. 2.** $\log_2$ Posterior Success Bound vs. standard deviation $\sigma$ (128-bit Secret, Identical Inference)

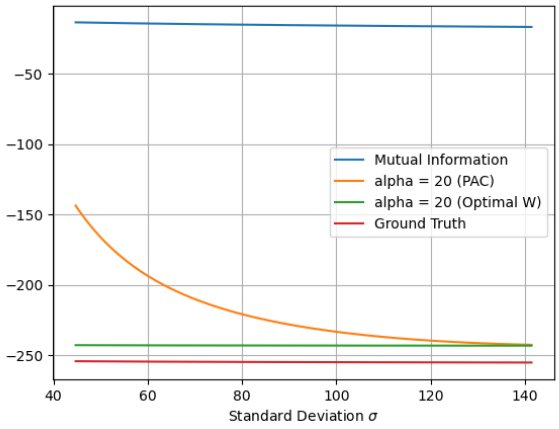


**Fig. 3.** $\log_2$ Posterior Success Bound vs. standard deviation $\sigma$ (256-bit Secret, Identical Inference)

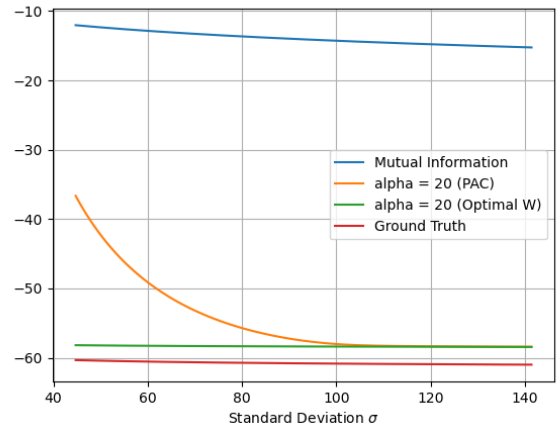


**Fig. 4.** $\log_2$ Posterior Success Bound vs. standard deviation $\sigma$ (128-bit Secret, $\leq$ 16-Bit Error)

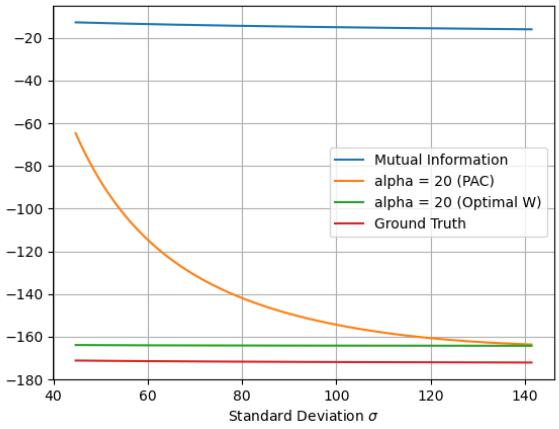


**Fig. 5.** $\log_2$ Posterior Success Bound vs. standard deviation $\sigma$ (256-bit Secret, $\leq$ 16-Bit Error)

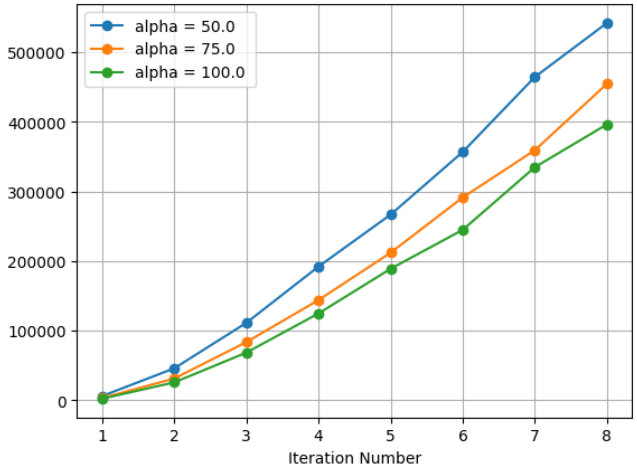


**Fig. 6.** Per-Round Average Noise vs. Iteration Number (64-bit Secret, Target Posterior Success $\leq 2^{-60}$)

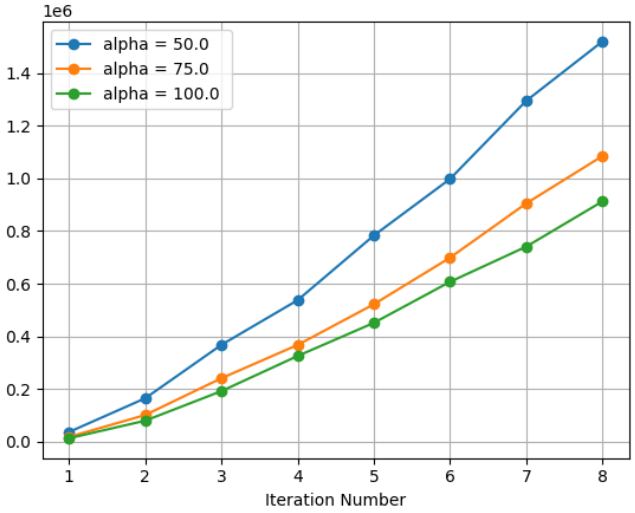


**Fig. 7.** Per-Round Average Noise vs. Iteration Number (63-bit Secret, Target Posterior Success $\leq 2^{-60}$)

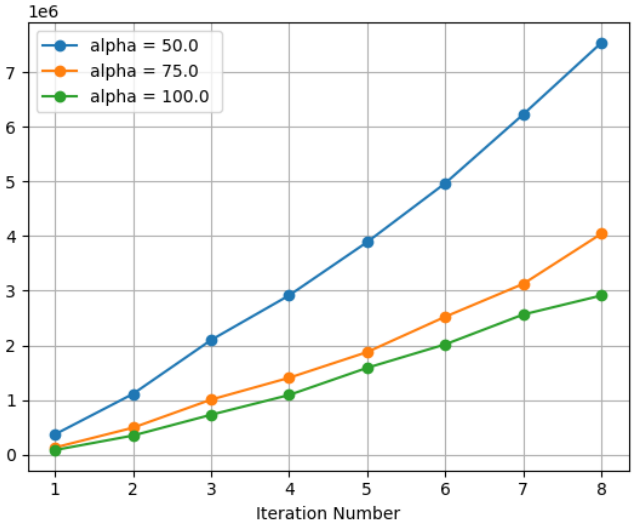


**Fig. 8.** Per-Round Average Noise vs. Iteration Number (62-bit Secret, Target Posterior Success $\leq 2^{-60}$)

## Q Analogous Results based on KL Divergence

For the KL divergence, we apply Lemma 2 with $f(u) = u \log u$ to bound $\delta_\rho$. We use the similar technique: consider further decomposing the global control problem into a collection of *local* subproblems, each of which depends only on the current branch and can be solved on the local branch.

**Theorem 12 (Adaptive Composition of KL-divergence).**

*Under the setting of Definition 8 with a single secret* $\mathsf{x} \sim \mathcal{D}$ *reused across all* $T$ *rounds. For any possible query sequence, consider an* auxiliary reference process

$$W_{1:T}(\mathsf{v}_{1:T}, \tau_{1:T}) := (W_1(\mathsf{v}_1, \tau_1), \ldots, W_T(\mathsf{v}_T, \tau_T)) \in \mathcal{O}^T$$

*which is independent of* $\mathsf{x}$.

*Define the* prefix coefficient

$$A_{t-1}(\hat{\mathsf{x}}) := \prod_{s=1}^{t-1} \frac{\Pr\big(\mathcal{M}_s(\hat{\mathsf{x}}, \mathsf{v}_s, \tau_s) = \mathsf{o}_s\big)}{\Pr\big(W_s(\mathsf{v}_s, \tau_s) = \mathsf{o}_s\big)}$$

*and*

$$D_t(\hat{\mathsf{x}}) := \mathcal{D}_{KL}(\mathcal{M}_t(\hat{\mathsf{x}}, \mathsf{v}_t, \tau_t) \,\|\, W_t(\mathsf{v}_t, \tau_t)).$$

*If for every possible transcript* $\tau_t$ *there exists such a choice of reference process* $W_{1:T}(\mathsf{v}_{1:T})$ *such that for all* $t = 1, \ldots, T$,

$$\mathbb{E}_{\mathsf{x}\sim D}\big[A_{t-1}(\mathsf{x})\, D_t(\mathsf{x})\big] \;\le\; r_t \tag{92}$$

*then the final interactive mechanism* $\mathcal{M}^{\mathrm{A}dv}$ *satisfies for any* A$dv$*, the optimal posterior success rate* $\delta_\rho$ *satisfies*

$$D_{KL}\big(\mathbf{1}_{\delta_\rho} \,\big\|\, \mathbf{1}_{\delta_{o,\rho}}\big) \;\le\; \sum_{t=1}^{T} r_t. \tag{93}$$

The KL condition above admits an intuitive weighted-sum interpretation. For each round $t$, the integrand is the conditional KL divergence between the round-$t$ output law $\Pr(\mathcal{M}_t(\mathsf{x}, \mathsf{v}_t, \tau_t) = \cdot)$ and the corresponding reference conditional law $\Pr(W_t(\mathsf{v}_t, \tau_t) = \cdot)$. This conditional KL term is multiplied by the prefix likelihood ratio $A_{t-1}(\mathsf{x})$ which can be viewed as an importance weight induced by the past transcript.

Just like the case for $\alpha$-divergence, implementing Theorem 12 in practice requires specifying a concrete reference process $W_{1:T}(v_{1:T})$. For KL, the best choice would be the marginal output law induced by the mechanism, since it yields the tightest divergence bound. To be more specific, we have the following theorem:

**Theorem 13 (Optimal selection of reference for KL divergence).** *For any* $\mathsf{x} \sim D$ *and mechanism* $\mathcal{M}(\mathsf{x})$*, let* $W$ *denote the marginal distribution of* $M(\mathsf{x})$ *over* $\mathsf{x} \sim D$*, then for arbitrary distribution* $Q$*, we have*

$$\mathbb{E}_{\mathsf{x}\sim D}\mathcal{D}_{KL}(P_{\mathcal{M}(\mathsf{x})} \| P_W) \le \mathbb{E}_{\mathsf{x}\sim D}\mathcal{D}_{KL}(P_{\mathcal{M}(\mathsf{x})} \| P_Q) \tag{94}$$

So we know the optimal reference distribution $W$ for KL divergence is the marginal distribution of $\mathcal{M}(\mathsf{x})$.

Finally, to further make Theorem 12 implementable from finite sampling, we establish Algorithm 3 (similar to Algorithm 2) by iteratively running Algorithm 1.

We provide the provable guarantees of Algorithm 2 in Theorem 9.

**Theorem 14 (Provable guarantees of Algorithm 3).** *Define the* prefix coefficient

$$A_{t-1}(\hat{\mathsf{x}}) := \prod_{s=1}^{t-1} \frac{\Pr\big(\mathcal{M}_s(\hat{\mathsf{x}}, \mathsf{v}_s, \tau_s) = \mathsf{o}_s\big)}{\Pr\big(W_s(\mathsf{v}_s, \tau_s) = \mathsf{o}_s\big)}$$

$$D_t(\hat{\mathsf{x}}) := \mathcal{D}_{KL}(\mathcal{M}_t(\hat{\mathsf{x}}, \mathsf{v}_t, \tau_t) \,\|\, W_t(\mathsf{v}_t, \tau_t)),$$

*and the privacy-loss function*

$$\mathcal{G}_t(\mathcal{D}_e, \hat{\mathsf{x}}) \;:= A_{t-1}(\hat{\mathsf{x}})\, D_t(\hat{\mathsf{x}}) \tag{95}$$

*Suppose that for every $\hat{\mathsf{x}} \in \mathcal{X}$ and every admissible $\mathcal{D}_e$ we have $\mathcal{G}_t(\mathcal{D}_e, \hat{\mathsf{x}}) \le B(\mathcal{D}_e)$ for some finite bound $B(\mathcal{D}_e)$ given per-round budget $r_t$, Let $\mathcal{M}_{[1:T]}$ be the interactive mechanism produced by Algorithm 3. Then for any $\mathcal{A}dv$, the optimal posterior success rate $\delta_\rho$ satisfies*

$$D_{KL}\big(\mathbf{1}_{\delta_\rho} \,\big\|\, \mathbf{1}_{\delta_{o,\rho}}\big) \;\le\; \sum_{t=1}^{T} r_t. \tag{96}$$

This result is similar to Theorem 9. The difference here is the one we select $\alpha$ divergence to control while for the other we select Kl divergence to control. Since when *alpha* $\to 1$, *alpha* divergence will converges to KL, Theorem 14 can viewed as a special case of Theorem 9. Notice when $\alpha \to 1$ the bound is tight so this intuitively suggests that Kl divergence is a very loose control since "1" is much smaller than "$\infty$".

## R Proof of Theorem 12

Apply Lemma 2, and we have

$$D_{KL}\big(\mathbf{1}_{\delta_\rho} \,\big\|\, \mathbf{1}_{\delta_{o,\rho}}\big) \le \mathbb{E}_{\mathsf{x}\sim D} \int_{\mathcal{O}^T \times \mathcal{V}^T} \cdot \prod_{t=1}^{T} \Pr\big(\mathrm{Adv}_t(\tau_t) = \mathsf{v}_t\big) \tag{97}$$

$$\big(\Pr\mathcal{M}_t(\mathsf{x}, \mathsf{v}_t, \tau_t) = \mathsf{o}_t\big) \cdot \prod_{t=1}^{T} \log\big(\frac{\Pr\mathcal{M}_t(\mathsf{x}, \mathsf{v}_t, \tau_t) = \mathsf{o}_t}{\Pr W_t(\mathsf{v}_t, \tau_t) = \mathsf{o}_t}\big) \mathrm{d}\mathsf{o}_{[1:T]}\, \mathrm{d}\mathsf{v}_{[1:T]} \tag{98}$$

**Algorithm 3** Randomization Optimizer under Adversarial Composition for KL divergence

1: **Input:** number of rounds $T$; leakage functions $\{\mathcal{F}_t : \mathcal{X} \times \mathcal{V} \to \mathcal{O}\}_{t=1}^{T}$; input distribution $D$; per-round $\alpha$-divergence budgets $\{r_t\}_{t=1}^{T}$; sample size $m$; global bound $B(\mathcal{D}_e)$.
2: Sample a secret input $\hat{\mathsf{x}} \sim D$.
3: Choose per-round targets $\{r'_t\}_{t=1}^{T}$, failure probabilities $\{\gamma_t\}_{t=1}^{T}$, and a fallback bound $R$ such that for every $t \in [T]$,

$$(1-\gamma_t)\, r'_t + \gamma_t\, R \le r_t.$$

4: **for** $t = 1, 2, \ldots, T$ **do**
5: Receive adversarial parameter $v_t$.
6: Define the *prefix coefficient*

$$A_{t-1}(\hat{\mathsf{x}}) := \Pr(\mathsf{x} = \hat{\mathsf{x}})^{\alpha-1} \left( \prod_{s=1}^{t-1} \frac{\Pr\big(\mathcal{M}_s(\hat{\mathsf{x}}, \mathsf{v}_s, \tau_s) = \mathsf{o}_s\big)}{\Pr\big(W_s(\mathsf{v}_s, \tau_s) = \mathsf{o}_s\big)} \right)^{\alpha}$$

$$D_t(\hat{\mathsf{x}}) := \mathcal{D}_\alpha(\mathcal{M}_t(\hat{\mathsf{x}}, \mathsf{v}_t, \tau_t) \,\|\, W_t(\mathsf{v}_t, \tau_t)),$$

and

$$\mathcal{G}_t(\mathcal{D}_e, \hat{\mathsf{x}}) \;:= A_{t-1}(\hat{\mathsf{x}})\, D_t(\hat{\mathsf{x}})$$

7: Take $\mathcal{F}(\cdot, \mathsf{v}_t), g_t(\mathcal{D}_e, \hat{\mathsf{x}}), D, r_t, B(\mathcal{D}_e)$ as the input for Algorithm 1
8: Sample noise $e_t$ from the calibrated distribution.
9: Output $\mathsf{o}_t \leftarrow \mathcal{F}_t(\hat{\mathsf{x}}, \mathsf{v}_t) + e_t$.
10: **end for**

For simplicity but without loss of generality, suppose $T = 2$, notice that inequality (98) can be rewritten as

$$D_{KL}\big(\mathbf{1}_{\delta_\rho} \,\big\|\, \mathbf{1}_{\delta_{o,\rho}}\big) \tag{99}$$

$$\le \mathbb{E}_{\mathsf{x}\sim D} \int_{\mathcal{O}\times\mathcal{V}} \Pr\big(\mathrm{Adv}_1(\tau_1) = \mathsf{v}_1\big) \Pr\big(W_1(\mathsf{v}_1, \tau_1) = \mathsf{o}_1\big) \int_{\mathcal{V}} \Pr\big(\mathrm{Adv}_1(\tau_2) = \mathsf{v}_2\big)$$
$$\cdot \frac{\Pr\big(\mathcal{M}_1(\mathsf{x}, \mathsf{v}_1, \tau_1) = \mathsf{o}_1\big)}{\Pr\big(W_1(\mathsf{v}_1, \tau_1) = \mathsf{o}_1\big)} \int_{\mathcal{O}} \Pr\big(\mathcal{M}_2(\mathsf{x}, \mathsf{v}_2, \tau_2) = \mathsf{o}_2\big) \Big(\log\Big(\frac{\Pr\big(\mathcal{M}_2(\mathsf{x}, \mathsf{v}_2, \tau_2) = \mathsf{o}_2\big)}{\Pr\big(W_2(\mathsf{v}_2, \tau_2) = \mathsf{o}_2\big)}\Big) + \log\Big(\frac{\Pr\big(\mathcal{M}_2(\mathsf{x}, \mathsf{v}_2, \tau_2) = \mathsf{o}_2\big)}{\Pr\big(W_2(\mathsf{v}_2, \tau_2) = \mathsf{o}_2\big)}\Big) \tag{100}$$

$$= \int_{\mathcal{O}\times\mathcal{V}} \Pr\big(\mathrm{Adv}_1(\tau_1) = \mathsf{v}_1\big) \Pr\big(W_1(\mathsf{v}_1, \tau_1) = \mathsf{o}_1\big) \int_{\mathcal{V}} \Pr\big(\mathrm{Adv}_1(\tau_2) = \mathsf{v}_2\big)$$
$$\cdot \mathbb{E}_{\mathsf{x}\sim D} \underbrace{\frac{\Pr\big(\mathcal{M}_1(\mathsf{x}, \mathsf{v}_1, \tau_1) = \mathsf{o}_1\big)}{\Pr\big(W_1(\mathsf{v}_1, \tau_1) = \mathsf{o}_1\big)}}_{(A)} \underbrace{\int_{\mathcal{O}} \Pr\big(\mathcal{M}_2(\mathsf{x}, \mathsf{v}_2, \tau_2) = \mathsf{o}_2\big) \log\Big(\frac{\Pr\big(\mathcal{M}_2(\mathsf{x}, \mathsf{v}_2, \tau_2) = \mathsf{o}_2\big)}{\Pr\big(W_2(\mathsf{v}_2, \tau_2) = \mathsf{o}_2\big)}\Big)\, \mathrm{d}\mathsf{o}_2}_{(B)}\, \mathrm{d}\mathsf{v}_2\, \mathrm{d}\mathsf{o}_1\, \mathrm{d}\mathsf{v}_1$$
$$+ \int_{\mathcal{O}\times\mathcal{V}} \Pr\big(\mathrm{Adv}_1(\tau_1) = \mathsf{v}_1\big) \underbrace{\mathbb{E}_{\mathsf{x}\sim D} \Pr\big(\mathcal{M}_1(\mathsf{x}, \mathsf{v}_1, \tau_1) = \mathsf{o}_1\big) \log\Big(\frac{\Pr\big(\mathcal{M}_1(\mathsf{x}, \mathsf{v}_1, \tau_1) = \mathsf{o}_1\big)}{\Pr\big(W_1(\mathsf{v}_1, \tau_1) = \mathsf{o}_1\big)}\Big)}_{(C)}$$
$$\underbrace{\int_{\mathcal{O}\times\mathcal{V}} \Pr\big(\mathcal{M}_2(\mathsf{x}, \mathsf{v}_2, \tau_2) = \mathsf{o}_2\big) \Pr\big(\mathrm{Adv}_1(\tau_2) = \mathsf{v}_2\big)\, \mathrm{d}\mathsf{o}_2\, \mathrm{d}\mathsf{v}_2}_{(D)}\, \mathrm{d}\mathsf{o}_1\, \mathrm{d}\mathsf{v}_1 \tag{101}$$

Notice that term (A) is $A_1(\mathsf{x})$, term (B) is $D_2(\mathsf{x})$, term (C) $\leq r_1$ and term (D) equals to 1, given the condition in the theorem, we know

$$\mathbb{E}_{\mathsf{x}\sim D}\big[A_1(\mathsf{x})\, D_2(\mathsf{x})\big] \;\leq\; r_2 \tag{102}$$

so

$$\begin{aligned}(101) \leq & r_2 \cdot \int_{\mathcal{O}\times\mathcal{V}} \Pr\big(\mathrm{Adv}_1(\tau_1) = \mathsf{v}_1\big)\Pr\big(W_1(\mathsf{v}_1,\tau_1) = \mathsf{o}_1\big) \int_{\mathcal{V}} \Pr\big(\mathrm{Adv}_1(\tau_2) = \mathsf{v}_2\big)\, \mathrm{d}\mathsf{v}_2\, \mathrm{d}\mathsf{o}_1\, \mathrm{d}\mathsf{v}_1 \\ & + r_1 \cdot \int_{\mathcal{O}\times\mathcal{V}} \Pr\big(\mathrm{Adv}_1(\tau_1) = \mathsf{v}_1\big)\mathrm{d}\mathsf{v}_1 \qquad (103) \\ = & \; r_1 + r_2 \qquad (104)\end{aligned}$$

## S Proof of Theorem 13

For each $\mathsf{x} \in \mathcal{X}$, write $P_{\mathsf{x}} := P_{\mathcal{M}(\mathsf{x})}$ for the output distribution of $\mathcal{M}$ given input $\mathsf{x}$. Let $W$ be the marginal (mixture) of these conditionals under $D$:

$$P_W(o) = \sum_{\mathsf{x}\in\mathcal{X}} \Pr_D(\mathsf{x})\, P_{\mathsf{x}}(o), \qquad o \in \mathcal{O}.$$

Let $Q$ be any other reference distribution on $\mathcal{O}$. For simplicity we assume $\mathcal{O}$ is finite or countable; the general case follows by replacing sums with integrals.

We start by expanding both sides:

$$\begin{aligned}\mathbb{E}_{\mathsf{x}\sim D}\mathcal{D}_{KL}(P_{\mathsf{x}}\|P_Q) &= \sum_{\mathsf{x}\in\mathcal{X}} \Pr_D(\mathsf{x}) \sum_{o\in\mathcal{O}} P_{\mathsf{x}}(o) \log \frac{P_{\mathsf{x}}(o)}{Q(o)}, \\ \mathbb{E}_{\mathsf{x}\sim D}\mathcal{D}_{KL}(P_{\mathsf{x}}\|P_W) &= \sum_{\mathsf{x}\in\mathcal{X}} \Pr_D(\mathsf{x}) \sum_{o\in\mathcal{O}} P_{\mathsf{x}}(o) \log \frac{P_{\mathsf{x}}(o)}{W(o)}.\end{aligned}$$

Subtracting the second from the first gives

$$\begin{aligned}&\mathbb{E}_{\mathsf{x}\sim D}\mathcal{D}_{KL}(P_{\mathsf{x}}\|P_Q) - \mathbb{E}_{\mathsf{x}\sim D}\mathcal{D}_{KL}(P_{\mathsf{x}}\|P_W) \\ &= \sum_{\mathsf{x}\in\mathcal{X}} \Pr_D(\mathsf{x}) \sum_{o\in\mathcal{O}} P_{\mathsf{x}}(o) \left(\log \frac{P_{\mathsf{x}}(o)}{Q(o)} - \log \frac{P_{\mathsf{x}}(o)}{W(o)}\right) \\ &= \sum_{\mathsf{x}\in\mathcal{X}} \Pr_D(\mathsf{x}) \sum_{o\in\mathcal{O}} P_{\mathsf{x}}(o) \log \frac{W(o)}{Q(o)} \\ &= \sum_{o\in\mathcal{O}} \left(\sum_{\mathsf{x}\in\mathcal{X}} \Pr_D(\mathsf{x}) P_{\mathsf{x}}(o)\right) \log \frac{W(o)}{Q(o)} \\ &= \sum_{o\in\mathcal{O}} W(o) \log \frac{W(o)}{Q(o)} = \mathcal{D}_{KL}(P_W\|P_Q) \;\geq\; 0,\end{aligned}$$

where the last inequality is Gibbs' inequality for KL divergence.

Rearranging,

$$\mathbb{E}_{\mathsf{x}\sim D}\mathcal{D}_{KL}(P_{\mathcal{M}(\mathsf{x})}\|P_W) \;\leq\; \mathbb{E}_{\mathsf{x}\sim D}\mathcal{D}_{KL}(P_{\mathcal{M}(\mathsf{x})}\|P_Q),$$

which proves that the marginal $W$ minimizes the average KL divergence over all choices of $Q$.

## T Proof of Theorem 14

By applying Theorem 1, we have

$$\mathbb{E}_{\mathsf{x}\sim D}\Big[\mathcal{G}(\mathcal{D}_e,\mathsf{x})\Big] \;\leq\; r_t.$$

Use $\mathcal{G}_t(\mathcal{D}_e,\hat{\mathsf{x}}) \;:=\; A_{t-1}(\hat{\mathsf{x}})\,D_t(\hat{\mathsf{x}})$ into the above equation for every t and every selection of we have

$$\mathbb{E}_{\mathsf{x}\sim D}\big[A_{t-1}(\mathsf{x})\,D_t(\mathsf{x})\big] \;\leq\; r_t \tag{105}$$

Consequently the condition 92 is held so we apply Theorem 12 we have:

$$D_{KL}\big(\mathbf{1}_{\delta_\rho}\,\big\|\,\mathbf{1}_{\delta_{o,\rho}}\big) \;\leq\; \sum_{t=1}^{T} r_t. \tag{106}$$