ARTIFACT EVALUATED
USENIX
AVAILABLE

Sliced Rényi Pufferfish Privacy: Tractable Privatization Mechanism and Private Learning with Gradient Clipping

Tao Zhang
Washington University in St. Louis

Yevgeniy Vorobeychik
Washington University in St. Louis

Abstract

We study the design of a privatization mechanism and privacy accounting in the Pufferfish Privacy (PP) family. Specifically, motivated by the curse of dimensionality and lack of practical composition tools for iterative learning in the recent Rényi Pufferfish Privacy (RPP) framework, we propose *Sliced Rényi Pufferfish Privacy (SRPP)*. SRPP preserves PP/RPP semantics (customizable secrets with probability-aware secret–dataset relationships) while replacing high-dimensional Rényi divergence with projection-based quantification via two sliced measures, *Average SRPP* and *Joint SRPP*. We develop *sliced Wasserstein mechanisms*, yielding sound SRPP certificates and closed-form Gaussian noise calibration. For iterative learning systems, we introduce an SRPP-SGD scheme with gradient clipping and new accountants based on *History-Uniform Caps (HUC)* and a *subsampling-aware* variant (sa-HUC), enabling decompose-then-compose privatization and additive composition under a common slicing geometry. Experiments on static and iterative privatization show that the proposed framework exhibits favorable privacy–utility trade-offs, as well as practical scalability.

1 Introduction

Differential Privacy (DP) [12] is the de facto standard for measuring data privacy risk. Under DP semantics, the secret is an individual record's presence or absence, and an adversary with arbitrary side information cannot draw substantially different conclusions from the released output about that inclusion. DP quantifies risk via ε-DP or (ε,δ)-DP, i.e., worst-case indistinguishability between neighboring datasets, equivalently expressed through bounds on max divergence and hockey-stick divergence, respectively. Rényi DP (RDP) [27, 32] preserves DP semantics but measures privacy using (α,ε)-Rényi divergence.

Pufferfish Privacy (PP) [21] generalizes DP semantics by allowing customizable secrets and modeling probabilistic secret–dataset relationships through *priors* that may represent subjective adversarial beliefs or an objective secret-data-generation process [46]. ε-PP and (ε,δ)-PP [49] adopt the same indistinguishability measures as DP, while Rényi Pufferfish Privacy (RPP) [34] shares PP semantics but uses (α,ε)-Rényi divergence.

Beyond semantics and a quantification measure, a useful privacy framework should provide implementable *privatization mechanisms*. For DP, Laplace [12] and Gaussian [11] mechanisms calibrate noise to query sensitivity. However, computing sensitivity is generally NP-hard [43]. DP and RDP composition properties underpin DP-SGD with gradient clipping [1], a *decompose-then-compose* approach for deep learning without sensitivity computation: each iteration enforces bounded sensitivity via gradient clipping, adds subsampled Gaussian noise, and a moments accountant converts per-iteration RDP costs into an overall (ε,δ)-DP guarantee.

The key technical difference between DP and PP is that PP involves a *probabilistic secret-dataset relationship*. Because a query only takes the realized dataset as input, PP analysis must account for the induced mapping from secret to output through the prior distributions. Consequently, practical PP calibration depends on prior-driven distributional quantities and is typically carried out from finite samples and estimators, so error-aware, assumption-dependent guarantees are often unavoidable for useful privacy-utility trade-offs.

For Pufferfish privatization, Song et al. [38] provide the first general-purpose privatization mechanism, the *Wasserstein mechanism*, which uses the ∞-Wasserstein distance as an analogue of sensitivity and generalizes the Laplace mechanism. Ding [10] proposes a Kantorovich-based exponential (and Gaussian) additive-noise mechanism for PP calibrated to 1-Wasserstein sensitivity. RPP [34] introduces the General Wasserstein Mechanism (GWM), which uses the shift-reduction lemma [14] to calibrate additive noise via ∞-Wasserstein distance. Extensions of GWM include a distribution-aware variant (DAGWM) using p-Wasserstein distances to capture average-case data geometry. While these Wasserstein mechanisms significantly advance the applicability of the PP family frameworks, they inherit the curse

of dimensionality from optimal transport (OT): estimating Wasserstein distances and sensitivities remains statistically and computationally expensive in high dimensions, limiting practical applicability. The tractable general mechanism design principles for PP family remain largely open [18].

Moreover, PP/RPP lack practical, general-purpose composition theorems for privacy accounting in stochastic optimization algorithms such as SGD. Existing results rely on restrictive structure (e.g., universally composable evolution scenarios [21] or Bayesian networks via Markov quilts [38]). Recent work [46] shows the obstacle is fundamental: probabilistic secret–dataset relationships induce inter-mechanism dependencies that break standard composition; their inverse composition requires solving an infinite-dimensional convex program, making integration with iterative learning nontrivial. RPP's Privacy Amplification by Iteration (PABI) [34] enables privacy guarantees for iterative convex optimization under contractivity, but still depends on Wasserstein distance computation, reintroducing high-dimensional OT costs.

Therefore, making PP-family guarantees operational in practice requires resolving two obstacles, both rooted in the probabilistic secret–dataset relationships encoded by priors: (i) a *curse of dimensionality in mechanism design*, namely calibrating noise without intractable high-dimensional Wasserstein computations; and (ii) a *lack of graceful composition for iterative optimization and learning*, namely enabling decompose-then-compose privatization for stochastic methods such as SGD under Pufferfish semantics.

Contributions. We address these obstacles by introducing *Sliced Rényi Pufferfish Privacy (SRPP)* (Sec. 4), a new Pufferfish-family privacy notion, together with two privatization routes with provable guarantees:

(i) the *Sliced Wasserstein Mechanism (SWM)* (Sec. 4.3), a direct mechanism calibrated by an efficiently computable *sliced Wasserstein sensitivity* (*SW-sensitivity*) (Sec. 3) that circumvents the curse of dimensionality of OT;

(ii) a decompose-then-compose scheme for iterative learning, based on *History-Uniform Caps (HUC)* and their *subsampling-aware* variant (*sa-HUC*) under gradient clipping, with matching SRPP/sa-SRPP accountants that avoid Wasserstein sensitivity computation altogether.

Sliced quantification. Calibrating a Wasserstein mechanism through full-dimensional Wasserstein sensitivity, as in RPP's General Wasserstein Mechanism, is bottlenecked by high-dimensional OT. Building on the sliced Wasserstein distance [3, 29, 35], we replace this sensitivity with SW-sensitivity, computed from one-dimensional (1-D) projections, and design SWM around it (Sec. 4.3). To align the privacy quantification with this sensitivity, we introduce two SRPP notions: *Average SRPP* (Sec. 4.1) and *Joint SRPP* (Sec. 4.2). Both share the Pufferfish semantics of PP/RPP but quantify privacy using *Average* and *Joint Sliced Rényi Divergence* (Ave-SRD and Joint-SRD), respectively.

Iterative learning. We develop an SRPP privatization route for iterative learning via HUC and sa-HUC, enabling DP-SGD-style privatization with gradient clipping *without* requiring contractivity of the update map. A central element of the contribution is that HUC/sa-HUC isolate a single scenario-dependent quantity—the *discrepancy cap* (in deterministic form K_t and subsampling-aware form $\overline{K}_t^2$)—that governs the per-iteration secret-induced shift. This identification (i) avoids the utility-draining worst-case "group-privacy" [34, 38] bound (see App. A.4 for a formal discussion), and (ii) enables a decompose-then-compose accountant for SGD with gradient clipping, without computing (sliced) Wasserstein sensitivities.

Composition of the SRPP-SGD mechanisms. For the HUC/sa-HUC-based SRPP-SGD mechanisms developed in this paper, we prove additive composition under a shared slice profile: when several such mechanisms act on the same dataset under a common slicing geometry, the total (sa-)Ave-/Joint-SRPP cost is bounded by the sum of the per-mechanism costs (Theorem 5.8). This gives modular accounting for pipelines and cascades built from these mechanisms.

Experiments on static query release and iterative learning demonstrate favorable privacy–utility trade-offs and practical scalability. A complete appendix, including formal proofs and detailed analysis, appears in the online extended version of this paper [48].

1.1 Related Work

Privacy Beyond Individual Data Records. One-sided DP [23] protects a designated sensitive subset of the dataset while allowing minimal or no protection for the remaining records. Fernandes et al. [15] use generalized DP with Earth Mover's Distance on word embeddings to obfuscate authorship. Hardt and Roth [16] study private singular vector computation under entry-level privacy, where each matrix entry is a secret. PAC Privacy [39, 41] and Residual PAC Privacy [47] bound the information-theoretic hardness of inference for general secrets, enabling simulation-based measurement and control. Pufferfish Privacy (PP) [21, 38] and Rényi Pufferfish Privacy (RPP) [34] generalize DP and RDP to arbitrary secrets within the same indistinguishability framework. Related distribution-level notions protect properties of the data distribution rather than individual records [4].

Instantiations of Pufferfish Privacy. PP has been instantiated for correlated data [38], attribute privacy [49], and confounding settings where secrets are data-derived [46]. Related lines include Blowfish, which uses policy constraints to flexibly specify what is protected and what is known [17], and mutual-information PP, an information-theoretic formulation based on conditional mutual information with improved composition [31]. Applied deployments include smart-meter time-series release and local energy markets [20] and temporally correlated time-series trading (HORAE) [30].

2 Preliminaries

In this section, we review Pufferfish-type privacy, with a main focus on Rényi Pufferfish Privacy (RPP) [34] and Wasserstein mechanisms, laying the necessary background for Sliced Rényi Pufferfish Privacy (SRPP). Formal introductions of DP and DP-SGD are deferred to App. A.1.

A distinguishability-based privacy framework can be viewed through three layers. The first is a *semantic interpretation*, specifying the protected secrets, the secret pairs to be made indistinguishable, and the assumptions under which the guarantee is required. The second is a *quantification measure*, specifying how distinguishability between the corresponding mechanism-output distributions is bounded. Together, these two layers characterize the privacy notion. The third layer consists of compatible *privatization mechanisms*, which provide practical ways to implement the guarantee. Guided by this perspective, Sec. 2.4 clarifies the status and interpretation of SRPP by separating its Pufferfish semantic goal from its sliced Rényi divergence quantification. It also explains how SRPP relates to DP, PP, and RPP. Sec. 2.5 then provides concrete intuition and real-world examples illustrating what types of information are, and are not, protected under these definitions.

2.1 Semantic Interpretation

DP has a clear semantic interpretation [11, 19]: *regardless of auxiliary knowledge, an adversary who observes the mechanism output should not reach substantially different conclusions depending on whether an individual record was included in the input database.*

A *Pufferfish scenario* $(\mathcal{S}, Q, \Theta)$ specifies the semantic layer. $\mathcal{S}$ is the set of protected secrets, and $Q \subseteq \mathcal{S} \times \mathcal{S}$ specifies which pairs of secrets should be made indistinguishable. Unlike DP, whose neighboring relation is defined directly over realized datasets, the Pufferfish-type framework models X as a random variable drawn from a distribution $\theta \in \Theta$, where each θ is a *prior*, i.e., a probabilistic model capturing either adversarial beliefs about the data or an admissible data-generating scenario, including secret–data relationships. Crucially, the Pufferfish family is not a prior-dependent relaxation of DP: privacy is defined uniformly over an admissible family of priors Θ and over secret pairs in Q. A Pufferfish-type guarantee therefore requires specifying Θ; practical mechanisms may additionally require estimating the relevant prior-driven sensitivity or output-distribution quantities.

Given $(\mathcal{S}, Q, \Theta)$, we can state a Pufferfish semantic interpretation (Pufferfish semantics) in the spirit of DP: *for every admissible prior $\theta \in \Theta$, an adversary observing the mechanism output should reach nearly indistinguishable conclusions about whether s_i or s_j holds, for every $(s_i, s_j) \in Q$.*

This Pufferfish semantics includes DP as a special case. In particular, DP can be represented through a Pufferfish scenario $(\mathcal{S}_{\mathrm{DP}}, Q_{\mathrm{DP}}, \Theta_{\mathrm{DP}})$ where $\mathcal{S}_{\mathrm{DP}}$ contains secrets about individual records, Q_{DP} pairs alternatives that differ only in one individual record, and Θ_{DP} restricts records to be mutually independent. Under this independence structure, conditioning on one individual's secret does not change the distribution of the remaining records, so the Pufferfish comparison reduces to the usual DP comparison between neighboring datasets. Thus, DP semantics can be viewed as the record-level, distribution-free special case of the broader Pufferfish semantics.

2.2 Quantification Measure and Definition

Any formal distinguishability-based privacy definition requires a quantification measure that specifies how indistinguishability is mathematically bounded. Although DP and PP have different semantics and scopes, classic PP [21, 49] uses the same DP-style classes of indistinguishability measures, primarily *max divergence* and *hockey-stick divergence*, but applies them to secret-conditioned output distributions induced by a Pufferfish scenario. RPP retains the Pufferfish semantics of PP while using *Rényi divergence* as the quantification measure for privacy loss, in direct analogy with how Rényi Differential Privacy (RDP) [27] uses Rényi divergence under the DP semantics.

For each $\theta \in \Theta$, let $P_\theta(S, X)$ denote the joint distribution over secrets S and datasets X, with marginals P_θ^S and P_θ^X, and conditional $P_\theta(x \mid s) = P_\theta(s, x)/P_\theta^S(s)$. We write $\Pr[\cdot \mid s, \theta]$ to denote the probability of an event under the conditional distribution $P_\theta(\cdot \mid s)$.

Definition 2.1 (Rényi Pufferfish Privacy (RPP) [34])**.** *Let $\alpha > 1$ and $\varepsilon \geq 0$. A mechanism $\mathcal{M}$ is (α, ε)-RPP in $(\mathcal{S}, Q, \Theta)$ if for all $\theta \in \Theta$ and $(s_i, s_j) \in Q$ with $P_\theta^S(s_i), P_\theta^S(s_j) > 0$,*

$$\mathtt{D}_\alpha\big(\Pr[\mathcal{M}(X) \mid S = s_i, \theta] \,\big\|\, \Pr[\mathcal{M}(X) \mid S = s_j, \theta]\big) \;\leq\; \varepsilon,$$

where $\mathtt{D}_\alpha$ is the Rényi divergence of order α:

$$\mathtt{D}_\alpha(\mathrm{P} \,\|\, \mathrm{Q}) := \tfrac{1}{\alpha - 1} \log \mathbb{E}_{Z \sim \mathrm{Q}}\left[\left(\tfrac{\mathrm{P}(Z)}{\mathrm{Q}(Z)}\right)^{\alpha}\right], \tag{1}$$

for distributions P *and* Q *on a common measurable space.*

RPP enjoys post-processing immunity. However, like classic PP, it does not provide the graceful composition behavior of the DP family [34]. This limits the direct use of RPP or PP to privatize iterative learning in a DP-SGD style.

We recall the PP definition here to highlight that PP and RPP share the same Pufferfish semantics but use different quantification measures. A mechanism $\mathcal{M}$ satisfies $(\hat{\varepsilon}, \delta)$-PP if for all $\theta \in \Theta$, all $(s_i, s_j) \in Q$ such that $P_\theta^S(s_i) > 0$ and $P_\theta^S(s_j) > 0$, and all measurable output events $\mathcal{T} \subseteq \mathcal{Y}$,

$$\Pr\big[\mathcal{M}(X) \in \mathcal{T} \mid S = s_i, \theta\big] \leq e^{\hat{\varepsilon}} \Pr\big[\mathcal{M}(X) \in \mathcal{T} \mid S = s_j, \theta\big] + \delta,$$

where $\hat{\varepsilon} \geq 0$ and $\delta \in [0, 1]$. This condition is mathematically equivalent to bounding the hockey-stick divergence. When $\delta = 0$, the mechanism satisfies pure $\hat{\varepsilon}$-PP, and the underlying quantification measure simplifies to the max divergence.

2.3 Wasserstein Mechanisms

Wasserstein mechanisms form a general class of privatization mechanisms for PP [38] and RPP [34]. In the three-layer perspective above, they belong to the mechanism layer: they do not change the protected secrets or the divergence used to quantify privacy, but provide a way to calibrate the additive noise to a Wasserstein-distance-based analogue of DP sensitivity, which we refer to as the *Wasserstein sensitivity*. We first recall the optimal transport (OT) notions: couplings and the ∞-Wasserstein distance.

Definition 2.2 (Coupling)**.** *Let μ and ν be probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. A coupling of (μ,ν) is any probability measure π on $(\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}(\mathbb{R}^d) \otimes \mathcal{B}(\mathbb{R}^d))$ whose marginals are μ and ν. We denote the set of all couplings by $\Pi(\mu,\nu)$.*

Definition 2.3 (∞-Wasserstein Distance)**.** *Fix a norm $\|\cdot\|$ on $\mathbb{R}^d$. For probability measures μ and ν on $\mathbb{R}^d$, the ∞-Wasserstein distance (∞-WD) is $W_\infty(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \sup_{(x,y)\in\mathrm{supp}(\pi)} \|x-y\|$.*

For a *query function* $f : \mathcal{X} \to \mathbb{R}^d$, the *$\infty$-Wasserstein sensitivity* is then defined as

$$\Delta_\infty := \max_{(s_i,s_j)\in Q,\ \theta\in\Theta} W_\infty\big(\Pr[f(X) \mid s_i,\theta], \Pr[f(X) \mid s_j,\theta]\big). \quad (2)$$

Let ζ be the noise distribution. Pierquin et al. [34] provide the General Wasserstein Mechanism (GWM), a general formulation describing how Δ_∞, ζ, and α together determine an upper bound on the Rényi divergence of $\mathcal{M}(X) = f(X) + N$, with $N \sim \zeta$, based on the *shift reduction lemma* [14]. When $\zeta = \mathcal{N}\left(0, \frac{\alpha\Delta_\infty^2}{2\varepsilon} I_d\right)$, with Δ_∞ computed with respect to the ℓ_2 norm, $\mathcal{M}(X)$ satisfies (α,ε)-RPP. The main limitation of Wasserstein mechanisms is the scalability bottleneck of optimal transport (OT), which is exacerbated in high dimensions. In particular, computing W_∞, or more generally W_p (see Definition A.2 in App. A.3) between empirical distributions requires solving large-scale OT problems whose complexity grows superlinearly with the sample size. This motivates replacing Δ_∞ by a sliced Wasserstein sensitivity, introduced in Sec. 3.

2.4 SRPP: Overview

We introduce Sliced Rényi Pufferfish Privacy (SRPP). To clarify its status relative to existing frameworks, SRPP retains the broad Pufferfish semantics shared by classic PP and RPP, while replacing the full-dimensional Rényi divergence used by RPP with *Sliced Rényi Divergence* (SRD). We develop SRD under two variants: Average SRD (Definition 4.1) and Joint SRD (Definition 4.3). Under the three-layer perspective above, DP/RDP differ from Pufferfish-family notions primarily in semantic scope: DP/RDP protect record-level neighboring-dataset secrets, whereas PP/RPP/SRPP protect the secret pairs specified by a Pufferfish scenario. Within the Pufferfish semantics, PP, RPP, and SRPP notions differ in the quantification measure used to bound distinguishability.

To establish its status as a formal privacy notion, SRPP should be understood as a change in the quantification layer, while keeping the semantic privacy goal of PP/RPP intact. Semantically, because the dataset X is the input to a query or training algorithm, leakage about a protected secret is reflected in the difference between the secret-conditioned output distributions, under the admissible priors in Θ. Bounding this distinguishability is a meaningful privacy principle: if the output distributions for two protected secrets are hard to distinguish, then an adversary observing the release has limited ability to infer which secret holds. This shared semantic objective unites PP, RPP, and SRPP; they differ in how the output-distribution distinguishability is quantified.

To quantify this distinguishability, SRPP uses SRD. The use of sliced quantification is motivated by the sliced Wasserstein distance (SWD) literature: SWD compares high-dimensional distributions by aggregating discrepancies of 1-D projections. Analogously, SRD evaluates Rényi divergence after 1-D projections and aggregates these per-slice divergences according to a slice profile. Thus, SRPP is not an approximation of RPP obtained by estimating the full-dimensional Rényi divergence. Rather, SRPP is a distinct Pufferfish-family guarantee: it keeps the Pufferfish semantics but uses SRD as its quantification measure. The slice profile is part of this quantification measure, not an assumption that the adversary only observes projected outputs.

The primary motivation for introducing SRPP is operational tractability. As noted in Sec. 2.3, existing Wasserstein mechanisms are calibrated through the full-dimensional ∞-Wasserstein sensitivity Δ_∞, or related p-Wasserstein quantities [34]. Computing these quantities requires large-scale OT computations and suffers from the curse of dimensionality, making direct calibration difficult for high-dimensional functions. By moving to a sliced quantification measure, SRPP aligns the privacy definition with *sliced Wasserstein sensitivity* (Sec. 3). This replaces full-dimensional OT sensitivity computation by 1-D projected sensitivity computations. The change from RPP to SRPP is therefore a change in the quantification layer and the resulting calibration formula, not a change in the basic perturbation format: the mechanism may still release $f(X)+N$ with full-dimensional Gaussian noise. For iterative learning, we develop SRPP-SGD with gradient clipping and HUC/sa-HUC accountants, providing a DP-SGD-style route to tractable per-iteration calibration and additive accounting under Pufferfish semantics.

Threat Model. We consider a standard trusted-curator setting: the data holder runs a randomized mechanism $\mathcal{M}$ on the dataset and releases its output (e.g., a query answer or trained model). The adversary observes the release and may know both $\mathcal{M}$ and the Pufferfish scenario $(\mathcal{S}, Q, \Theta)$. Side information about the data-generating process or secret–data relationships is modeled through the admissible prior set Θ.

Privacy is required for all $(s_i, s_j) \in Q$ uniformly over $\theta \in \Theta$. In the SGD setting, the adversary observes the trained model or updates released by our algorithm, but not internal randomness beyond what the release reveals. The guarantee is therefore relative to the specified Pufferfish scenario: if Θ omits plausible secret-data relationships, the guarantee need not cover those omitted relationships. In this paper, we take $(\mathcal{S}, Q, \Theta)$ as the intrinsic modeling setting and focus on the resulting mechanistic and computational challenges of privatization mechanism design.

2.5 Examples

We give two examples to illustrate the semantic coverage of Pufferfish-family guarantees. In both cases, the protected fact is not simply the inclusion or value of one independent record, but a secret whose truth changes the distribution of the data through temporal correlation or aggregate structure.

Example 1: Temporal correlation in time-series [38]. Consider releasing an activity histogram over a long time series $X = \{X_1, \ldots, X_T\}$ while protecting the activity at a specific time t ($X_t = a \in \mathcal{A}$). A Pufferfish scenario can take $\mathcal{S} = \{s_a^t : s_a^t = \{X_t = a\}, a \in \mathcal{A}, t \in [T]\}$, $Q = \{(s_a^t, s_b^t) : a \neq b, t \in [T]\}$, and Θ as a class of time-homogeneous Markov chains capturing temporal correlations in human behavior [38]. Here, the probabilistic secret–dataset relationship is explicit: conditioning on the secret $s_a^t = \{X_t = a\}$ changes the conditional distribution of nearby time points through the Markov transition model in Θ. This protects activity-at-time-t secrets under the models in Θ. In this setting, standard record-level DP does not directly apply because the time series belongs to a single subject; record-level DP does not erase evidence of activities that persist across correlated time points, whereas the group-privacy baseline used in [38] adds noise that scales with T and can destroy utility. Pufferfish-family notions instead use Θ to calibrate noise to the modeled secret-induced influence across the time series.

Example 2: Macro-level attribute privacy [49]. Consider releasing aggregate statistics, such as average height and weight, over a clinical cohort while protecting a dataset-level attribute, such as the cohort's gender distribution. A Pufferfish scenario can specify aggregate secrets such as $s_1 = \{\text{fraction of female patients} \geq 0.7\}$ and $s_2 = \{\text{fraction of female patients} \leq 0.3\}$, include (s_1, s_2) in Q, and let Θ model correlations among biometric features and the protected aggregate attribute, e.g., via a Gaussian model [49]. This protects a macro-level property of the dataset. Record-level DP does not directly target this goal, whereas Pufferfish-family notions calibrate noise to reduce distinguishability between the specified aggregate secrets under Θ.

These examples illustrate that the Pufferfish family is useful when the protected fact is not merely the inclusion or value of one independent record, but a secret for which conditioning on different secret values induces different data distributions under the specified scenario. In such settings, the scenario $(\mathcal{S}, Q, \Theta)$ explicitly specifies the protected secrets, the alternatives to be made indistinguishable, and the admissible secret–data relationships. DP can be recovered as a special record-level case under independence assumptions, but Pufferfish-type notions allow the privacy statement to target temporally correlated secrets or dataset-level aggregate attributes directly.

What is not directly certified. The coverage of any Pufferfish-family guarantee is limited to the specified scenario $(\mathcal{S}, Q, \Theta)$. For example, an SRPP guarantee targeting a global demographic attribute does not automatically protect unrelated temporal patterns or individual-level secrets unless those secrets and relationships are included in the scenario. Thus, the scenario must be chosen to match the real-world privacy risks that the deployment intends to protect.

3 Sliced Wasserstein Sensitivity

In this section, we introduce the *Sliced Wasserstein sensitivity (SW-sensitivity)*, obtained by slicing the ∞-WD, to mitigate the computational bottlenecks of Wasserstein mechanisms in high dimensions.

We first introduce notation. Let $f : \mathcal{X} \to \mathbb{R}^d$ be a *query* function. For each $\theta \in \Theta$ and $s \in \mathcal{S}$, let $P_\theta^{f,s}$ denote the conditional distribution of the (non-perturbed) query output $f(X)$ given $S = s$, i.e., $P_\theta^{f,s} := P_\theta(f(X) \mid s)$. For a measurable map $g : \mathcal{W} \to \mathcal{C}$ and a Borel measure μ on $\mathcal{W}$, the *pushforward* $g_\# \mu$ on $\mathcal{C}$ is defined by

$$g_\#\mu(A) \;=\; \mu\big(g^{-1}(A)\big), \qquad \forall A \in \mathcal{B}(\mathcal{C}), \tag{3}$$

where $\mathcal{B}(\mathcal{C})$ denotes the Borel σ-algebra on $\mathcal{C}$. Let $\mathbb{S}^{d-1} = \{u \in \mathbb{R}^d : \|u\|_2 = 1\}$ be the unit sphere and let ω be a probability measure on $\mathbb{S}^{d-1}$. For each *direction* $u \in \mathbb{S}^{d-1}$, define the projection $\Psi^u : \mathbb{R}^d \to \mathbb{R}$ by $\Psi^u(a) = \langle a, u \rangle$. Then, for any Borel probability measure P on $\mathbb{R}^d$, the pushforward $\Psi^u_\# \mathrm{P}$ is the 1-D distribution of $\langle \hat{X}, u \rangle$ when $\hat{X} \sim \mathrm{P}$. For a direction u, we refer to the induced 1-D projection Ψ^u as a *slice*.

Next, we define the *sliced Wasserstein distance*, which defines a metric for comparing probability measures on $\mathbb{R}^d$ by integrating WD over all 1-D projections.

Definition 3.1 (Sliced Wasserstein Distance (SWD) [35])**.** *Let $\mathcal{P}_p(\mathbb{R}^d)$ denote the set of all Borel probability measures on $\mathbb{R}^d$ with finite p-th moment, i.e., $\int_{\mathbb{R}^d} \|x\|^p \, \mathrm{d}\mu(x) < \infty$ for all $\mu \in \mathcal{P}_p(\mathbb{R}^d)$. For $p \geq 1$ and $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, the p-SWD is defined by*

$$\mathrm{SW}_p^p(\mu, \nu) \;:=\; \int_{\mathbb{S}^{d-1}} W_p^p(\Psi^u_\#\mu,\, \Psi^u_\#\nu)\, \mathrm{d}\lambda(u), \tag{4}$$

where λ is the uniform probability measure on $\mathbb{S}^{d-1}$.

The SWD is appealing because it replaces a high-dimensional OT comparison by an expectation over 1-D projections, where W_p admits a closed form (including $p = \infty$) in

terms of quantile functions. By the Cramér-Wold theorem [6], a probability measure on $\mathbb{R}^d$ is uniquely determined by the totality of its 1-D projections across all directions $u \in \mathbb{S}^{d-1}$. Consequently, the SWD is a proper metric.

In practice, SW_p is typically estimated by Monte Carlo sampling of directions and sorting the projected samples, reducing the computation to projections and 1-D Wasserstein evaluations. This yields *finite* slicing (or empirical slice distribution) and thus an approximate SWD estimator (See Sec. 4.3.2 for formal finite-sample analyses). The resulting finite-slice quantity remains a pseudometric (it is nonnegative, symmetric, and satisfies the triangle inequality).

In one dimension, $W_p(\hat{\mu}, \hat{\nu})$ admits the representation [33, Remark 2.30]:

$$W_p^p(\hat{\mu}, \hat{\nu}) = \int_0^1 \left| \mathsf{F}_{\hat{\mu}}^{-1}(u) - \mathsf{F}_{\hat{\nu}}^{-1}(u) \right|^p \mathrm{d}u, \tag{5}$$

where $\mathsf{F}_{\hat{\mu}}^{-1}$ and $\mathsf{F}_{\hat{\nu}}^{-1}$ denote the quantile functions of $\hat{\mu}$ and $\hat{\nu}$. The map $\hat{\mu} \mapsto \mathsf{F}_{\hat{\mu}}^{-1}$ (the same for $\hat{\nu} \mapsto \mathsf{F}_{\hat{\nu}}^{-1}$) gives an isometric embedding of $(\mathcal{P}_p(\mathbb{R}), W_p)$ into $L_p([0,1])$. For $p = 2$ the metric is induced by the L_2 inner product on this embedded subset, yielding a simple, essentially linear geometry compared to the much richer geometry in higher dimensions.

SW_∞ can be viewed as the $p \to \infty$ endpoint of the sliced Wasserstein family SW_p.[1] At the slice level, SW_∞ provides a natural analogue of the W_∞ (Definition 2.3) for sensitivity analysis in PP/RPP: it captures the largest possible shift in the projected query output induced by changing the secret.

Given a Pufferfish scenario $(\mathcal{S}, Q, \Theta)$, for each direction $u \in \mathbb{S}^{d-1}$, we define the *per-slice ∞-Wasserstein sensitivity (SW-sensitivity)* as

$$\Delta_\infty^u := \max_{(s_i, s_j) \in Q,\ \theta \in \Theta} W_\infty\Big(\Psi_\#^u P_\theta^{f, s_i},\ \Psi_\#^u P_\theta^{f, s_j} \Big), \tag{6}$$

which is the worst-case ∞-WD between the 1-D projections of $P_\theta^{s_i}$ and $P_\theta^{s_j}$ along direction u. Equivalently, it quantifies the maximal shift in the projected query output when the secret changes from s_i to s_j under any $\theta \in \Theta$.

Conceptually, we might hope to replace the direction-dependent sensitivity Δ_∞^u in (6) by a single scalar derived from SW_p. However, SW_p is an *average* over directions of 1-D transport costs and therefore does not provide a uniform, worst-case bound across all $u \in \mathbb{S}^{d-1}$. In contrast, our per-slice Rényi envelope (9) (Sec. 4.3) requires controlling each slice, and within each slice taking the worst case over admissible secret pairs and priors. In particular, an average can be small even when a subset of directions has large sensitivity.

Moreover, an adversary observes the d-dimensional query output, not a 1-D projection. Thus, calibrating noise using a sliced quantity (e.g., an average of Δ_∞^u over u) while retaining the original RPP definition introduces a mismatch between the geometry used for calibration and the divergence geometry in the privacy definition, and can therefore lead to unsound privacy bounds. In particular, for any $u \in \mathbb{S}^{d-1}$, the projection $\Psi^u(x) = \langle x, u \rangle$ is 1-Lipschitz with respect to the Euclidean norm, i.e., $|\Psi^u(x) - \Psi^u(y)| \le \|x - y\|_2$. Hence, for any probability measures μ, ν on $\mathbb{R}^d$, $W_\infty\big(\Psi_\#^u \mu,\ \Psi_\#^u \nu\big) \le W_\infty(\mu, \nu)$, where W_∞ on $\mathbb{R}^d$ is computed with the ℓ_2 norm. Indeed, for any coupling $\gamma \in \Pi(\mu, \nu)$ we have $|\Psi^u(x) - \Psi^u(y)| \le \|x - y\|_2$ γ-almost surely. Taking the essential supremum over $(x, y) \in \mathrm{supp}(\gamma)$ and then the infimum over γ yields the inequality. Applying this bound to $\mu = P_\theta^{s_i}$ and $\nu = P_\theta^{s_j}$, and then taking the maximum over $(s_i, s_j) \in Q$ and $\theta \in \Theta$, gives $\Delta_\infty^u \le \Delta_\infty$, where Δ_∞ is defined in (2).

This raises a natural question: which privacy notion is compatible with sliced Wasserstein sensitivity? We answer it by defining *Sliced Rényi Pufferfish Privacy* in the next section.

4 Sliced Rényi Pufferfish Privacy

In this section, we introduce *Sliced Rényi Pufferfish Privacy* (SRPP), a directional refinement of RPP defined relative to a *slice profile* $(\mathcal{U}, \omega)$, where $\mathcal{U} \subseteq \mathbb{S}^{d-1}$ is a set of directions and ω is a probability measure on $\mathbb{S}^{d-1}$ supported on $\mathcal{U}$. In the finite case, $\mathcal{U} = u_1, \ldots, u_m$ and ω is discrete with weights $(\omega_1, \ldots, \omega_m)$, so that $\int_{\mathbb{S}^{d-1}} f(u)\, \mathrm{d}\omega(u) = \sum_{i=1}^m \omega_i f(u_i)$.

We define two variants: Average SRPP (Ave-SRPP; Sec. 4.1) and Joint SRPP (Joint-SRPP; Sec. 4.2). Ave-SRPP bounds the ω-average of per-slice Rényi divergences. Joint-SRPP aggregates per-slice divergences through a joint log-moment bound, making it more sensitive to rare high-risk directions. Sec. 4.3 formally introduce *Sliced Wasserstein Mechanisms* using SW-sensitivity defined by Sec. 3. We defer additional discussions of SRPP, including the properties such as post-processing immunity and the conversion to *sliced PP*, and the difference between Ave- and Joint-SRPP in App. B.

4.1 Average SRPP

Our Ave-SRPP aggregates per-slice Rényi divergences by averaging, where the quantification measure is *Average Sliced Rényi Divergence.*

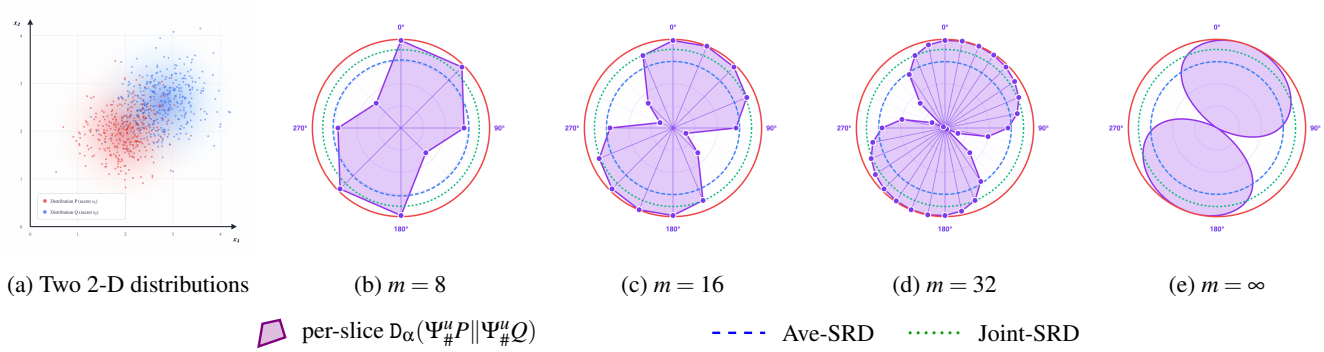
Definition 4.1 (Average Sliced Rényi Divergence (Ave-SRD))**.** *For two probability distributions* P *and* Q *on* $\mathbb{R}^d$*, and a slice profile* $(\mathcal{U}, \omega)$*, the Ave-SRD of order* $\alpha > 1$ *with respect to* ω *is*

$$\mathtt{AveSD}_\alpha^\omega(\mathrm{P} \parallel \mathrm{Q}) := \int_{\mathbb{S}^{d-1}} \mathtt{D}_\alpha(\Psi_\#^u \mathrm{P} \parallel \Psi_\#^u \mathrm{Q})\, \mathrm{d}\omega(u), \tag{7}$$

where $\mathtt{D}_\alpha$ *is the standard Rényi divergence.*

Ave-SRD averages the 1-D Rényi divergences between the projected measures $\Psi_\#^u \mathrm{P}$ and $\Psi_\#^u \mathrm{Q}$ over directions $u \sim \omega$. When $d = 1$, $\mathtt{AveSD}_\alpha^\omega$ reduces to the standard $\mathtt{D}_\alpha(\mathrm{P} \| \mathrm{Q})$.

[1] Formally, this corresponds to taking the $p \to \infty$ limit of the per-slice L_p aggregation of 1-D Wasserstein distances.



(a) Two 2-D distributions (b) $m = 8$ (c) $m = 16$ (d) $m = 32$ (e) $m = \infty$

Figure 1: SRD profiles in $\mathbb{R}^2$ at order $\alpha = 4$ under a uniform slice distribution ω. (a) Two 2-D Gaussian distributions P (red) and Q (blue). (b)–(d) Discrete slice profiles with $m \in \{8, 16, 32\}$ directions: the filled purple polygon traces the per-direction divergences $\mathtt{D}_\alpha(\Psi^u_\# P \| \Psi^u_\# Q)$. The blue dashed circle is Ave-SRD (directional mean), and the green dotted circle is Joint-SRD (log-moment aggregation, emphasizing large divergences). (e) The continuous-direction limit $m = \infty$.

Definition 4.2 ($(\alpha, \varepsilon, \omega)$-Ave-SRPP)**.** *Let $\alpha > 1$ and $\varepsilon \geq 0$. Given a slice profile $(\mathcal{U}, \omega)$, a mechanism $\mathcal{M}$ satisfies $(\alpha, \varepsilon, \omega)$-Ave-SRPP in $(\mathcal{S}, Q, \Theta)$ if for all $\theta \in \Theta$ and all $(s_i, s_j) \in Q$ with $P^S_\theta(s_i), P^S_\theta(s_j) > 0$,*

$$\mathtt{AveSD}^\omega_\alpha\big(\Pr[\mathcal{M}(X) \mid s_i, \theta] \,\big\|\, \Pr[\mathcal{M}(X) \mid s_j, \theta]\big) \leq \varepsilon.$$

Thus, Ave-SRPP bounds the ω-average of per-slice Rényi divergences between the output distributions for any admissible secret pair (uniformly over $\theta \in \Theta$).

4.2 Joint SRPP

Joint-SRPP aggregates per-slice divergences through a joint log-moment (Rényi) bound. Its quantification measure is the *Joint Sliced Rényi Divergence*.

Definition 4.3 (Joint Sliced Rényi Divergence (Joint-SRD))**.** *For two probability distributions* P *and* Q *on $\mathbb{R}^d$ and a slice profile $(\mathcal{U}, \omega)$, the Joint-SRD of order $\alpha > 1$ is*

$$\mathtt{JSD}^\omega_\alpha(\mathrm{P} \,\|\, \mathrm{Q}) := \tfrac{1}{\alpha - 1} \log \int_{\mathbb{S}^{d-1}} \exp\Big((\alpha - 1)\mathtt{D}_\alpha(\Psi^u_\# \mathrm{P} \,\|\, \Psi^u_\# \mathrm{Q})\Big)\, \mathrm{d}\omega(u). \tag{8}$$

Joint-SRD aggregates per-slice divergences via a log-moment transform, making it more sensitive than Ave-SRD to rare directions with large 1-D divergence (see Figure 1). When $d = 1$, $\mathtt{JSD}^\omega_\alpha(\mathrm{P} \| \mathrm{Q})$ reduces to the standard $\mathtt{D}_\alpha(\mathrm{P} \| \mathrm{Q})$.

Definition 4.4 ($(\alpha, \varepsilon, \omega)$-Joint-SRPP)**.** *Let $\alpha > 1$ and $\varepsilon \geq 0$. Given a slice profile $(\mathcal{U}, \omega)$, a mechanism $\mathcal{M}$ satisfies $(\alpha, \varepsilon, \omega)$-Joint-SRPP in $(\mathcal{S}, Q, \Theta)$ if for all $\theta \in \Theta$ and all $(s_i, s_j) \in Q$ with $P^S_\theta(s_i), P^S_\theta(s_j) > 0$,*

$$\mathtt{JSD}^\omega_\alpha\big(\Pr(\mathcal{M}(X) \mid s_i, \theta) \,\big\|\, \Pr(\mathcal{M}(X) \mid s_j, \theta)\big) \leq \varepsilon.$$

4.3 Sliced Wasserstein Mechanism

In this section, we present the *Sliced Wasserstein Mechanism (SWM)* to obtain SRPP guarantee by leveraging the computational tractability afforded by SW-sensitivity of Sec. 3.

Let $N \sim \zeta$ denote additive noise with distribution ζ on $\mathbb{R}^d$. For $a \in \mathbb{R}^d$, write ζ_{-a} for the shifted distribution (i.e., the distribution of $U - a$ for $U \sim \zeta$), where $\zeta_{-a}(w) = \zeta(w - a)$. For any $\pi : \mathbb{S}^{d-1} \to \mathbb{R}_+$ and direction $u \in \mathbb{S}^{d-1}$, we define the *per-slice Rényi envelope* in as

$$R_\alpha(\zeta, \pi(u)) = \sup_{\|a\| < \pi(u)} \mathtt{D}_\alpha\big(\zeta_{-a}, \zeta\big). \tag{9}$$

Recall $P^{f,s}_\theta := P_\theta(f(X) \mid s)$ for a query function f. When $\mathcal{M}(X) = f(X) + N$ with $N \sim \zeta$ independent of $f(X)$, let $\mathcal{M}^\theta_s = P^{f,s}_\theta * \zeta$ be the distribution of $\mathcal{M}(X)$ where $*$ is the convolution operator.

Lemma 4.1. *Fix a Pufferfish scenario $(\mathcal{S}, Q, \Theta)$ and a slice profile $(\mathcal{U}, \omega)$. Let $\mathcal{M}(X) = f(X) + N$ with $N \sim \zeta$ independent of X. For any $\theta \in \Theta$ and any $(s_i, s_j) \in Q$ with $P^S_\theta(s_i) > 0$ and $P^S_\theta(s_j) > 0$, we have for all $\alpha > 1$ and all $u \in \mathcal{U}$,*

$$\mathtt{D}_\alpha\Big(\Psi^u_\# \mathcal{M}^\theta_{s_i} \,\Big\|\, \Psi^u_\# \mathcal{M}^\theta_{s_j}\Big) \;\leq\; R_\alpha\big(\zeta, z^u_\infty(\theta, s_i, s_j)\big),$$

where $z^u_\infty(\theta, s_i, s_j) := W_\infty\Big(\Psi^u_\# P^{f,s_i}_\theta,\ \Psi^u_\# P^{f,s_j}_\theta\Big)$.

Lemma 4.1 states that, along each direction u, the per-slice Rényi envelope evaluated at the shift radius z^u_∞ upper-bounds the actual Rényi divergence between the projected output distributions under secrets s_i and s_j.

We then define the *averaged Rényi envelope* as

$$\mathtt{AR}^\infty_{\alpha,\omega}\big(\mathcal{M}^\theta_{s_i} \,\big\|\, \mathcal{M}^\theta_{s_j}\big) := \int_{\mathbb{S}^{d-1}} R_\alpha\big(\zeta, z^u_\infty\big)\, \mathrm{d}\omega(u). \tag{10}$$

This quantity aggregates the per-slice Rényi envelope bounds over slice $u \sim \omega$ and will serve as our sliced Rényi envelope

surrogate for the true averaged sliced Rényi divergence. Similarly, we define the *joint sliced Rényi envelope* by

$$\begin{aligned}&\mathtt{JR}^{\infty}_{\alpha,\omega}\big(\mathcal{M}^{\theta}_{s_i}\|\mathcal{M}^{\theta}_{s_j}\big)\\&:=\tfrac{1}{\alpha-1}\log\left(\int_{\mathbb{S}^{d-1}}\exp\big((\alpha-1)R_\alpha(\zeta,z^u_\infty)\big)d\omega(u)\right).\end{aligned}\tag{11}$$

Here, instead of averaging the per-slice envelope values linearly, we aggregate them via an exponential moment over $V\sim\omega$, followed by a log and $1/(\alpha-1)$ rescaling. This construction mirrors the passage from an average of Rényi divergences to a joint Rényi divergence for the pair $(V,\Psi^V(\mathcal{M}(X)))$. As a result, $\mathtt{JR}^{\infty}_{\alpha,\omega}$ is more sensitive to rare directions with large envelope values than $\mathtt{AR}^{\infty}_{\alpha,\omega}$.

Corollary 4.1. *Fix a continuous or finite slice profile $\{\mathcal{U},\omega\}$. Let $\alpha>1$ and $\varepsilon\geq 0$. For a query function f, let $\mathcal{M}(X)=f(X)+N$ with $N\sim\zeta$. Then, (i) $\mathcal{M}(X)$ satisfies $(\alpha,\varepsilon,\omega)$-Ave-SRPP if $\mathtt{AR}^{\infty}_{\alpha,\omega}\big(\mathcal{M}^{\theta}_{s_i}\,\|\,\mathcal{M}^{\theta}_{s_j}\big)\leq\varepsilon$; and (ii) $\mathcal{M}(X)$ satisfies $(\alpha,\varepsilon,\omega)$-Joint-SRPP if $\mathtt{JR}^{\infty}_{\alpha,\omega}\big(\mathcal{M}^{\theta}_{s_i}\,\|\,\mathcal{M}^{\theta}_{s_j}\big)\leq\varepsilon$.*

Corollary 4.1 follows immediately from Lemma 4.1. In particular, an ε-bounded Rényi envelope under (α,ω) is sufficient to guarantee $(\alpha,\varepsilon,\omega)$-SRPP, for both the average and joint variants. The envelopes $\mathtt{AR}^{\infty}_{\alpha,\omega}$ and $\mathtt{JR}^{\infty}_{\alpha,\omega}$ are especially useful operational targets: our Sliced Wasserstein Mechanism (SWM), calibrated via SW-sensitivity, provides a direct and tractable way to enforce $\mathtt{AR}^{\infty}_{\alpha,\omega}\big(\mathcal{M}^{\theta}_{s_i}\,\|\,\mathcal{M}^{\theta}_{s_j}\big)\leq\varepsilon$ and $\mathtt{JR}^{\infty}_{\alpha,\omega}\big(\mathcal{M}^{\theta}_{s_i}\,\|\,\mathcal{M}^{\theta}_{s_j}\big)\leq\varepsilon$ as we detail next.

4.3.1 Gaussian Sliced Wasserstein Mechanism

We specialize to Gaussian noise, $N\sim\mathcal{N}(0,\sigma^2I_d)$. Given a slice profile $(\mathcal{U},\omega)$ (continuous or finite) and the per-slice SW-sensitivity Δ^u_∞ in (6), define

$$\overline{\Delta}^2:=\int_{\mathbb{S}^{d-1}}\big(\Delta^u_\infty\big)^2\,\mathrm{d}\omega(u),\qquad \Delta^2_\star:=\sup_{u\in\mathcal{U}}\big(\Delta^u_\infty\big)^2.$$

Here, $\overline{\Delta}^2$ is the second moment of SW-sensitivity and $\Delta^2_\star$ is the largest squared directional sensitivity over $u\in\mathcal{U}$.

Theorem 4.5. *Fix a continuous or finite slice profile $\{\mathcal{U},\omega\}$. Let $\alpha>1$ and $\varepsilon\geq 0$. A mechanism $\mathcal{M}(X)=f(X)+N$ with $N\sim\zeta$ satisfies $(\alpha,\varepsilon,\omega)$-Ave-SRPP in $(\mathcal{S},\mathcal{Q},\Theta)$ if $\zeta=\mathcal{N}(0,\frac{\alpha\overline{\Delta}^2}{2\varepsilon}I_d)$.*

Theorem 4.5 provides a closed-form Gaussian noise calibration for Ave-SRPP under a fixed slice profile. Specifically, setting the isotropic variance to $\sigma^2=\frac{\alpha\overline{\Delta}^2}{2\varepsilon}$ ensures that the averaged Rényi envelope $\mathtt{AR}^{\infty}_{\alpha,\omega}$ is bounded by ε uniformly over $\theta\in\Theta$ and $(s_i,s_j)\in\mathcal{Q}$. By Corollary 4.1, this envelop bound is sufficient to guarantee $(\alpha,\varepsilon,\omega)$-Ave-SRPP. Moreover, the per-slice Rényi envelope is attained by the Gaussian SWM.

Theorem 4.6. *Fix a continuous or finite slice profile $\{\mathcal{U},\omega\}$. Let $\alpha>1$ and $\varepsilon\geq 0$. A mechanism $\mathcal{M}(X)=f(X)+N$ with $N\sim\zeta$ satisfies $(\alpha,\varepsilon,\omega)$-Joint-SRPP in $(\mathcal{S},\mathcal{Q},\Theta)$ if $\zeta=\mathcal{N}(0,\frac{\alpha\Delta^2_\star}{2\varepsilon}I_d)$.*

Theorem 4.6 gives the analogous closed-form noise calibration for Joint-SRPP. Taking $\sigma^2=\frac{\alpha\Delta^2_\star}{2\varepsilon}$ ensures that the joint Rényi envelope $\mathtt{JR}^{\infty}_{\alpha,\omega}$ is at most ε uniformly over $\theta\in\Theta$ and $(s_i,s_j)\in\mathcal{Q}$, and hence $\mathcal{M}$ satisfies $(\alpha,\varepsilon,\omega)$-Joint-SRPP by Corollary 4.1. Compared to the Gaussian SWM for the Ave-SRPP, this choice depends on the worst directional sensitivity $\Delta^2_\star$, rather than the profile average $\overline{\Delta}^2$, making it more conservative and typically requiring larger noise to protect against rare high-sensitivity directions.

4.3.2 Computational Properties of SWM

The SRPP guarantees of the SWM in Theorems 4.5 and 4.6 hold for both *continuous* slice profiles ω on $\mathbb{S}^{d-1}$ and *finite* profiles. In practice, even when ω is conceptually continuous, SWM is instantiated by sampling finitely many directions and the conditional distributions $P^{f,s}_\theta$ are typically accessed only through samples. This induces (i) a *finite-slice* (Monte Carlo) approximation of the profile aggregates and (ii) a *finite-sample* approximation of the per-slice sensitivities Δ^u_∞.

Notations. Fix a query function $f:\mathcal{X}\to\mathbb{R}^d$ and a Pufferfish scenario $(\mathcal{S},\mathcal{Q},\Theta)$. Fix finite instantiated families $\widehat{\Theta}\subseteq\Theta$ and $\widehat{\mathcal{Q}}\subseteq\mathcal{Q}$. For each $(s,\theta)\in\mathcal{S}\times\widehat{\Theta}$, draw n i.i.d. samples $Z^{(\theta,s)}_{1:n}\sim P^{f,s}_\theta$. For a direction u, write $V^{\theta,s}_k(u):=\langle Z^{(\theta,s)}_k,u\rangle$, and let $\widehat{F}^u_{\theta,s}$ denote the empirical CDF of $\{V^{\theta,s}_k(u)\}_{k=1}^n$, with generalized inverse $(\widehat{F})^{-1}(t):=\inf\{x:\widehat{F}(x)\geq t\}$. Let ω be a continuous slice profile on $\mathbb{S}^{d-1}$ and draw i.i.d. directions $u_1,\dots,u_m\sim\omega$. Define the empirical profile $\omega_m:=\frac{1}{m}\sum_{\ell=1}^m\delta_{u_\ell}$ and the profile aggregates $\overline{\Delta}^2(\omega):=\mathbb{E}_{U\sim\omega}\big[(\Delta^U_\infty)^2\big]$ and $\overline{\Delta}^2(\omega_m)=\frac{1}{m}\sum_{\ell=1}^m(\Delta^{u_\ell}_\infty)^2$.

For $\widetilde{\rho}\in(0,1)$, define $\mathsf{e}_n(\widetilde{\rho}):=\sqrt{\frac{\log(2/\widetilde{\rho})}{2n}}$ [2]. For $u\in\mathbb{S}^{d-1}$ and $(\theta,(s_i,s_j))\in\widehat{\Theta}\times\widehat{\mathcal{Q}}$, define

$$\begin{aligned}\widehat{W}^{\theta,s_i,s_j}_\infty(u,\widetilde{\rho}):=\sup_{t\in[\mathsf{e}_n(\widetilde{\rho}/2),\,1-\mathsf{e}_n(\widetilde{\rho}/2)]}\Big|&\big(\widehat{F}^u_{\theta,s_i}\big)^{-1}\big(t+\mathsf{e}_n(\widetilde{\rho}/2)\big)\\&-\big(\widehat{F}^u_{\theta,s_j}\big)^{-1}\big(t-\mathsf{e}_n(\widetilde{\rho}/2)\big)\Big|.\end{aligned}$$

For each sampled direction u_ℓ, define $\widehat{\Delta}_\infty(u_\ell):=\max_{(\theta,(s_i,s_j))\in\widehat{\Theta}\times\widehat{\mathcal{Q}}}\widehat{W}^{\theta,s_i,s_j}_\infty\big(u_\ell,\widetilde{\rho}\big)$.

Theorems 4.7 and 4.8 show that SWM admits PAC-style calibrations under *both* Monte Carlo slicing and finite-sample estimation of $P^{f,s}_\theta$, yielding $(\alpha,\varepsilon,\omega)$-Ave/Joint-SRPP guarantee with probability at least $1-\gamma$.

[2] $\sup_x|\widehat{F}(x)-F(x)|\leq\mathsf{e}_n(\widetilde{\rho})$ holds with probability at least $1-\widetilde{\rho}$ due to the Dvoretzky–Kiefer–Wolfowitz inequality.

Theorem 4.7. *Fix a Pufferfish scenario $(\mathcal{S}, Q, \Theta)$, a query $f : \mathcal{X} \to \mathbb{R}^d$, and a continuous slice distribution ω on $\mathbb{S}^{d-1}$. Draw i.i.d. directions $U_{1:m} \sim \omega$. Assume $0 \le \Delta_\infty^u \le \Delta_0$ for all u. Fix $\alpha > 1$, $\varepsilon \ge 0$, and $\gamma \in (0,1)$. Let $\widehat{\Delta}_\infty(U_\ell)$ be random quantities satisfying $\Pr\left(\Delta_\infty^{U_\ell} \le \widehat{\Delta}_\infty(U_\ell) \text{ for all } \ell \in [m]\right) \ge 1-\gamma/2$, where the probability is over the conditional sampling used to form $\widehat{\Delta}_\infty(U_\ell)$. If Gaussian SWM uses $N \sim \mathcal{N}(0, \sigma^2 I_d)$ with*

$$\sigma^2 \;\ge\; \frac{\alpha}{2\varepsilon}\left(\frac{1}{m}\sum_{\ell=1}^{m}\left(\widehat{\Delta}_\infty(U_\ell)\right)^2 + \Delta_0^2\sqrt{\frac{\log(4/\gamma)}{2m}}\right),$$

then with probability at least $1-\gamma$ (over both the conditional sampling and the draw of $U_{1:m}$), $\mathcal{M}(X) = f(X) + N$ satisfies $(\alpha, \varepsilon, \omega)$-Ave-SRPP.

Theorem 4.8. *Fix a Pufferfish scenario $(\mathcal{S}, Q, \Theta)$, a query $f : \mathcal{X} \to \mathbb{R}^d$, and a continuous slice distribution ω on $\mathbb{S}^{d-1}$. Draw i.i.d. directions $U_{1:m} \sim \omega$. Assume $0 \le \Delta_\infty^u \le \Delta_0$ for all u. Fix $\alpha > 1$, $\varepsilon \ge 0$, and $\gamma \in (0,1)$. Let $\widehat{\Delta}_\infty(U_\ell)$ be random quantities satisfying $\Pr\left(\Delta_\infty^{U_\ell} \le \widehat{\Delta}_\infty(U_\ell) \text{ for all } \ell \in [m]\right) \ge 1-\gamma/2$, where the probability is over the conditional sampling used to form $\widehat{\Delta}_\infty(U_\ell)$. For $\sigma^2 > 0$, define $c(\sigma) := \frac{\alpha(\alpha-1)}{2\sigma^2}$, $\widehat{\mu}_m(\sigma) := \frac{1}{m}\sum_{\ell=1}^{m}\exp\left(c(\sigma)\left(\widehat{\Delta}_\infty(U_\ell)\right)^2\right)$, and $b(\sigma) := \exp\left(c(\sigma)\Delta_0^2\right)$. Gaussian SWM uses $N \sim \mathcal{N}(0, \sigma^2 I_d)$ with σ^2 satisfying*

$$\frac{1}{\alpha-1}\log\left(\widehat{\mu}_m(\sigma) + \left(b(\sigma)-1\right)\sqrt{\frac{\log(4/\gamma)}{2m}}\right) \le \varepsilon, \qquad (12)$$

then with probability at least $1-\gamma$ (over both the conditional sampling and the draw of $U_{1:m}$), $\mathcal{M}(X) = f(X) + \mathcal{N}(0, \sigma^2 I_d)$ satisfies $(\alpha, \varepsilon, \omega)$-Joint-SRPP.

Since $c(\sigma) = \frac{\alpha(\alpha-1)}{(2\sigma^2)}$ is decreasing in σ, both $\widehat{\mu}_m(\sigma)$ and $b(\sigma)$ are decreasing in σ, hence the left-hand side of (12) is decreasing. Therefore the Joint-SRPP calibration condition $\Phi(\sigma) \le \varepsilon$ can be enforced by a 1-D bisection search over σ.

Finite ω. If ω is discrete on $\mathcal{U} = \{u_1, \ldots, u_m\}$ with weights $\{\omega_\ell\}$ (no Monte Carlo directions), for Ave-SRPP, we can replace $\frac{1}{m}\sum_{\ell=1}^{m}(\cdot)$ by $\sum_{\ell=1}^{m}\omega_\ell(\cdot)$ and drop the Hoeffding term $\Delta_0^2\sqrt{\log(4/\gamma)/(2m)}$. For Joint-SRPP, the same statement applies with $U_\ell = u_\ell$. Thus, only the finite-sample (distribution-estimation) layer remains.

Proposition 4.1 (Computational Advantage of SMD, Informal)**.** *With n samples per conditional distribution $P_\theta^{f,s}$, query dimension d, $\mathsf{M} := |\widehat{\Theta}|\,|\widehat{Q}|$ instances, and m slice directions, the runtime of computing SW-sensitivity is $O\big(\mathsf{M}\,m\,(dn + n\log n)\big)$. When the unsliced WM is computed by the standard discrete OT (Kantorovich) formulation on empirical measures, the runtime is $O\big(\mathsf{M}\,(dn^2 + n^3\log n)\big)$.*

Proposition 4.1 shows that slicing converts d-dimensional OT into m 1-D computations, reducing runtime from quadratic/cubic in n to nearly linear (up to sorting) (see App. P for the formal version). Beyond SWD, another mainstream scalable OT approximation is entropic OT via the Sinkhorn algorithm [5, 7, 24, 40]. App. A.5 describes a Sinkhorn–Wasserstein mechanism for RPP and summarizes why it is less suitable for privacy calibration and privatization.

4.4 Discussion on Slice Profile

In the SRPP framework, the privacy notion is parameterized by a slice profile $(\mathcal{U}, \omega)$, where $\mathcal{U} \subseteq \mathbb{S}^{d-1}$ is a set of projection directions and ω is a probability measure over those directions. To clarify the status of the profile and address how finite slicing affects the privacy guarantee, it is necessary to separate the slice profile's role in the *definition of the privacy notion* from its role in the *implementation of the privatization mechanism*. Depending on the operational context, a finite set of directions used during SWM calibration can be interpreted through two distinct modes.

Mode-I: Exact Finite-Profile SRPP. In this mode, the target guarantee is SRPP with a specified finite slice profile $(\mathcal{U}, \omega)$. The mechanism, e.g., SWM, uses the same $(\mathcal{U}, \omega)$. Hence, the resulting guarantee is an exact SRPP guarantee for the discrete distribution ω over $\mathcal{U}$, not an approximation to another slice profile. In particular, there is no finite-slicing approximation error relative to this finite-profile SRPP guarantee. Because the specified slice profile is part of the SRPP quantification, changing the finite profile changes the finite-profile SRPP guarantee being stated: it changes how the sliced Rényi divergence is evaluated and aggregated, not the mathematical form of the per-slice Rényi bound. Therefore, two finite-profile SRPP guarantees should not be interpreted as providing the same protection merely because their native numerical (α, ε) values are equal. Their protection strength should be compared through a common operational metric, such as the induced Bayes-optimal attack accuracy (e.g., Experiment 2 of Sec. 4.5). The same logic applies when the target guarantee uses a continuous slice profile and the corresponding SW-sensitivity can be computed exactly.

Mode-II: Continuous-Profile SRPP via Finite Estimation. In this mode, the target guarantee is SRPP with a continuous slice profile, such as uniform ω on $\mathbb{S}^{d-1}$. The finite directions used in the mechanism are not the slice profile in the SRPP definition. Instead, they are used to estimate the continuous-profile quantities needed for Gaussian SWM calibration. For Ave-SRPP, this quantity is the second moment $\overline{\Delta}^2(\omega) = \mathbb{E}_{U\sim\omega}[(\Delta_\infty^U)^2]$. For Joint-SRPP, it is the continuous-profile log-moment envelope appearing in the Joint-SRPP calibration, not the worst-direction quantity $\Delta_\star^2$ used in the closed-form fixed-profile bound. Therefore, the resulting guarantee must account for both Monte Carlo slicing error and finite-sample error in the estimated per-slice SW-sensitivities. These errors are handled by Theorems 4.7–4.8, which give high-probability SRPP guarantees for the target continuous slice

profile. In this mode, different finite slice sets are not different finite-profile SRPP guarantees; they are different finite-slice and finite-sample estimates of the same continuous-profile target. A smaller number of sampled directions typically yields looser PAC-style calibration bounds, so the sufficient Gaussian noise scale in these bounds can increase in order to obtain the target SRPP guarantee with high probability.

This two-mode distinction highlights that the same computational object, a finite set of projection directions, has different roles in different settings. In **Mode-I**, the finite directions are the slice profile used in the SRPP definition. In **Mode-II**, by contrast, the finite directions are only used to estimate the quantities needed for a continuous-profile SRPP guarantee. This is analogous to how different choices of the protected secret-pair set Q in a Pufferfish scenario specify different privacy requirements: changing Q changes the Pufferfish semantic requirement, whereas changing the slice profile changes the quantification of the SRPP guarantee, not the underlying Pufferfish semantics.

The Effect of Finite Slicing. In **Mode-I**, a finite slice profile affects the directional resolution of the SRPP guarantee, not its validity for that profile. Both continuous-profile and finite-profile SRPP mechanisms perturb the full output vector using isotropic Gaussian noise; the slice profile is used to compute the noise variance and to specify the projections over which the sliced Rényi divergence is evaluated. It does not leave any output coordinate unperturbed. Under the standard finite-sensitivity assumption (common in DP), for any direction $u \in \mathbb{S}^{d-1}$, whether or not $u \in \mathcal{U}$, the per-slice SW-sensitivity Δ_∞^u is finite. Once the finite-profile SRPP mechanism fixes the Gaussian noise variance using the chosen profile $(\mathcal{U}, \omega)$, it defines a single full-dimensional randomized release $\mathcal{M}(X) = f(X) + N$, regardless of which projection direction is later used for analysis. Hence, directions outside $\mathcal{U}$ do not create unperturbed or infinite-leakage channels; they are simply not included in the finite-profile SRPP aggregation. Thus, compared with continuous profiles, a finite slice profile yields lower directional resolution in the SRPP quantification: the guarantee evaluates and aggregates the per-direction Rényi envelope bounds only over the chosen slice profile $(\mathcal{U}, \omega)$.

Approximation and estimation in Pufferfish-family mechanisms. The finite-slice and finite-sample bounds above should be understood in the broader context of practical Pufferfish-family privatization. Mechanisms for PP/RPP also depend on prior-driven, distribution-dependent sensitivity quantities, which are rarely available in closed form outside simple cases. Although PP/RPP does not use slicing, estimating its Wasserstein sensitivity in high dimensions can suffer from severe sample complexity and the curse of dimensionality. SRPP does not eliminate these scenario-dependent estimation requirements; rather, it moves the relevant calibration problem into sliced, 1-D forms for which finite-sample and finite-slice errors can be explicitly controlled.

4.4.1 Guidance on Slice Profile Selection.

The choice of slice profile determines how sliced Rényi divergence is evaluated and aggregated in the SRPP guarantee.

Uniform Default. A uniform profile over the sphere is a natural default when there is no prior reason to weight some projection directions more than others in the SRPP quantification. It treats directions isotropically and aligns with common practice in the sliced Wasserstein distance (SWD) literature. Finite profiles drawn uniformly from the sphere provide a practical and scalable implementation of this principle.

Application-Specific Profiles. SRPP does not require the slice profile to be uniform. An application-specific profile means that the SRPP guarantee places more weight on projection directions that are independently identified as relevant for the released output. This does not change the protected secrets or the Pufferfish scenario; it changes how the sliced Rényi divergence is evaluated and aggregated. For example, domain knowledge or sensitivity analysis may indicate that certain coordinates, features, or linear combinations of outputs carry more secret-relevant variation, motivating a structured profile.

This is consistent with the SWD literature: although standard SWD uses uniform directions, non-uniform variants optimize, weight, or adapt projection distributions to problem geometry [8, 28]. In SRPP, the analogous choice is the slice profile used in the sliced Rényi quantification. This is also conceptually related to geometry-aware DP mechanisms, such as anisotropic Gaussian mechanisms [42], but the roles differ: DP geometry shapes the noise covariance, whereas SRPP geometry shapes the sliced divergence being evaluated. Such geometry must be justified independently of the privacy definition. A structured-profile SRPP guarantee remains valid SRPP, just as a properly calibrated anisotropic Gaussian mechanism remains valid DP. Choosing such a profile is therefore an application-dependent design choice, not a search for a universally best approximation to PP/RPP. It should balance the computational cost of evaluating the selected directions with the intended resolution of the SRPP quantification for the deployment. Developing adaptive or application-aware methods for slice-profile selection is therefore a natural and important direction for future work.

4.5 Example: SRPP vs. RPP

In this section, we provide a tractable numerical experiment to illustrate the difference between full-dimensional RPP and finite-profile SRPP under **Mode-I** (Sec. 4.4). We quantify the gap between their native certificates (i.e., the numerical values of privacy parameters) and their operational security (via Bayes-optimal attack accuracy). See App. W for details.

Setup. Consider a Pufferfish scenario $(\mathcal{S}, Q, \Theta)$, with $\mathcal{S} = \{s_1, s_2, s_3\}$, $Q = \{(s_i, s_j) : i \neq j\}$, and $\Theta = \{\theta^A, \theta^B\}$. The data domain contains tuples $x = (i, b, z)$, where $i \in \{1,2,3\}$, $b \in \{A, B\}$, and $z \in \mathcal{Z} := \{(0,0,0), (0.1, 0.1, 0.1)\}$. Define a query

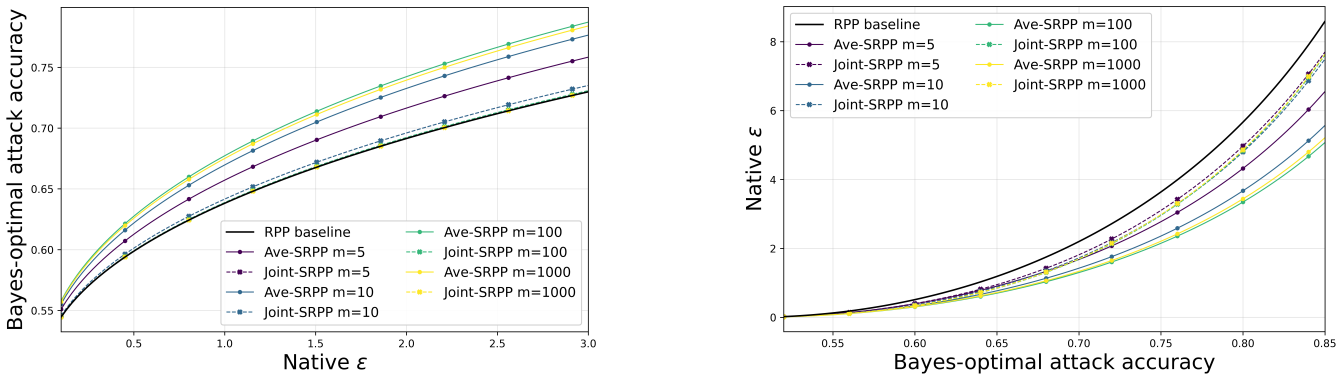


Figure 2: Comparison of RPP, Ave-SRPP, and Joint-SRPP under finite slice profiles: (Left) induced Bayes-optimal attack accuracy at the same native ε; (Right) native certificate values at the same Bayes-optimal attack accuracy.

function $f(x) := \mu_i^b + z \in \mathbb{R}^3$, where $\mu_1^A = e_1$, $\mu_2^A = e_2$, $\mu_3^A = e_3$, $\mu_1^B = e_1$, $\mu_2^B = \frac{1}{2}(e_1 + e_2)$, and $\mu_3^B = e_3$. Under prior θ^b, conditioning on secret s_i means that the data has the form $X = (i, b, Z)$, where Z is uniform on $\mathcal{Z}$. Hence, $P_{\theta^b}^{f,s_i}$ is the distribution of $\mu_i^b + Z$. We apply the Gaussian mechanism $\mathcal{M}(X) = f(X) + N$ with $N \sim \mathcal{N}(0, \sigma^2 I_3)$ and set $\alpha = 4$.

Sensitivities. The exact full-dimensional Wasserstein sensitivity for RPP is $\Delta_\infty^2 = \max_{b\in\{A,B\},(s_i,s_j)\in Q} \|\mu_i^b - \mu_j^b\|_2^2 = 2$. For SRPP, we evaluate nested finite slice profiles $\mathcal{U}_m \subset \mathbb{S}^2$ with uniform weights for $m \in \{5, 10, 100, 1000\}$, where $\mathcal{U}_m$ consists of the first m directions from a single size-1000 sample. For each direction u_ℓ, the per-slice SW-sensitivity is $\Delta_\infty^{u_\ell} = \max_{b\in\{A,B\},(s_i,s_j)\in Q} |\langle u_\ell, \mu_i^b - \mu_j^b\rangle|$. The Ave-SRPP and Joint-SRPP sensitivity quantities are $\overline{\Delta}^2(\mathcal{U}_m) = \frac{1}{m}\sum_{\ell=1}^m (\Delta_\infty^{u_\ell})^2$ and $\Delta_\star^2(\mathcal{U}_m) = \max_{\ell\in[m]} (\Delta_\infty^{u_\ell})^2$, respectively.

We conduct two reciprocal comparisons between RPP and SRPP:

Experiment 1: fixed certificate → protection strength. First, we fix the native privacy budget ε value across RPP, Ave-SRPP, and Joint-SRPP, calibrate their respective σ^2 values, using the closed-form worst-direction calibration for Joint-SRPP, and report the worst-pair Bayes-optimal attack accuracy, $\mathrm{Acc}^\star(\sigma^2) = \Phi(\sqrt{2}/(2\sigma))$. As shown in the left panel of Fig. 2, equal native numerical ε values at the same fixed α do not imply equal operational protection across RPP and finite-profile SRPP, because the certificates are defined using different quantification measures and, for SRPP, different slice profiles. Thus, finite-profile SRPP certificates should be interpreted relative to their specified slice profiles rather than compared to RPP solely by matching ε values.

Experiment 2: fixed protection strength → certificate gap. Second, we fix the target Bayes-optimal attack accuracy a, compute the physically required common noise variance $\sigma_a^2 = [2(\Phi^{-1}(a))^2]^{-1}$, and calculate the native certificate values assigned by each framework to this identical physical mechanism. Letting $t(a) = (\Phi^{-1}(a))^2$, the assigned certificates are: $\varepsilon_{\mathrm{RPP}}(a) = 8t(a)$, $\varepsilon_{\mathrm{Ave}}(a; \mathcal{U}_m) = 4\overline{\Delta}^2(\mathcal{U}_m)t(a)$, and $\varepsilon_{\mathrm{Joint}}(a; \mathcal{U}_m) = \frac{1}{3}\log\left[\frac{1}{m}\sum_{\ell=1}^m \exp\left(12(\Delta_\infty^{u_\ell})^2 t(a)\right)\right]$. The Joint-SRPP value $\varepsilon_{\mathrm{Joint}}(a; \mathcal{U}_m)$ is the tight log-mean-exp envelope; the closed-form Joint bound replaces it by the worst-direction upper bound (See App. W). As shown in the right panel of Fig. 2, once protection is fixed on a common operational scale, RPP and finite-profile SRPP assign different native certificate values to the same Gaussian mechanism. This illustrates that "weaker" or "stronger" protection is not determined by the native ε value alone when the underlying quantification measures differ; such comparisons require a common operational metric, such as Bayes-optimal attack accuracy.

5 SRPP-Learning with Gradient Clipping

For stochastic iterative algorithms such as SGD, directly bounding the end-to-end sensitivity of an entire training run is typically intractable under both DP and Pufferfish scenarios. Following the decompose–then–compose paradigm of DP-SGD with gradient clipping (Sec. A.2), we extend SRPP to stochastic learning by introducing SRPP-SGD with gradient clipping, an SRPP analogue of DP-SGD in the Pufferfish setting. We begin by fixing notation.

Dataset. Let $x = (x_1, \ldots, x_n) \in \mathcal{X} = \bar{\mathcal{X}}^n$ be the finite samples of data points, where each $x_i = (a_i, b_i) \in \bar{\mathcal{X}}$ consists of features a_i and label b_i. In this section, we take P_θ^X as the empirical (marginal) distribution of the data samples $x = (x_1, \ldots, x_n) \in \mathcal{X} = \bar{\mathcal{X}}^n$ for all $\theta \in \Theta$. At iteration t, we draw a random mini-batch $\mathsf{I}_t \subseteq [n]$ using a subsampling scheme $\{\eta, \rho\}$ characterized by a rate η ($\eta = \eta_t$ may vary with t) and a scheme type ρ, where $\rho \in \{\mathsf{WR}, \mathsf{WOR}, \mathsf{Poisson}\}$ denotes sampling with-replacement (WR), without-replacement (WOR), or via a Poisson process (Poisson), respectively. Let $X^t = (X_i)_{i\in\mathsf{I}_t}$ denote the random mini-batch subsampled from x, with x^t representing a particular realization.

Sampling randomness. Given a subsampling scheme $\{\eta, \rho\}$, let $(\mathcal{R}, \mathscr{R}, \mathbb{P}_{\eta,\rho})$ denote its probability space, with

Algorithm 1 SRPP-SGD with gradient clipping

Require: Data $x_1,\dots,x_n$; loss $\ell(\xi;x)$; subsampling scheme $\{\eta,\delta\}$; clip norm C; update maps $(T_t)_{t=1}^T$; noise covariances $(\Sigma_t)_{t=1}^T$.

1: Initialize parameters ξ_0; set history $y_{<1} \leftarrow \xi_0$
2: **for** $t = 1,\dots,T$ **do**
3: Sample minibatch index set $\mathsf{I}_t \subseteq [n]$ via scheme $\{\eta,\delta\}$; let $B_t \leftarrow |\mathsf{I}_t|$
4: For each $i \in \mathsf{I}_t$:

$$g_t(x_i) \leftarrow \nabla_\xi \ell(\xi_{t-1};x_i),$$
$$\tilde{g}_t(x_i) \leftarrow g_t(x_i)\cdot \min\left\{1,\ \frac{C}{\|g_t(x_i)\|_2}\right\}.$$

5: $\bar{g}_t \leftarrow \frac{1}{B_t}\sum_{i\in\mathsf{I}_t}\tilde{g}_t(x_i)$
6: Draw $N_t \sim \mathcal{N}(0,\Sigma_t)$ and set $\hat{g}_t \leftarrow \bar{g}_t + N_t$
7: $\xi_t \leftarrow T_t(\hat{g}_t;y_{<t})$
8: Update history $y_{<t+1}$
9: **end for**
10: **return** ξ_T; compute SRPP parameters from (Σ_t) via HUC/sa-HUC accountant.

$\mathbb{P}_{\eta,\rho}$ the corresponding probability measure. For a sampling draw $r \sim \mathbb{P}_{\eta,\rho}$, let $\mathsf{I}_t(r) \subseteq [n]$ denote the (multi)set of selected indices at iteration t and $B_t(r) := |\mathsf{I}_t(r)|$ its size. For $\rho \in \{\mathsf{WR},\mathsf{WOR}\}$, we have $B_t(r) \equiv B$ (deterministic). For Poisson subsampling ($\rho = \mathsf{Poisson}$; see App. E for details), $B_t(r)$ is random. We write $\mathcal{B} : \bar{\mathcal{X}}^n \times \mathcal{R} \to \bar{\mathcal{X}}^{B_t(r)}$ and $x^t = \mathcal{B}(x;r) := (x_i)_{i\in\mathsf{I}_t(r)}$ for the (mini-batch) subsampling operator.

Empirical Risk Minimization. In supervised learning, each example is $x_i = (a_i, b_i)$ with features a_i and label b_i. Let $\ell(\xi;x_i)$ denote the per-sample loss of a model with parameter ξ. The empirical risk minimization (ERM) problem is $\min_\xi F(\xi) := \frac{1}{n}\sum_{i=1}^n \ell(\xi;x_i)$.

(Clipped) SGD as a history-dependent query. At iteration t, a subsampling scheme $\{\eta,\rho\}$ draws randomness $r_t \sim \mathbb{P}_{\eta,\rho}$ and selects a minibatch $\mathcal{B}_t(x;r_t) = (x_i)_{i\in\mathsf{I}_t}$ of size B. Let $y_{<t} \in \mathcal{Y}_{<t}$ denote the history available at iteration t; for standard SGD, $y_{<t} = \xi_{t-1}$. Define the per-example gradients

$$g_t(x_i;\xi_{t-1}) := \nabla_\xi \ell(\xi_{t-1};x_i),$$

and the clipped batch gradient $\bar{g}_t(r_t) := \frac{1}{B}\sum_{i\in\mathsf{I}_t}\tilde{g}_t(x_i;\xi_{t-1})$, where $\tilde{g}_t(x_i)$ is the per-example clipped gradient (Line 4 of Algorithm 1). The pre-noise update (query) with step size $\kappa_t > 0$ is

$$f_t(x,y_{<t};r_t) := \xi_{t-1} - \kappa_t\,\bar{g}_t(r_t),$$

and we suppress r_t when unambiguous.

In DP-SGD, composition is key to the effectiveness of the DP privatization [1]. But PP/RPP generally lack graceful composition [21, 34, 38, 46] because conditioning on secrets via the prior can induce dependencies between the outputs of mechanisms that are otherwise independent [46]. The slicing process of SRPP does not circumvent such induced independence, and thus SRPP inherits the RPP's lack of proper composition properties. To address this and make SRPP applicable to stochastic learning, we introduce the *History-Uniform Cap*.

5.1 History-Uniform Cap

We define HUC for a fixed Pufferfish scenario $(\mathcal{S},\mathcal{Q},\Theta)$ and a slice profile $(\mathcal{U},\omega)$. For any $(s_i,s_j) \in \mathcal{Q}$ and $\theta \in \Theta$, let $\Pi(\mu^\theta_{s_i},\mu^\theta_{s_j})$ denote the set of couplings between $\mu^\theta_{s_i}$ and $\mu^\theta_{s_j}$ (Definition 2.2), where $\mu^\theta_{s_i}$ and $\mu^\theta_{s_j}$ are the distributions of X given $(S = s_i,\theta)$ and of X' given $(S = s_j,\theta)$, respectively.

Definition 5.1 (History-Uniform Cap (HUC))**.** *A vector $h_t = (h_{t,i})_{i=1}^m \in \mathbb{R}_+^m$ is a HUC for $\mathcal{U} = \{u_i\}_{i=1}^m$ at iteration t if, for all $\theta \in \Theta$, $(s_i,s_j) \in \mathcal{Q}$, history $y_{<t}$ in the support, and every coupling $\gamma \in \Pi(\mu^\theta_{s_i},\mu^\theta_{s_j})$, we have $\gamma\times\mathbb{P}_{\eta,\rho}$-a.s. in $((X,X'),R_t)$,*

$$\left|\langle f_t(X,y_{<t};R_t) - f_t(X',y_{<t};R_t),u_i\rangle\right| \le \sqrt{h_{t,i}},\forall i \in [m]. \quad (13)$$

Thus, a HUC $h_t = (h_{t,i})_{i=1}^m$ provides a *pathwise* bound that is uniform over histories: for every admissible coupling and every history $y_{<t}$ in support, the directional displacement along u_i is almost surely bounded by $\sqrt{h_{t,i}}$, and the constants $h_{t,i}$ do not depend on the particular value of $y_{<t}$. In matrix form, we may represent these caps by a PSD *HUC matrix* $H_t \succeq 0$ such that, $\gamma\times\mathbb{P}_{\eta,\rho}$-a.s. in $((X,X'),R_t)$,

$$\left|\left\langle f_t(X,y_{<t};R_t) - f_t(X',y_{<t};R_t),u\right\rangle\right|^2 \le u^\top H_t u, \quad \forall u \in \mathcal{U},$$

and in particular $h_{t,i} = u_i^\top H_t u_i$ for all $i \in [m]$. We view H_t as a deterministic, history-uniform sensitivity ellipsoid at iteration t, while h_t records its values on the slice set $\mathcal{U}$.

We next show that a finite HUC always exists. We start by introducing necessary notions.

Discrepancy cap. A general (probabilistic) secret–dataset relation may alter attributes, individual data points, or global properties of datasets input to algorithms as the secret varies. Under minibatch subsampling from a fixed finite dataset $x = (x_1,\dots,x_n)$, its effect on the iteration-t update is fully captured by *how many sampled points differ* between two coupled datasets.[3] For a sampling draw $r \sim \mathbb{P}_{\eta,\rho}$ with index (multi)set $\mathsf{I}_t(r)$ and size $B_t = |\mathsf{I}_t(r)|$[4], define the *discrepancy*

$$K_t(x,x';r) := \sum\nolimits_{j\in\mathsf{I}_t(r)} \mathbf{1}\{x_j \ne x'_j\}. \quad (14)$$

Given a coupling $\gamma \in \Pi(\mu^\theta_{s_i},\mu^\theta_{s_j})$, define the *iteration-$t$ discrepancy cap* as

$$K_t(\gamma) := \operatorname*{ess\,sup}_{((X,X'),R_t)\sim\gamma\times\mathbb{P}_{\eta,\rho}} K_t(X,X';R_t). \quad (15)$$

[3] App. G clarifies how this coupling-based viewpoint differs from *group* DP's worst-case k-neighbor model.

[4] We replace B_t by $B_t^{\max} := \operatorname{ess\,sup}_{R_t\sim\mathbb{P}_{\eta,\rho}} |\mathsf{I}_t(R_t)|$ if the size is random.

If a pair $(s_i, s_j) \in Q$ guarantees that X and X' differ in at most $\overline{K}$ coordinates γ-a.s., then $K_t(\gamma) \le \min\{\overline{K}, B_t\}$ at iteration t for the draw r. We say that a value K_t is a *feasible discrepancy cap* at iteration t (for the realized r) if $K_t \in [K_t(\gamma), B_t]$ for all $\theta \in \Theta$ and $(s_i, s_j) \in Q$. Intuitively, a feasible discrepancy cap provides a uniform upper bound on the number of differing sampled points that can arise at round t across all priors and admissible secret pairs. Operationally, estimating the discrepancy cap reduces to a counting problem: under a coupling of the two secret-conditioned data distributions, count how many minibatch records can differ when the protected secret is switched. The HUC cap controls this discrepancy count in a worst-case or tail sense. The subsampling-aware variant introduced below in Sec. 5.2 instead controls the corresponding second moment under minibatch sampling. App. D gives Monte Carlo and analytic estimators with finite-sample guarantees.

Update map and L_t-Lipschitz. At iteration t, we write the pre-perturbation update as a map $T_t : \mathbb{R}^d \times \mathcal{Y}_{<t} \to \mathbb{R}^d$ such that, for any realized sampling draw r,

$$f_t(x, y_{<t}; r) = T_t\big(\bar{g}_t(\mathcal{B}(x; r)); y_{<t}\big),$$

where $\bar{g}_t(\mathcal{B}(x;r)) \in \mathbb{R}^d$ is the clipped and averaged minibatch gradient (Line 5 of Algorithm 1). We say that T_t is *L_t-Lipschitz* in its first argument if there exists $L_t < \infty$ such that, for all $z, z' \in \mathbb{R}^d$ and all histories $y_{<t}$ in the support, $\big\|T_t(z; y_{<t}) - T_t(z'; y_{<t})\big\|_2 \le L_t \|z - z'\|_2$. App. C gives two common examples of L_t-Lipschitz update maps.

Assumption 5.2 (Slice-wise Lipschitz updates)**.** *For each iteration t and each direction $u_i \in \mathcal{U}$, there exists a finite constant $L_{t,i} < \infty$ such that, for all $z, z' \in \mathbb{R}^d$ and all histories $y_{<t}$ in the support, $\big|u_i^\top \big(T_t(z; y_{<t}) - T_t(z'; y_{<t})\big)\big| \le L_{t,i} \big|u_i^\top (z - z')\big|$.*

Assumption 5.2 enforces a directional non-expansiveness property: for each $u_i \in \mathcal{U}$, the u_i-component of the update cannot amplify the u_i-component of a gradient perturbation by more than $L_{t,i}$. This holds for standard SGD and for preconditioned updates with bounded linear preconditioners. Assumption 5.2 is a convenient sufficient condition for a fixed finite direction set $\mathcal{U}$. One may alternatively assume global ℓ_2-Lipschitzness with constant L_t of T_t, which implies Assumption 5.2 with $L_{t,i} \le L_t$ for all i, or impose other geometric constraints. App. F discusses structured (e.g., per-layer) clipping/perturbation variants that yield analogous slice-wise bounds with modified constants.

Proposition 5.1. *Fix a slice profile $(\mathcal{U}, \omega)$. Suppose per-example gradients are ℓ_2-clipped at threshold $C > 0$. Let K_t be a feasible discrepancy cap at iteration t, and let $B_t \ge 1$ denote the (deterministic) minibatch size. Suppose Assumption 5.2 holds. Then, the following is a valid HUC for each $u_i \in \mathcal{U}$,*

$$h_{t,i} := \Big(\frac{2K_t L_{t,i} C}{B_t}\Big)^2. \tag{16}$$

Proposition 5.1 provides an explicit worst-case, history-uniform bound on the directional displacement of the iteration-t update. If at most K_t sampled examples differ in a minibatch of size B_t, then the averaged clipped gradient can change by at most $\frac{2K_t C}{B_t}$ in any direction. Assumption 5.2 further implies that the u_i-component of the update can expand this change by at most a factor $L_{t,i}$. Thus, the directional shift along u_i is bounded by $\frac{2K_t L_{t,i} C}{B_t}$, and hence the squared shift is bounded by $(\frac{2K_t L_{t,i} C}{B_t})^2$, which matches the cap $h_{t,i}$ in (16). App. H characterizes HUC's *minimality* and *attainability*.

5.2 Subsampling-Aware HUC

Any feasible discrepancy cap K_t used in $h_{t,i}$ is a uniform worst-case upper bound on the random discrepancy $K_t(X, X'; R_t)$ under $\gamma \times \mathbb{P}_{\eta,\rho}$. While this worst-case viewpoint is sound, it can be overly conservative in SGD, where the subsampling randomness η is independent of the prior and has a genuine averaging effect. To better reflect this structure and improve utility, we introduce a *subsampling-aware HUC* (sa-HUC), which uniformly bounds the *expected* squared directional shifts over subsampling, while retaining worst-case uniformity over priors, secrets, and couplings.

Definition 5.3 (Subsampling-Aware History-Uniform Cap (sa-HUC))**.** *Fix t and a slicing set $\mathcal{U} = \{u_i\}_{i=1}^m \subset \mathbb{S}^{d-1}$. A vector $h_t^{\mathsf{sa}} = (h_{t,i}^{\mathsf{sa}})_{i=1}^m \in \mathbb{R}_+^m$ is a sa-HUC for $\mathcal{U}$ at iteration t if, for all $\theta \in \Theta$, all $(s_i, s_j) \in Q$, all histories $y_{<t}$ in the support, and every coupling $\gamma \in \Pi(\mu_{s_i}^\theta, \mu_{s_j}^\theta)$, we have, for γ-almost every (X, X'), all $i = 1, \dots, m$,*

$$\mathbb{E}_{\eta_t}\Big[\big|\langle f_t(X, y_{<t}; R_t) - f_t(X', y_{<t}; R_t), u_i\rangle\big|^2\Big] \le h_{t,i}^{\mathsf{sa}},$$

where the expectation is taken only over the subsampling randomness $R_t \sim \mathbb{P}_{\eta,\rho}$.

Let $\overline{K}_t^2$ be any scalar such that, for all $\theta \in \Theta$, $(s_i, s_j) \in Q$, all histories $y_{<t}$ in the support, and all couplings $\gamma \in \Pi(\mu_{s_i}^\theta, \mu_{s_j}^\theta)$, we have, for γ-almost every (X, X'),

$$\mathbb{E}_{\eta_t}\big[K_t(X, X'; R_t)^2\big] \le \overline{K}_t^2. \tag{17}$$

Proposition 5.2. *Fix an iteration t and a slicing set $\mathcal{U} = \{u_i\}_{i=1}^m \subset \mathbb{S}^{d-1}$. Assume:*

- *Per-example gradients are ℓ_2-clipped at radius $C > 0$.*
- *The minibatch size $B_t \ge 1$ is deterministic.*
- *Assumption 5.2 holds with constants $\{L_{t,i}\}_{i=1}^m$.*

For each $\theta \in \Theta$, $(s_i, s_j) \in Q$, history $y_{<t}$ in the support, and $\gamma \in \Pi(\mu_{s_i}^\theta, \mu_{s_j}^\theta)$, let $\overline{K}_t^2$ satisfy (17). Then the vector $h_t^{\mathsf{sa}} = (h_{t,i}^{\mathsf{sa}})_{i=1}^m$ with

$$h_{t,i}^{\mathsf{sa}} := \Big(\frac{2L_{t,i} C}{B_t}\Big)^2 \overline{K}_t^2, \qquad i = 1, \dots, m, \tag{18}$$

is a valid sa-HUC.

Proposition 5.2 shows that, under slice-wise Lipschitz updates and ℓ_2-clipping, the subsampling-aware discrepancy $\overline{K}_t^2$ directly yields a closed-form sa-HUC. In particular, if $\overline{K}_t^2$ uniformly bounds the expected squared discrepancy K_t^2 over subsampling, then the subsampling-aware directional shift along each slice u_i is controlled by $h_{t,i}^{\mathsf{sa}} = \left(\frac{2L_{t,i}C}{B_t}\right)^2 \overline{K}_t^2$. Thus h_t^{sa} is a valid sa-HUC.

To calibrate privacy using $\mathbb{E}_{\eta_t}[K_t^2]$ instead of a worst-case K_t, we now introduce subsampling-aware variants of Ave- and Joint-SRPP that match the sa-HUC notion. Let η denote all subsampling randomness used in SRPP-SGD, and for each fixed η, let $\mathcal{M}_s^{\theta,\eta}$ be the distribution of the mechanism's output under prior θ and secret s, conditional on η.

Definition 5.4 (Subsampling-Aware Average SRPP (sa-Ave-SRPP))**.** *Fix $\alpha > 1$ and a slice profile $(\mathcal{U}, \omega)$. A mechanism $\mathcal{M}$ with subsampling randomness η satisfies $(\alpha, \varepsilon, \omega)$-sa-Ave-SRPP if, for all $\theta \in \Theta$ and all $(s_i, s_j) \in \mathcal{Q}$,*

$$\mathbb{E}_\eta\left[\mathtt{ASD}_\alpha^\omega\left(\mathcal{M}_{s_i}^{\theta,\eta} \,\middle\|\, \mathcal{M}_{s_j}^{\theta,\eta}\right)\right] \le \varepsilon.$$

Here the expectation is taken only over η.

Compared to standard Ave-SRPP, sa-Ave-SRPP keeps the same sliced Rényi divergence and the same uniformity over priors and secrets, but relaxes the guarantee from a worst-case bound for each subsampling pattern η to an *on-average* bound over the subsampling randomness.

Definition 5.5 (Subsampling-Aware Joint SRPP (sa-Joint-SRPP))**.** *Under the same setup as Definition 5.4, a mechanism $\mathcal{M}$ satisfies $(\alpha, \varepsilon, \omega)$-sa-Joint-SRPP if, for all $\theta \in \Theta$ and all $(s_i, s_j) \in \mathcal{Q}$,*

$$\mathbb{E}_\eta\left[\mathtt{JSD}_\alpha^\omega\left(\mathcal{M}_{s_i}^{\theta,\eta} \,\middle\|\, \mathcal{M}_{s_j}^{\theta,\eta}\right)\right] \le \varepsilon.$$

Here, sa-Joint-SRPP is the subsampling-aware analogue of Joint-SRPP: it still treats the slice index and projection jointly, but now requires that the joint sliced Rényi divergence be small *on average over subsampling* rather than for every fixed realization of η. As with sa-Ave-SRPP, the only change from the standard notion is replacing a worst-case guarantee in η by an expectation over η.

5.3 HUC Accountant for Gaussian Noise

Similar to DP-SGD, an important issue of SRPP-SGD is computing the aggregated privacy cost of the entire training process, i.e., the *composition part* of SRPP-SGD as a decompose-then-compose privatization scheme. In this section, we set out to show how HUC can circumvent the RPP/SRPP's lack of graceful composition properties and enable a privacy *accountant* process to compute the privacy costs along the SGD-based training process. We refer to it as *HUC accountant*.

We start by introducing the Gaussian slice Rényi Envelope, which gives a per-iteration, per-slice RPP cost after noise perturbation.

Lemma 5.1. *Fix t and a slicing direction $u_i \in \mathbb{S}^{d-1}$. Let $N_t \sim \mathcal{N}(0, \Sigma_t)$ be independent of $((X, X'), R_t)$ and set $Y_t = f_t(X, y_{<t}; R_t) + N_t$. Define $v_{t,i} := u_i^\top \Sigma_t u_i$, and let $h_t = (h_{t,1}, \ldots, h_{t,m})$ be any HUC vector. Assume that $v_{t,i} > 0$. Then, for any $\alpha > 1$, we have*

$$\mathtt{D}_\alpha\left(\Pr\left(\langle Y_t, u_i\rangle \mid s_i, \theta\right) \middle\| \Pr\left(\langle Y_t, u_i\rangle \mid s_j, \theta\right)\right) \le \frac{\alpha}{2}\frac{h_{t,i}}{v_{t,i}}. \quad (19)$$

For ease of exposition, we absorb any step-size κ and preconditioning into the covariance Σ_t of the Gaussian noise. Lemma 5.1 gives a per-slice Rényi cost at iteration t directly in terms of the HUC component $h_{t,i}$ and the variance $v_{t,i}$ of the projected noise.

Theorem 5.6. *Fix a slicing profile $\{\mathcal{U}, \omega\}$. Let $h = \{h_1, \ldots, h_T\}$ with each $h_t = (h_{t,1}, \ldots, h_{t,m})$ as a* valid HUC *vector. Let $N_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 I_d)$. Then, for any $\alpha > 1$, we have:*

(i) **Ave-SRPP:** *For any $\sigma^2 \ge \frac{\alpha}{2\varepsilon}\sum_{t=1}^{T}\sum_{\ell=1}^{m}\omega_\ell h_{t,\ell}$, Algorithm 1 is $(\alpha, \varepsilon, \omega)$-Ave-SRPP.*

(ii) **Joint-SRPP:** *For any $\sigma^2 \ge \frac{\alpha}{2\varepsilon}\sum_{t=1}^{T}\max_\ell h_{t,\ell}$, Algorithm 1 is $(\alpha, \varepsilon, \omega)$-Joint-SRPP.*

Theorem 5.6 shows that once a valid sequence of $\{h_t\}_{t=1}^T$ has been constructed for a fixed slicing set $\mathcal{U}$, the overall SRPP guarantee for the entire SGD trajectory reduces to explicit, closed-form calibration conditions on the Gaussian noise. Both Ave-SRPP and Joint-SRPP are ensured by choosing a single variance parameter σ^2 large enough to offset the cumulative per-step HUC contributions, either in the ω-weighted average form (part (i)) or under the worst-slice aggregation (part (ii)).

Theorem 5.7. *Fix a slicing profile $\{\mathcal{U}, \omega\}$ with $\mathcal{U} = \{u_\ell\}_{\ell=1}^m$ and $\omega \in \Delta(\mathcal{U})$. Let $h^{\mathsf{sa}} = \{h_1^{\mathsf{sa}}, \ldots, h_T^{\mathsf{sa}}\}$ with each $h_t^{\mathsf{sa}} = (h_{t,1}^{\mathsf{sa}}, \ldots, h_{t,m}^{\mathsf{sa}})$ a* valid sa-HUC *vector. Let $N_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 I_d)$. Then, for any $\alpha > 1$, we have:*

(i) **sa-Ave-SRPP:** *If $\sigma^2 \ge \frac{\alpha}{2\varepsilon}\sum_{t=1}^{T}\sum_{\ell=1}^{m}\omega_\ell h_{t,\ell}^{\mathsf{sa}}$, then lgorithm 1 satisfies $(\alpha, \varepsilon, \omega)$-sa-Ave-SRPP.*

(ii) **sa-Joint-SRPP:** *If $\sigma^2 \ge \frac{\alpha}{2\varepsilon}\sum_{t=1}^{T}\max_\ell h_{t,\ell}^{\mathsf{sa}}$, then Algorithm 1 satisfies $(\alpha, \varepsilon, \omega)$-sa-Joint-SRPP.*

Theorem 5.7 is the subsampling-aware analogue of Theorem 5.6. Once we have a sequence of sa-HUC caps h_t^{sa} that bound the *average* (over subsampling) squared directional shifts at each iteration, the overall sa-SRPP guarantee for the full SGD trajectory reduces to the same kind of closed-form calibration on a single noise level σ^2. Part (i) ensures an $(\alpha, \varepsilon, \omega)$ sa-Ave-SRPP bound by matching the ω-weighted sum of sa-HUCs, while part (ii) does the same for

sa-Joint-SRPP using the worst-slice aggregation. Compared to the standard HUC case, the structure of the accountant is unchanged; the only difference is that the guarantee is now on-average over subsampling randomness. In App. F, we characterize SGD with per-layer gradient clipping [26], where we impose an assumption of layer-wise Lipschitz updates instead of Assumption 5.2.

Proposition 5.3 (Utility of sa-SRPP, informal)**.** *Under mild conditions, subsampling-aware discrepancy caps yield more utility than worst-case caps at the same true SRPP of the mechanism.* (see Proposition J.1 in App. J for formal results)

Proposition 5.3 confirms that averaging over the subsampling randomness improves utility at the same true SRPP. App. J provides formal and detailed analysis.

5.4 Composition of SRPP-SGD Mechanisms

In this section, we show that the HUC- and sa-HUC-based SRPP/sa-SRPP-SGD mechanisms developed above admit additive composition under a shared slice profile. This result should not be interpreted as a general composition theorem for arbitrary SRPP mechanisms. As in PP/RPP, arbitrary mechanisms may fail to compose gracefully because conditioning on Pufferfish secrets can induce dependencies across releases. The result below applies to the structured SRPP-SGD mechanisms considered in this section, where the per-iteration secret-induced shifts are controlled by HUC or sa-HUC sequences and the mechanisms are calibrated under the same slice profile. For clarity, we present the setup and results in terms of HUC and SRPP; the same arguments apply to sa-HUC and sa-SRPP after replacing the corresponding notions.

We consider J SRPP-SGD mechanisms $\mathcal{M}_1, \dots, \mathcal{M}_J$ acting on the same dataset X. Without loss of generality, we take the primitive output of each mechanism to be a vector in a common parameter space $\mathbb{R}^d$ (e.g., a final clipped update or parameter vector), and we fix a shared slice profile $(\mathcal{U}, \omega)$ with $\mathcal{U} = \{u_\ell\}_{\ell=1}^m \subset \mathbb{S}^{d-1}$ and $\omega \in \Delta(\mathcal{U})$. In this section, we restrict to the following mechanisms: for each $j \in \{1, \dots, J\}$, the run $\mathcal{M}_j : \mathcal{X} \to \mathbb{R}^d$ is an SRPP-SGD mechanism with iterates $\theta_t^{(j)} \in \mathbb{R}^d$ $(t = 0, \dots, T_j)$, additive Gaussian noises $N_t^{(j)} \sim \mathcal{N}(0, \Sigma_t^{(j)})$, and an associated HUC (or sa-HUC) sequence $h^{(j)} = \{h_t^{(j)}\}_{t=1}^{T_j}$ that satisfies the conditions of Theorem 5.6 (resp. Theorem 5.7). The composition is

$$\vec{\mathcal{M}}(X) := \big(\mathcal{M}_1(X), \dots, \mathcal{M}_J(X)\big) \in (\mathbb{R}^d)^J. \tag{20}$$

Theorem 5.8. *Fix a slice profile $\{\mathcal{U}, \omega\}$ and a Pufferfish scenario $\{\mathcal{S}, \mathcal{Q}, \Theta\}$. Let $\vec{\mathcal{M}}$ be the composition in (20). Fix any $\Xi \in \{\text{Ave}, \text{Joint}, \text{sa-Ave}, \text{sa-Joint}\}$. Suppose that, for each $j \in \{1, \dots, J\}$, $\mathcal{M}_j$ is an HUC-based or sa-HUC-based SRPP-SGD mechanism satisfying the corresponding assumptions of Theorem 5.6 or Theorem 5.7, and suppose that $\mathcal{M}_j$ is $(\alpha, \varepsilon_j, \omega)$-$\Xi$-SRPP under the shared slice profile $\{\mathcal{U}, \omega\}$. Then the composed mechanism $\vec{\mathcal{M}}$ is $(\alpha, \sum_{j=1}^J \varepsilon_j, \omega)$-$\Xi$-SRPP.*

In other words, for any fixed SRPP notion (Ave, Joint, or their subsampling-aware counterparts), the privacy costs of J SRPP-SGD mechanisms acting on the same data and slice profile add up linearly: the composed mechanism has the same α and ω, with total privacy cost $\sum_j \varepsilon_j$.

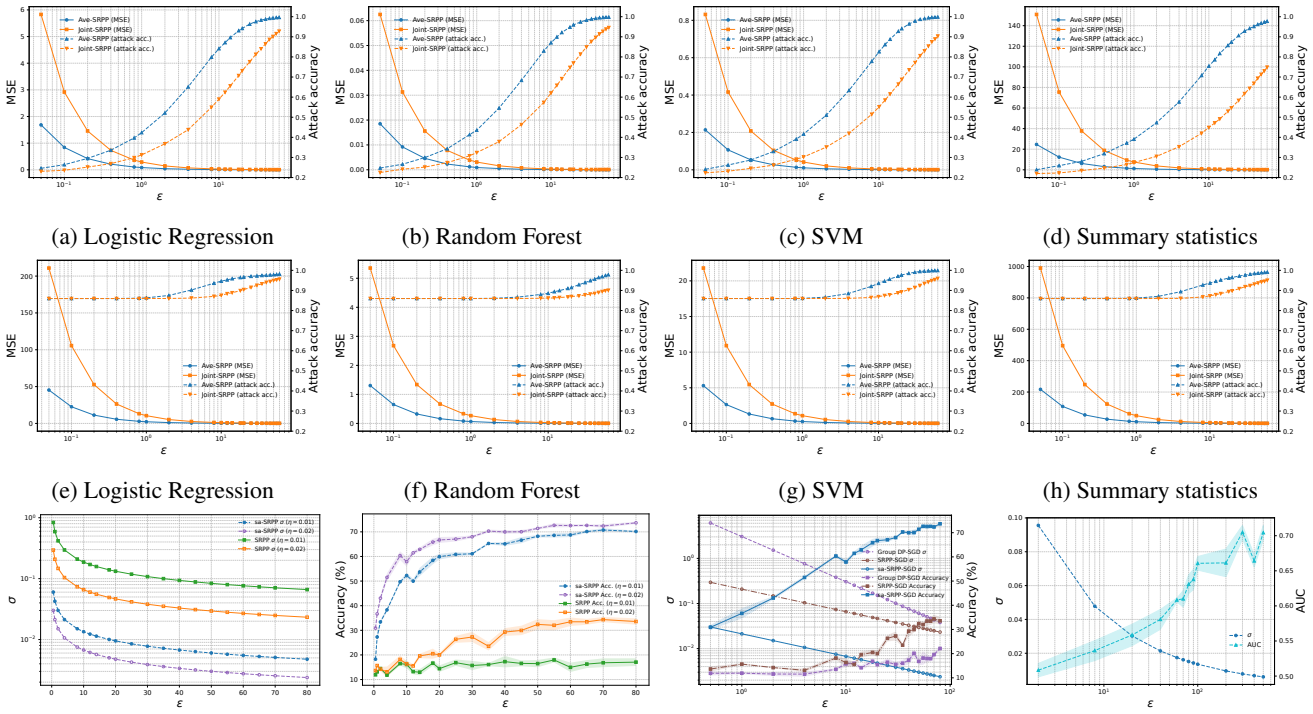
6 Experiments

We empirically evaluate our Gaussian sliced Wasserstein mechanisms in both the static and stochastic-learning settings. For static queries, we instantiate SRPP guarantees (Section 4), and for iterative training we use SRPP/sa-SRPP together with gradient clipping to privatize SGD (Section 5). Full experimental details and additional plots are in App. X.

6.1 Gaussian Sliced Wasserstein Mechanisms

Dataset and Pufferfish scenario. We evaluate our sliced Wasserstein mechanisms on three standard tabular benchmarks in an attribute-privacy setting [49]: Adult Census [2] (race as secret, $k = 5$), Cleveland Heart Disease [9], and Student Performance [36], with Pufferfish query set $\mathcal{Q} = \{(s_i, s_j) : s_i \neq s_j\}$. We present results for Adult and App. X shows results for the other datasets. For Adult, we consider both the natural imbalanced race prior (majority race probability 0.86) and a balanced, stratified subsample with a uniform 0.20 prior over the five race categories. We study two query families: (i) low-dimensional summary statistics (means, standard deviations, and rates of selected features) conditioned on each secret value; and (ii) machine learning (logistic regression, random forest, and SVM) queries on Adult, trained separately per secret group. In all experiments, we use Gaussian mechanisms $Y = \mathcal{M}(X) + N$, where $N \sim \mathcal{N}(0, I_d \sigma)$ with σ calibrated via our sliced SRPP. Utility is measured by mean squared error and related norms between $\mathcal{M}(X)$ and Y. We also use a prior-aware Gaussian MAP attack on the secret attribute to perform empirical privacy auditing, reporting accuracy and advantage over the prior baseline (i.e., 0.86 for imbalanced Adult, 0.20 for balanced Adult).

Figures 3a-3d and 3e-3h show utility-privacy trade-off as a function of the SRPP budget ε on balanced and imbalanced Adult datasets for four query types. Across all four queries, the behavior is consistent for both balanced and imbalanced settings. For very strong privacy (small ε), both Ave- and Joint-SRPP force high noise and the MAP attack accuracy stays close to the prior baselines (i.e., 0.2 and 0.86), indicating that the mechanism reveals little additional information about the race. As ε increases, MSE decreases toward zero while attack accuracy rises monotonically, reflecting the expected utility-privacy trade-off. In addition, at any fixed ε, Joint-SRPP is more conservative than Ave-SRPP: it induces larger MSE but



(a) Logistic Regression (b) Random Forest (c) SVM (d) Summary statistics

(e) Logistic Regression (f) Random Forest (g) SVM (h) Summary statistics

(i) ResNet: σ (j) ResNet: Accuracy (k) SRPP vs. Group DP (l) Empirical Attack Auditing

Figure 3: Privacy–utility tradeoffs with the Gaussian sliced Wasserstein mechanism. **(a)–(d)** ($\alpha = 4$) Ave-/Joint-SRPP on balanced Adult (uniform 0.2 prior) for four query families: logistic regression, SVM, random forest feature importances, and summary statistics; each panel plots utility and attack accuracy vs. the SRPP budget ε. **(e)–(h)** ($\alpha = 4$) the same on imbalanced Adult (0.86 prior). **(i)–(l)** use $\alpha = 16$ on CIFAR-10. **(i)** calibrated per-iteration noise σ vs. ε for SRPP-SGD and sa-SRPP-SGD at $\eta \in 0.01, 0.02$. **(j)** corresponding test accuracy vs. ε. **(k)** σ and test accuracy for group-DP-SGD, SRPP-SGD, and sa-SRPP-SGD under matched clipping, subsampling, and model. **(l)** overfitted SRPP-SGD ($\eta = 0.02$): empirical MIA attack AUC vs. the sa-SRPP budget ε.

yields uniformly lower attacker accuracy. This gap is most visible in the moderate privacy regime, where the Ave-SRPP curves already allow the attacker to reach substantially higher accuracy, while Joint-SRPP keeps the attack much closer to the prior baseline.

6.2 SRPP-SGD Using Gaussian Noise

We evaluate our SRPP-SGD and sa-SRPP-SGD mechanisms on image classification with Gaussian noise added to clipped gradients in a DP-SGD-style training pipeline. In our implementation, the HUC envelope is instantiated using a *global* ℓ_2-Lipschitz bound on the (clipped, averaged) gradient update map. Specifically, for each iteration t we bound the change in the update under neighboring worlds by a scalar constant L_t, so the global sensitivity of the update in Euclidean norm is at most L_t. This global envelope automatically upper-bounds all 1-D projections and therefore induces valid SRPP guarantees for any slicing profile $(\mathcal{U}, \omega)$ via the linear-update setting. Hence, Ave- and Joint-SRPP-SGD would have the same HUC. App. X.2 provides more details.

Dataset and Pufferfish scenario. We use CIFAR-10 (50,000 training and 10,000 test images). We create a two-world Pufferfish scenario, where `cat` is designated as the secret class, and construct two worlds s_0 and s_1 that differ only in the prevalence of this class: in the prior, the `cat` frequency shifts from $p_{\text{low}} = 0.10000$ to $p_{\text{high}} = 0.10004$. The realized world s_0 is then fixed and encoded by overwriting the CIFAR-10 training labels with the realized label vector; all training procedures (private and non-private) operate only on this realized world.

Models. Our main model is a ResNet-22 similar to [42], implemented with ELU activations, GroupNorm-based layers, a 3×3 adaptive average pooling, and a linear classifier on a standardized $64 \times 3 \times 3$ feature vector. The architecture is made differentially private–compatible (ready for Opacus [45]) by replacing non–DP-friendly components (such as non-picklable shortcuts) with equivalent modules.

SRPP-SGD and sa-SRPP-SGD training. Training follows the standard DP-SGD template: at each iteration, per-sample gradients are clipped to an ℓ_2 bound C and then perturbed by isotropic Gaussian noise. We use a worst-case dis-

crepancy cap K_{cap} for SRPP-SGD and use $\mathbb{E}_\eta[K_t^2]$ for sa-SRPP-SGD. We pre-compute, for each target SRPP and sa-SRPP budget ε at a fixed Rényi order $\alpha = 16$, a per-step noise scale $\sigma(\varepsilon)$ that ensures the resulting Gaussian mechanisms satisfies the SRPP- and sa-SRPP-SGD under the given Pufferfish scenario. In addition, we include a diagnostic group-DP-SGD curve under the same experimental setup, differing only in the noise calibration. For this curve, we follow the standard DP-SGD calibration [1] and convert it to a group-DP-style calibration using an effective group-size proxy derived from the sa-HUC mean-square discrepancy cap $\overline{K}_t^2$, namely $k_{\text{eff}} = \lceil \sqrt{\overline{K}_t^2} \rceil$; see App. G for details.

Overfitted regime and empirical privacy attacker. In addition to the above accuracy-focused experiments on the full training set, we run a separate overfitted regime tailored to empirical privacy evaluation. We restrict the training data to a small random subset of CIFAR-10 (on the order of 10^3 images), substantially increase the number of epochs, and remove or reduce weight decay, so that the resulting private SRPP-SGD models exhibit markedly higher training accuracy than test accuracy. On these overfitted models, we apply the loss-threshold membership inference attack of Yeom et al. [44]: for each trained model, we draw disjoint sets of member and non-member points, compute per-example cross-entropy losses, and use the negative loss as a scalar membership score (lower loss $\Rightarrow$ more likely member). We report the AUC of this score as a function of ε and compare against a non-private overfitted baseline trained on the same subset.

As shown in Fig. 3i–3j, across all ε and for both step sizes, sa-SRPP consistently requires much smaller Gaussian noise than worst-case SRPP: the sa-SRPP σ curves lie well below their SRPP counterparts, often by a factor of several and approaching an order of magnitude around $\varepsilon \approx 20$–40. This tighter calibration directly translates into higher utility. For small budgets both methods yield low accuracy, but as ε increases the sa-SRPP models continue to improve, reaching roughly 60%–70% test accuracy at the largest budgets, whereas SRPP saturates around 20%–35%. The same pattern holds for both $\eta = 0.01$ and $\eta = 0.02$: sa-SRPP dominates SRPP in terms of the accuracy–privacy tradeoff, demonstrating that replacing a worst-case discrepancy cap K_{cap} by the subsampling-aware cap $\mathbb{E}_\eta[K_t^2]$ yields substantially milder noise for the same ε.

Although group-DP, SRPP, and sa-SRPP are different privacy notions with different semantic goals, Fig. 3k reports a diagnostic group-DP-style calibration applied to the same intended Pufferfish secret. Specifically, we run DP-SGD with an effective group-size proxy derived from the sa-HUC mean-square discrepancy cap $\overline{K}_t^2$; see App. G for the calibration and for the distinction between this surrogate and the SRPP/sa-SRPP. This curve should not be read as a direct baseline for SRPP-SGD or sa-SRPP-SGD, since it corresponds to a different formal privacy guarantee. Rather, it is a diagnostic surrogate showing what happens when the Pufferfish secret-level dependency is translated into an effective group-size parameter for DP-SGD. In this setting, that translation yields conservative noise levels across the range of ε. The SRPP/sa-SRPP calibration developed in Sec. 5 instead keeps the calibration within the Pufferfish framework by using HUC/sa-HUC discrepancy caps rather than a group-size surrogate.

In the overfitted setting, the calibrated noise scale σ exhibits the same privacy–utility behavior as in Fig. 3i: tighter sa-SRPP budgets (smaller ε) require larger σ. Fig. 3l shows the corresponding stress-test of the empirical membership-inference attack. Tighter budgets (larger noise) yield weaker membership inference, while looser budgets reduce the noise at the cost of stronger membership-inference signals.

7 Conclusion

We introduced *Sliced Rényi Pufferfish Privacy* (SRPP), a PP/RPP-semantics-preserving framework that replaces high-dimensional Rényi divergence with projection-based quantification via *Average*- and *Joint*-SRPP. To mitigate the curse of dimensionality in PP-family mechanism design, we developed *sliced Wasserstein mechanisms* calibrated by a computable sliced Wasserstein sensitivity, yielding provable SRPP guarantees and closed-form Gaussian noise calibration without incurring the computational burden of full optimal transport. We further provided finite-sample characterizations, including conditions under which Gaussian noise achieves SRPP with high confidence.

To enable practical privacy accounting for iterative learning under probabilistic secret–dataset relationships, we proposed SRPP-SGD with gradient clipping and new accountants based on *History-Uniform Caps* (HUC) and a subsampling-aware variant (sa-HUC) that improves utility. This supports a decompose-then-compose privatization strategy without requiring Wasserstein sensitivity computation or contractivity assumptions. Under a shared slicing profile, we proved graceful additive composition for Ave/Joint SRPP and their subsampling-aware counterparts, enabling modular accounting for multi-stage pipelines and cascades. Empirically, our experiments on both static privatization and SRPP-SGD demonstrate favorable privacy–utility trade-offs and scalability.

Ethical Considerations

We propose Sliced Rényi Pufferfish Privacy (SRPP) together with practical sliced mechanisms (e.g., Ave-/Joint-SRPP and SRPP-SGD) to make Pufferfish-style, scenario-based privacy more tractable to certify and deploy in high dimensions while preserving the intended semantics. Our contributions are methodological: we do not collect new data, do not scrape data, do not interact with production systems or APIs, and do not discover or disclose vulnerabilities. We organize ethical

considerations along the generic pipeline

$$\textbf{Data} \rightarrow \textbf{Mechanisms} \rightarrow \textbf{Downstream Decision-Making},$$

and focus on the benefits and risks of releasing privatized query answers or trained models in a trusted-curator setting.

Stakeholders

Data subjects and curators. Although our experiments use public benchmarks, SRPP-like mechanisms may be applied to sensitive real-world datasets (e.g., health, education, finance). Relevant stakeholders include data subjects, dataset curators, and data stewards responsible for governance and consent.
Practitioners and auditors. Data scientists, ML engineers, privacy/security teams, and external auditors may rely on SRPP guarantees to decide whether (and how) to release statistics or models under a specified Pufferfish scenario $(\mathcal{S}, Q, \Theta)$.
Affected individuals and oversight bodies. People impacted by downstream decisions (e.g., lending, admissions, clinical risk scoring) and regulators/standards bodies that interpret privacy claims and assess compliance.

Potential Harms and Mitigations

Misinterpretation and privacy-washing. SRPP is scenario-based: its protection depends on the declared secrets $\mathcal{S}$, discriminative pairs Q, and admissible worlds Θ. A key harm is misrepresenting a scenario-specific guarantee as universal privacy, or obscuring the assumptions behind the guarantee (e.g., which secret pairs are protected). We mitigate this by clearly stating the threat model and emphasizing that deployments must document the chosen scenario, parameters, and scope, and should be reviewed by privacy/security stakeholders.
Utility degradation and downstream harms. As with other rigorous privacy frameworks, achieving stronger privacy typically requires adding noise, which can reduce accuracy and may cause harmful errors in high-stakes contexts. Our work does not deploy SRPP in real decision pipelines; we study trade-offs on benchmarks. For real deployments, we recommend conservative parameter choices when harms are high, explicit evaluation of downstream impact (including error costs and subgroup effects when appropriate), and governance processes that do not treat formal bounds as the sole determinant of acceptability.
Implementation and approximation pitfalls. Practical instantiations may involve Monte Carlo estimation, finite slice profiles, numerical optimization, or other approximations. Under-estimating sensitivity surrogates or using unstable solvers can lead to insufficient noise and invalid guarantees. We mitigate this risk by highlighting approximation failure modes, recommending conservative numerical tolerances and sanity checks, and encouraging reproducible implementations and independent auditing for high-stakes deployments.

Decision to Conduct and Publish

We believe this work has positive expected impact by making scenario-based privacy more practical in high dimensions and by promoting transparent, auditable privacy specifications tied to explicit assumptions. At the same time, SRPP does not eliminate privacy risk and should not be used to justify aggressive data use without clear communication of scope and limits, careful parameter selection, and appropriate oversight. Deployment decisions remain context-dependent.

Open Science

Our artifacts consist of source code archived on Zenodo: https://zenodo.org/records/20433043. The repository includes implementations and reproduction scripts, with public benchmarks downloaded automatically from official sources. Experiments used Python 3.x and PyTorch on a single NVIDIA A40 48 GB GPU.

Acknowledgements

This work was partially supported by the NSF (IIS-2214141, ITE-2452834), ARO (W911NF-25-1-0059), ONR (N000142412663), and Amazon.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, page 308–318, 2016.

[2] Barry Becker and Ronny Kohavi. Adult. *UCI Machine Learning Repository*, 10:C5XW20, 1996.

[3] Clément Bonet, Lucas Drumetz, and Nicolas Courty. Sliced-Wasserstein Distances and Flows on Cartan-Hadamard Manifolds. *Journal of Machine Learning Research*, 26(32):1–76, 2025.

[4] Michelle Chen and Olga Ohrimenko. Protecting global properties of datasets with distribution privacy mechanisms. In *International Conference on Artificial Intelligence and Statistics*, pages 7472–7491. PMLR, 2023.

[5] Lenaic Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré. Faster wasserstein distance estimation with the sinkhorn divergence. *Advances in neural information processing systems*, 33:2257–2269, 2020.

[6] Harald Cramér and Herman Wold. Some theorems on distribution functions. *Journal of the London Mathematical Society*, 1(4):290–294, 1936.

[7] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[8] Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander G Schwing. Max-sliced wasserstein distance and its use for gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10648–10656, 2019.

[9] Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H Guppy, Stella Lee, and Victor Froelicher. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American journal of cardiology*, 64(5):304–310, 1989.

[10] Ni Ding. Kantorovich mechanism for pufferfish privacy. In *International Conference on Artificial Intelligence and Statistics*, pages 5084–5103. PMLR, 2022.

[11] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual international conference on the theory and applications of cryptographic techniques*, pages 486–503. Springer, 2006.

[12] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[13] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[14] Vitaly Feldman, Ilya Mironov, Kunal Talwar, and Abhradeep Thakurta. Privacy amplification by iteration. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 521–532. IEEE, 2018.

[15] Natasha Fernandes, Mark Dras, and Annabelle McIver. Generalised differential privacy for text document processing. In *International Conference on Principles of Security and Trust*, pages 123–148. Springer International Publishing Cham, 2019.

[16] Moritz Hardt and Aaron Roth. Beyond worst-case analysis in private singular vector computation. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 331–340, 2013.

[17] Xi He, Ashwin Machanavajjhala, and Bolin Ding. Blowfish privacy: Tuning privacy-utility trade-offs using policies. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, page 1447–1458. Association for Computing Machinery, 2014.

[18] Sushil Jajodia, Pierangela Samarati, and Moti Yung. *Encyclopedia of Cryptography, Security and Privacy*. Springer, 2025.

[19] Shiva P Kasiviswanathan and Adam Smith. On the'semantics' of differential privacy: A bayesian formulation. *Journal of Privacy and Confidentiality*, 6(1), 2014.

[20] Stephan Kessler, Erik Buchmann, and Klemens Böhm. Deploying and evaluating pufferfish privacy for smart meter data. In *IEEE 12th Int. Conf. Ubiquitous Intell. Comput. IEEE 12th Int. Conf. Autonomic Trusted Comput. IEEE 15th Int. Conf. Scalable Comput. Commun. Associated Workshops (UIC-ATC-ScalCom)*, pages 229–238, 2015.

[21] Daniel Kifer and Ashwin Machanavajjhala. Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems (TODS)*, 39(1):1–36, 2014.

[22] Philip A Knight. The sinkhorn–knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008.

[23] Ios Kotsogiannis, Stelios Doudalis, Sam Haney, Ashwin Machanavajjhala, and Sharad Mehrotra. One-sided differential privacy. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 493–504. IEEE, 2020.

[24] Giulia Luise, Alessandro Rudi, Massimiliano Pontil, and Carlo Ciliberto. Differential properties of sinkhorn approximation for learning with wasserstein distance. *Advances in Neural Information Processing Systems*, 31, 2018.

[25] Pascal Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pages 1269–1283, 1990.

[26] H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *International Conference on Learning Representations*, 2018.

[27] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275, 2017.

[28] Khai Nguyen, Nhat Ho, Tung Pham, and Hung Bui. Distributional sliced-wasserstein and applications to generative modeling. *arXiv preprint arXiv:2002.07367*, 2020.

[29] Sloan Nietert, Ziv Goldfeld, Ritwik Sadhu, and Kengo Kato. Statistical, robustness, and computational guarantees for sliced wasserstein distances. *Advances in Neural Information Processing Systems*, 35:28179–28193, 2022.

[30] Chaoyue Niu, Zhenzhe Zheng, Shaojie Tang, Xiaofeng Gao, and Fan Wu. Making big money from small sensors: Trading time-series data under pufferfish privacy. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pages 568–576. IEEE, 2019.

[31] Theshani Nuradha and Ziv Goldfeld. Pufferfish privacy: An information-theoretic study. *IEEE Transactions on Information Theory*, 69(11):7336–7356, 2023.

[32] Nicolas Papernot and Thomas Steinke. Hyperparameter tuning with renyi differential privacy. *arXiv preprint arXiv:2110.03620*, 2021.

[33] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[34] Clément Pierquin, Aurélien Bellet, Marc Tommasi, and Matthieu Boussard. Rényi pufferfish privacy: General additive noise mechanisms and privacy amplification by iteration via shift reduction lemmas. In *International Conference on Machine Learning (ICML 2024)*, 2024.

[35] Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2011.

[36] Alice Silva. Using data mining to predict secondary school student performance. 2008.

[37] Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE global conference on signal and information processing*, pages 245–248. IEEE, 2013.

[38] Shuang Song, Yizhen Wang, and Kamalika Chaudhuri. Pufferfish privacy mechanisms for correlated data. In *Proceedings of the 2017 ACM International Conference on Management of Data*, SIGMOD '17, page 1291–1306, 2017.

[39] Mayuri Sridhar, Hanshen Xiao, and Srinivas Devadas. Pac-private algorithms. In *2025 IEEE Symposium on Security and Privacy (SP)*, pages 3839–3857. IEEE, 2025.

[40] Eric Wong, Frank Schmidt, and Zico Kolter. Wasserstein adversarial examples via projected sinkhorn iterations. In *International conference on machine learning*, pages 6808–6817. PMLR, 2019.

[41] Hanshen Xiao and Srinivas Devadas. Pac privacy: Automatic privacy measurement and control of data processing. In *Annual International Cryptology Conference*, pages 611–644. Springer, 2023.

[42] Hanshen Xiao, Jun Wan, and Srinivas Devadas. Geometry of sensitivity: Twice sampling and hybrid clipping in differential privacy with optimal gaussian noise and application to deep learning. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 2636–2650, 2023.

[43] Xiaokui Xiao and Yufei Tao. Output perturbation with query relaxation. *Proceedings of the VLDB Endowment*, 1(1):857–869, 2008.

[44] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE, 2018.

[45] Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, et al. Opacus: User-friendly differential privacy library in pytorch. *arXiv preprint arXiv:2109.12298*, 2021.

[46] Tao Zhang, Bradley A Malin, Netanel Raviv, and Yevgeniy Vorobeychik. Differential confounding privacy and inverse composition. In *2025 IEEE International Symposium on Information Theory (ISIT)*, pages 1–6. IEEE, 2025.

[47] Tao Zhang and Yevgeniy Vorobeychik. Breaking the gaussian barrier: Residual-pac privacy for automatic privatization. *arXiv preprint arXiv:2506.06530*, 2025.

[48] Tao Zhang and Yevgeniy Vorobeychik. Sliced rényi pufferfish privacy: Tractable privatization mechanism and private learning with gradient clipping. Cryptology ePrint Archive, Paper 2026/321, 2026.

[49] Wanrong Zhang, Olga Ohrimenko, and Rachel Cummings. Attribute privacy: Framework and mechanisms. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 757–766, New York, NY, USA, 2022. Association for Computing Machinery.

Layer & Question	**Main objects**	**Role in the paper**
Semantic interpretation: *what is protected?*	Pufferfish scenario $(\mathcal{S}, Q_{,}, \Theta)$ (Sec. 2.1)	Specifies the protected secrets, secret pairs, and admissible priors about secret–data relationships. DP corresponds to a record-level special case, while PP/RPP/SRPP use the broader Pufferfish semantic interpretation.
Quantification measure: *how is leakage measured?*	PP, RPP, Ave-SRPP, Joint-SRPP; SRD under slice profile $(\mathcal{U}, \omega)$ (Sec. 4)	PP, RPP, and SRPP share the Pufferfish semantic interpretation but use different measures of output-distribution distinguishability. SRPP replaces full-dimensional Rényi divergence used by RPP with sliced Rényi divergence (SRD).
Privatization mechanism: *how is noise calibrated?*	Wasserstein sensitivity Δ_∞; sliced Wasserstein sensitivities Δ_∞^u (Sec. 3); Gaussian SWM (Sec. 4.3.1)	Classical Wasserstein mechanisms calibrate to full-dimensional OT sensitivity. SWM aligns SRPP with sliced Wasserstein sensitivity, replacing high-dimensional OT computations with one-dimensional projected sensitivity computations.
Finite profiles: *what does finite slicing mean?*	Mode-I finite-profile SRPP; Mode-II continuous-profile SRPP via finite estimation (Sec. 4.4)	The same finite set of directions has two interpretations. In Mode-I it is the slice profile in the SRPP definition. In Mode-II it estimates quantities for a target continuous slice profile and requires PAC-style finite-slice and finite-sample corrections.
Iterative learning: *how is SGD privatized?*	Gradient clipping; HUC; sa-HUC; SRPP-SGD / sa-SRPP-SGD (Sec. 5.1–5.2)	Direct per-iteration Pufferfish sensitivity computation is impractical and arbitrary Pufferfish mechanisms do not compose gracefully. HUC/sa-HUC use structured discrepancy caps that support tractable per-iteration calibration and additive accounting under stated conditions.
Composition: *what composes?*	HUC/sa-HUC-based SRPP-SGD mechanisms under a shared slice profile (Sec. 5.4)	Provides additive accounting for the HUC/sa-HUC-based SRPP-SGD mechanisms developed in the paper under a shared slice profile; this is the structured composition property needed for the iterative-learning construction.
Experiments: what is evaluated?	Static SWM experiments and SRPP-SGD experiments (Sec. 6); diagnostic SRPP-vs-RPP comparison (Sec. 4.5)	Evaluates the scalability and privacy–utility behavior of SRPP mechanisms in static query release and iterative learning. The diagnostic comparison further illustrates how native certificate values relate to common operational metrics.

Table 1: Summary of the paper's conceptual layers. The table separates the semantic layer, quantification layer, mechanism layer, finite-profile interpretation, iterative-learning construction, and empirical evaluation.

Section	Notation	Meaning
Preliminaries		
Sec. 2.1	$(\mathcal{S}, Q, \Theta)$	Pufferfish scenario: secret set, protected secret pairs, and admissible priors.
Sec. 2.1	s_i, s_j	Protected secrets, with $(s_i, s_j) \in Q$.
Sec. 2.1	$\theta \in \Theta$	An admissible prior or data-generating model in the Pufferfish scenario.
Sec. 2.1	$X \in \mathcal{X}$	Dataset.
Sec. 2.2	$\mathtt{D}_\alpha(\mathsf{P}\|\mathsf{Q})$	Rényi divergence of order $\alpha > 1$.
Sec. 2.3	$W_\infty(\mu, \nu)$	(Full-dimensional) ∞-Wasserstein Distance (∞-WD).
Sec. 2.3	Δ_∞	(Full-dimensional) ∞-Wasserstein sensitivity.
Slicing and sliced Wasserstein sensitivity		
Sec. 3	$P_\theta^{f,s}$	Conditional law of the clean query output $f(X)$ given secret s under prior θ.
Sec. 3	$g_\# \mu(A)$	Pushforward on $\mathcal{C}$, where $g : \mathcal{W} \to \mathcal{C}$, μ is a Borel measure on $\mathcal{W}$.
Sec. 3	$\mathrm{SW}_p^p(\mu, \nu)$	Sliced Wasserstein Distance (SWD).
Sec. 3	$u \in \mathbb{S}^{d-1}$	Projection direction on the unit sphere $\mathbb{S}^{d-1}$.
Sec. 3	$\Psi^u(a) = \langle a, u \rangle$	One-dimensional projection map along direction u.
Sec. 3	$\Psi^u_\# P$	Pushforward law of P under the projection Ψ^u.
Sec. 3	Δ_∞^u	Per-slice ∞-Wasserstein sensitivity (SW-sensitivity) along direction u.
SRPP definitions and Sliced Wasserstein Mechanisms		
Sec. 4	$(\mathcal{U}, \omega)$	Slice profile: a set or distribution of projection directions and associated weights.
Sec. 4.1	$\mathrm{AveSD}_\alpha^\omega(P\|Q)$	Average sliced Rényi divergence under slice profile ω.
Sec. 4.1 Sec. 4.2	$\mathrm{JSD}_\alpha^\omega(P\|Q)$	Joint sliced Rényi divergence, using a log-moment aggregation over slices.
Sec. 4.3.1	Δ_ω^2	Slice-profile average of squared per-slice sensitivities.
Sec. 4.3.1	$\Delta_\star^2$	Worst-case squared per-slice sensitivity over the slice profile.
SRPP-SGD		
Sec. 5	η	subsampling rate.
Sec. 5	ρ	$\rho \in \{\mathsf{WR}, \mathsf{WOR}, \mathsf{Possion}\}$ notes sampling type.
Sec. 5	$r \sim \mathbb{P}_{\eta,\rho}$	sampling draw.
Sec. 5	$\mathsf{I}_t(r)$	(multi)set of selected indices at iteration t.
Sec. 5	$B_t(r)$	$B_t(r) = \|\mathsf{I}_t(r)\|$ size of $I_t(r)$. When $\rho \in \{\mathsf{WR}, \mathsf{WOR}\}$, $B_t(r) = B$ is deterministic.
Sec. 5	ξ_t	Gradient parameter at iteration t.
Sec. 5	f_t	Pre-noise update function.
Sec. 5	$g_t(x; \xi_{t-1})$	Per-example gradients at iteration t.
Sec. 5.1	$h_t = (h_{t,i})_{i=1}^m$	History-Uniform Cap (HUC) at iteration t when there are m slices in $\mathcal{U}$.
Sec. 5.1	K_t	Discrepancy cap.
Sec. 5.2	$h_t^{\mathsf{sa}} = (h_t^{\mathsf{sa}})_{i=1}^m$	Subsampling-Aware History-Uniform Cap (sa-HUC).
Sec. 5.2	$\overline{K}_t$	Subsampleing-aware discrepancy cap.
Sec. 5.4	$\vec{\mathcal{M}}(X)$	composition of J SRPP-SGD mechanisms $\mathcal{M}_1, \ldots, \mathcal{M}_J$ acting on the same dataset X.

Table 2: Summary of important notation.

Outline of Appendix

Tables

- Table 1 provides a conceptual overview of the paper, organizing the framework by layer—semantic, quantification, and mechanism—together with the finite-profile interpretation, iterative-learning construction, and composition results. It is intended as a navigational aid for the technical material that follows.

- Table 2 lists important notations.

Background

App. A provides additional preliminaries on differential privacy (App. A.1), DP-SGD (App. A.2), and a relaxation of RPP's general Wasserstein mechanism (GWM), termed the *Distribution-Aware General Wasserstein Mechanism* (App. A.3). Appendix A.4 introduces *worst-case group privacy*, which captures a robust (prior-free) RPP guarantee that does not depend on the prior family Θ.

App. A.5.1 formalizes an RPP variant in which the corresponding ∞-Wasserstein distance is approximated via entropically regularized (Sinkhorn) optimal transport. We refer to the resulting privatization method as the *Sinkhorn–Wasserstein mechanism*, and discuss key limitations for privacy quantification and accounting.

SRPP Related Discussion and Analysis

App. B presents detailed properties of SRPP, including ordering (App. B.1), post-processing immunity (App. B.2), a deeper comparison between Ave- and Joint-SRPP (App. B.3), and conversions from SRPP to *sliced Pufferfish privacy* (App. B.4).

App. C gives two examples of update maps that satisfy the L_t-Lipschitz assumption in Sec. 5.1. App. D provides analyses and practical approaches for estimating the discrepancy cap in (15). App. E remarks on Poisson subsampling in HUC/sa-HUC-based accounting.

App. F characterizes *per-layer gradient clipping* as an alternative setting to the slice-wise assumptions used in Sec. 5.1. App. G clarifies the relationship between SRPP/sa-SRPP and group DP-SGD; note that *group DP* here differs from the *worst-case group privacy* notion in App. A.4.

App. H establishes the *minimality* and *attainability* of HUCs (Definition 5.1 in Sec. 5.1). App. I compares minibatch sampling with replacement (WR) and without replacement (WOR) in the HUC/sa-HUC framework. App. J provides formal utility characterizations showing why subsampling-aware bounds (sa-HUC/sa-SRPP) can be strictly less conservative.

Proofs and Experimental Details

App. K–V contain the formal proofs of the main-body results. App. X provides additional experimental details.

A Additional Preliminaries

A.1 Differential Privacy

This section revisits the Differential Privacy (DP). Let X be the input dataset. Each data point x_i is defined over some measurable domain $\mathcal{X}^\dagger$, so that $x = (x_1, x_2, \ldots, x_n) \in \mathcal{X} = (\mathcal{X}^\dagger)^n$. We say two datasets $x, x' \in \mathcal{X}$ are *adjacent* if they differ in exactly one data point.

Definition A.1 ((ε, δ)-Differential Privacy [12])**.** *A randomized mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ is said to be (ε, δ)-differentially private (DP), with $\varepsilon \geq 0$ and $\delta \in [0,1]$, if for any pair of adjacent datasets x, x', and for all $\mathcal{T} \subseteq \mathcal{Y}$, we have* $\Pr[\mathcal{M}(x) \in \mathcal{T}] \leq e^{\varepsilon} \Pr[\mathcal{M}(x') \in \mathcal{T}] + \delta$.

The parameter ε is usually referred to as the *privacy budget*, and $\delta \in (0,1]$ represents the failure probability. DP represents a worst-case, input-agnostic adversarial worst-case framework.

A.2 DP-SGD with Gradient Clipping

In this section, we describe the Differentially private stochastic gradient descent (DP-SGD) [1, 37] has become the standard approach for training neural networks with formal privacy guarantees. First, We provide the following basic setup for DP-SGD similar to Sec. 5 for completeness.

Dataset. Let $x = (x_1, \ldots, x_n) \in \mathcal{X} = \bar{\mathcal{X}}^n$ be the finite samples of data points, where each $x_i = (a_i, b_i) \in \bar{\mathcal{X}}$ consists of features a_i and label b_i. In this section, we take P_θ^X as the empirical (marginal) distribution of the data samples $x = (x_1, \ldots, x_n) \in \mathcal{X} = \bar{\mathcal{X}}^n$ for all $\theta \in \Theta$. At iteration t, we draw a random mini-batch $\mathsf{I}_t \subseteq [n]$ using a subsampling scheme $\{\eta, \rho\}$ characterized by a rate η and a scheme type ρ, where $\rho \in \{\mathsf{WR}, \mathsf{WOR}, \mathsf{Poisson}\}$ denotes sampling with-replacement (WR), without-replacement (WOR), or via a Poisson process (Poisson), respectively. Let $X^t = (X_i)_{i \in \mathsf{I}_t}$ denote the random mini-batch subsampled from x, with x^t representing a particular realization.

Empirical Risk Minimization. In supervised learning, each example is $x_i = (a_i, b_i)$ with features a_i and label b_i. Let $\ell(\theta; x_i)$ denote the per-sample loss of a model with parameter ξ. The empirical risk minimization (ERM) problem is $\min_\xi F(\xi) := \frac{1}{n} \sum_{i=1}^n \ell(\theta; x_i)$.

Algorithm 2 Noise-perturbed SGD with gradient clipping

Require: Data $x_1,\dots,x_n$; loss $\ell(\xi;x)$; stepsizes $(\eta_t)_{t=0}^{T-1}$; batch size B; clip norm C; noise covariances $(\Sigma_t)_{t=1}^{T}$.

Initialize parameters ξ_0
for $t = 0,\dots,T-1$ **do**
  Sample minibatch $\mathsf{I}_t \subseteq [n]$ with $|\mathsf{I}_t| = B$
  For each $i \in \mathsf{I}_t$:

$$g_t(x_i) \leftarrow \nabla_\xi \ell(\xi_t;x_i),\ \tilde{g}_t(x_i) \leftarrow g_t(x_i)\min\left\{1, \frac{C}{\|g_t(x_i)\|_2}\right\}$$

  $\bar{g}_t \leftarrow \frac{1}{B}\sum_{i\in\mathsf{I}_t} \tilde{g}_t(x_i)$
  Draw $N_t \sim \mathcal{N}(0,\Sigma_t)$ and set $\hat{g}_t \leftarrow \bar{g}_t + N_t$
  $\xi_{t+1} \leftarrow \theta_t - \eta_t \hat{g}_t$
end for
return θ_T; compute privacy parameters from (Σ_t) according to DP guarantee.

Stochastic Gradient Descent. SGD approximately solves ERM by iteratively sampling minibatches and updating θ along noisy gradients. At iteration t, a subsampling scheme $\{\eta,\rho\}$ produces a random minibatch index set $\mathsf{I}_t \subseteq [n]$ of size B and corresponding minibatch $x^t = (x_i)_{i\in\mathsf{I}_t}$. The (unclipped) stochastic gradient is

$$g_t(x_i) := \nabla_\xi \ell(\xi_t;x_i), \qquad g_t := \frac{1}{B}\sum_{i\in\mathsf{I}_t} g_t(x_i),$$

and the parameters are updated as $\xi_{t+1} = \xi_t - \eta_t g_t$, for stepsize $\eta_t > 0$.

Gradient clipping and DP-SGD. DP-SGD [1, 37] modifies SGD in two ways: it clips per-example gradients to control sensitivity and adds noise to the averaged gradient for privacy. Each per-sample gradient is clipped at ℓ_2 radius C,

$$\tilde{g}_t(x_i) := g_t(x_i)\min\left\{1, \frac{C}{\|g_t(x_i)\|_2}\right\},$$

and the averaged clipped gradient is

$$\bar{g}_t := \frac{1}{B}\sum_{i\in\mathsf{I}_t} \tilde{g}_t(x_i).$$

Noise-perturbed SGD with gradient clipping (Algorithm 2) then draws Gaussian noise $N_t \sim \mathcal{N}(0,\Sigma_t)$, forms $\hat{g}_t = \bar{g}_t + N_t$, and updates $\xi_{t+1} = \xi_t - \eta_t\hat{g}_t$. In DP-SGD with isotropic Gaussian noise perturbation, the Gaussian noise variance is calibrated so that the resulting algorithm satisfies a target (ε,δ)-DP guarantee via, e.g., moment accountant [1].

Since the Pufferfish scenario considers probabilistic secret-dataset relationship captured by the priors, DP and SRPP are fundamentally different in both semantics as well as how the privacy leakages are aggregated iteratively in SGD algorithms. App. G provides detailed discussion between group DP-SGD and our SRPP-SGD.

A.3 Distribution Aware General Wasserstein Mechanism for RPP

In this section, we recall the Distribution Aware General Wasserstein Mechanism (DAGWM) of [34], which refines the General Wasserstein Mechanism (GWM) for Rényi Pufferfish Privacy (RPP) by replacing worst-case ∞-Wasserstein sensitivities with p-Wasserstein quantities that better capture average-case geometry. The DAGWM is proved to offer more utility than the GWM at no additional privacy cost [34].

Definition A.2 (p-Wasserstein distance)**.** *Fix a norm $\|\cdot\|$ on $\mathbb{R}^d$ and let $1 \le p < \infty$. For probability measures ν,μ on $\mathbb{R}^d$ with finite p-th moments, the p-Wasserstein distance is*

$$W_p(\nu,\mu) := \inf_{\pi\in\Pi(\nu,\mu)} \left(\int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^p d\pi(x,y)\right)^{1/p}$$

If either ν or μ lacks a finite p-th moment, set $W_p(\nu,\mu) = +\infty$.

For $1 \le p \le s \le \infty$, $W_p(\nu,\mu) \le W_s(\nu,\mu)$. Moreover, $W_p(\nu,\mu)$ is nondecreasing in p, and for measures with bounded support we can relate large-p Wasserstein distances to the ∞-Wasserstein distance.

To improve the utility of the GWM, Pierquin et al. [34] leverage p-Wasserstein distance to constrain the sensitivity instead of ∞-Wasserstein distances. This replaces the worst-case transportation cost between $\Pr(f(X) \mid s_i,\theta)$ and $\Pr(f(X) \mid s_j,\theta)$ by moments of the transportation cost. Let ζ be a noise distribution and, for $(s_i,s_j) \in \mathcal{Q}$ and $\theta \in \Theta$, write $\mu_i^\theta = P(f(X) \mid s_i,\theta)$. For $q \ge 1$ and $\alpha > 1$, define the p-Wasserstein sensitivity as

$$\Delta_G^{\zeta,q,\alpha} := \max_{(s_i,s_j),\theta} \inf_{(X,Y)\sim\Pi(\mu_i^\theta,\mu_j^\theta)} \mathbb{E}\Big[\exp\left(q(\alpha-1)\mathtt{D}_{q(\alpha-1)+1}(\zeta,\zeta*(X-Y))\right)\Big]. \tag{21}$$

The DAGWM calibration term is governed by full-dimensional p-Wasserstein distances between the conditional output distributions $\Pr(f(X) \mid s,\theta)$. When $f(X) \in \mathbb{R}^d$ is high-dimensional, accurately estimating these distances from empirical samples and computing them at scale is known to be difficult in general for optimal transport. Therefore, while DAGWM improves the distribution-awareness of RPP relative to worst-case W_∞-based calibration, *it still inherits the high-dimensional intractability associated with full-dimensional Wasserstein sensitivity.* This motivates replacing full-dimensional OT calibration by sliced constructions, where the mechanism depends only on one-dimensional projections and remains computable and calibratable in high dimensions.

A.4 Worst-Case Group Privacy

Following [34, 38], in this section, we define a *worst-case group privacy* baseline for a given Pufferfish scenario

$(\mathcal{S}, Q, \Theta)$, which is a *prior-free* (group-DP-type) quantification that does not exploit any distributional restrictions encoded by Θ.

Fix $n \in \mathbb{N}^*$ and group dimensions $d_1, \dots, d_n \in \mathbb{N}^*$. Let $\mathcal{X} \subset \mathbb{R}^d$ and define the block-structured dataset space

$$\mathcal{X}^{\text{all}} := \mathcal{X}^{d_1} \times \cdots \times \mathcal{X}^{d_n}.$$

For $x \in \mathcal{X}^{\text{all}}$, write $x = (x^{(1)}, \dots, x^{(n)})$ with $x^{(k)} \in \mathcal{X}^{d_k}$ the k-th group (block).

Group adjacency and worst-case group sensitivity. For each $k \in \{1, \dots, n\}$, define the group adjacency relation

$$D_k := \left\{ (x, x') \in (\mathcal{X}^{\text{all}})^2 \ : \ x^{(j)} = x'^{(j)} \ \ \forall j \neq k \right\},$$

i.e., $(x, x') \in D_k$ iff x and x' may differ arbitrarily within the k-th group and are identical outside that group. For any query $f : \mathcal{X}^{\text{all}} \to \mathbb{R}^m$, define the worst-case group sensitivity (with respect to $\|\cdot\|$ on $\mathbb{R}^m$) as

$$\Delta_{\text{GROUP}}(f) := \max_{k \in \{1,\dots,n\}} \sup_{(x,x') \in D_k} \|f(x) - f(x')\|.$$

Equivalently, letting $D := \bigcup_{k=1}^n D_k$, one has

$$\Delta_{\text{GROUP}}(f) = \sup_{(x,x') \in D} \|f(x) - f(x')\|.$$

Worst-case (prior-free) RPP under group adjacency. A prior-free (worst-case) instantiation of RPP requires the certificate to hold uniformly over the chosen adjacency class; here we take $D = \bigcup_{k=1}^n D_k$. Consider the additive-noise mechanism $M(x) = f(x) + Z$. We say that M satisfies *worst-case (prior-free) RPP* at order $\alpha > 1$ with budget ε if

$$\sup_{(x,x') \in D} \mathtt{D}_\alpha\big(M(x) \,\|\, M(x')\big) \ \leq \ \varepsilon.$$

In the noise-calibration form of [34], a sufficient condition is

$$R_\alpha\big(Z, \Delta_{\text{GROUP}}(f)\big) \leq \varepsilon$$

where

$$R_\alpha(Z, z) := \sup_{\|u\| \leq z} \mathtt{D}_\alpha(Z - u \,\|\, Z)$$

is the Rényi envelop (i.e., the worst-case Rényi divergence) between the noise distribution and its shifts by at most z.

Validity and Conservativeness. By construction, the above worst-case (prior-free) specification yields a *valid* RPP certificate: the privacy guarantee is quantified by a supremum over the entire adjacency class $D = \bigcup_{k=1}^n D_k$ and therefore does not rely on any distributional restriction (in particular, it does not exploit Θ). In particular, if $R_\alpha(Z, \Delta_{\text{GROUP}}(f)) \leq \varepsilon$, then the additive-noise mechanism $M(x) = f(x) + Z$ satisfies worst-case RPP at order $\alpha > 1$ with budget ε:

$$\sup_{(x,x') \in D} D_\alpha\big(M(x) \,\|\, M(x')\big) \leq \varepsilon.$$

However, this prior-free baseline is typically *utility-draining* because $\Delta_{\text{GROUP}}(f)$ is an *absolute worst-case* sensitivity over all group-adjacent dataset pairs; for highly correlated data, such group-level worst-case protections can force noise magnitudes that scale with the group size and may destroy utility [38]. Since $R_\alpha(Z, z)$ is monotone nondecreasing in z for common noise families, achieving the same privacy budget ε under the prior-free baseline generally requires larger noise, which in turn degrades utility.

A.5 Sinkhorn-Wasserstein Mechanism for RPP

In this section, we introduce the *Sinkhorn-Wasserstein (Sinkhorn-W) mechanism* as an estimated version of the standard full-dimensional WD, where the corresponding Wasserstein distance is estimated by the Sinkhorn algorithm [5, 7, 24, 40].

A.5.1 Entropy-Regularized Wasserstein Distance

We start by defining the *entropy-regularized* Wasserstein distance (ER-WD). The WD in Definition 2.3 can be equivalently represented as follows. Let

$$J_t := \left\{ (x, y) \in \mathbb{R}^d \times \mathbb{R}^d : \|x - y\| \leq t \right\},$$

$$\Pi_t(\nu, \mu) := \{ \pi \in \Pi(\nu, \mu) : \pi(J_t) = 1 \}.$$

Then,

$$W_\infty(\nu, \mu) = \inf\{ t \geq 0 : \Pi_t(\nu, \mu) \neq \emptyset \}. \tag{22}$$

The ER-WD re-formulates the WD in (22) by adding a regularization based on Kullback–Leibler (KL) divergence. For any regularization parameter $\hat{\varepsilon} > 0$, the ER-WD is then defined as

$$W_\infty^{(\hat{\varepsilon})}(\nu, \mu) := \inf_{t \geq 0, \pi \in \Pi(\nu,\mu)} \{ t + \hat{\varepsilon} \mathrm{KL}(\pi \| \nu \otimes \mu) : \pi(J_t) = 1 \}, \tag{23}$$

where

$$\mathrm{KL}(\pi \| \nu \otimes \mu) = \int \log\left(\tfrac{d\pi}{d(\nu \otimes \mu)} \right) d\pi$$

is the KL divergence with $\mathrm{KL}(\pi \| \nu \otimes \mu) = +\infty$ if $\pi \not\ll (\nu \otimes \mu)$. For ER-WD, we still minimizing the smallest radius t for which a coupling exists supported inside $\|x - y\| \leq t$, but we also penalize couplings that have low entropy (i.e., "too concentrated"). Equivalently, (23) can be viewed as

$$W_\infty^{(\hat{\varepsilon})}(\nu, \mu) = \inf_{t \geq 0} \left\{ t + \hat{\varepsilon} \inf_{\pi \in \Pi_t(\nu,\mu)} \mathrm{KL}(\pi \| \nu \otimes \mu) \right\}.$$

The discrete formulation follows immediately by restricting ν, μ to empirical measures and optimizing over transport matrices (π) with fixed marginals, with the same entropic KL regularizer and an ∞-cost implemented through a distance-threshold support constraint.

We can also define the more general entropy-regularized p-WD as

$$W^p_{p,\hat{\varepsilon}}(\nu,\mu) := \inf_{\pi\in\Pi(\nu,\mu)} \left\{ \int \|x-y\|^p d\pi(x,y) + \hat{\varepsilon}\mathrm{KL}(\pi\|\nu\otimes\mu) \right\}.$$

A.5.2 Sinkhorn Algorithm

For each $t \geq 0$, define the KL-minimizing coupling

$$\pi_t^\star \in \arg\min_{\pi\in\Pi_t(\nu,\mu)} \mathrm{KL}(\pi \,\|\, \nu\otimes\mu),$$

and the associated profile objective

$$J_{\hat{\varepsilon}}(t) \;:=\; t \;+\; \hat{\varepsilon}\,\mathrm{KL}(\pi_t^\star \,\|\, \nu\otimes\mu).$$

Then (23) can be written compactly as

$$W_\infty^{(\hat{\varepsilon})}(\nu,\mu) = \inf_{t\geq 0} J_{\hat{\varepsilon}}(t).$$

A standard result for KL projections onto marginal constraints is that the solution (when it exists) admits a multiplicative scaling representation: there exist strictly positive measurable functions $u_t : \mathbb{R}^d \to (0,\infty)$ and $v_t : \mathbb{R}^d \to (0,\infty)$, unique up to the rescaling $(u_t, v_t) \mapsto (c\,u_t, c^{-1}v_t)$ for $c > 0$, such that

$$d\pi_t(x,y) = u_t(x)\, v_t(y)\, \mathbf{1}\{\|x-y\| \leq t\}\, d\nu(x)\, d\mu(y).$$

Imposing the marginal constraints $\pi_t(\cdot,\mathbb{R}^d) = \nu$ and $\pi_t(\mathbb{R}^d,\cdot) = \mu$ yields the IPFP fixed-point equations

$$\int_{\mathbb{R}^d} v_t(y)\, \mathbf{1}\{\|x-y\| \leq t\}\, d\mu(y) \;=\; \frac{1}{u_t(x)} \quad \text{for } \nu\text{-a.e. } x,$$

$$\int_{\mathbb{R}^d} u_t(x)\, \mathbf{1}\{\|x-y\| \leq t\}\, d\nu(x) \;=\; \frac{1}{v_t(y)} \quad \text{for } \mu\text{-a.e. } y.$$

This motivates the (continuous) Sinkhorn–Knopp algorithm (also known as the iterative proportional fitting procedure, IPFP) [22]: initialize a measurable function $v_t^{(0)} : \mathbb{R}^d \to (0,\infty)$, and for $k \geq 0$ define measurable functions $u_t^{(k+1)}, v_t^{(k+1)} : \mathbb{R}^d \to (0,\infty)$ by

$$u_t^{(k+1)}(x) = \left(\int_{\mathbb{R}^d} \mathbf{1}\{\|x-y\| \leq t\}\, v_t^{(k)}(y)\, d\mu(y) \right)^{-1},$$

$$v_t^{(k+1)}(y) = \left(\int_{\mathbb{R}^d} \mathbf{1}\{\|x-y\| \leq t\}\, u_t^{(k+1)}(x)\, d\nu(x) \right)^{-1},$$

for ν-a.e. x and μ-a.e. y, respectively. If the iterates $(u_t^{(k)}, v_t^{(k)})$ converge to (u_t, v_t), then the coupling

$$d\pi_t(x,y) = u_t(x) v_t(y)\, \mathbf{1}\{\|x-y\| \leq t\}\, d\nu(x)\, d\mu(y)$$

has marginals ν and μ.

We write π_t^{SK} for the coupling induced by the Sinkhorn(-Knopp) iterates after termination, i.e.,

$$d\pi_t^{\mathrm{SK}}(x,y) = u_t^{(k)}(x) v_t^{(k)}(y) \mathbf{1}\{\|x-y\| \leq t\}\, d\nu(x)\, d\mu(y),$$

for the final iterate k. Given a candidate set $\mathcal{T} \subseteq [0,\infty)$, we define the Sinkhorn-computed ER-∞-WD by

$$W_\infty^{(\hat{\varepsilon}),\mathrm{SK}}(\nu,\mu;\mathcal{T}) \;:=\; \min_{t\in\mathcal{T}} \left\{ t + \hat{\varepsilon}\,\mathrm{KL}\left(\pi_t^{\mathrm{SK}} \,\middle\|\, \nu\otimes\mu\right) \right\}. \quad (24)$$

Discrete Case. Let $\nu = \sum_{i=1}^n a_i \delta_{x_i}$ and $\mu = \sum_{j=1}^m b_j \delta_{y_j}$ with $a \in \Delta_n$ and $b \in \Delta_m$. Define the admissibility mask $M(t) \in \{0,1\}^{n\times m}$ by $M_{ij}(t) = \mathbf{1}\{\|x_i - y_j\| \leq t\}$, and the reference matrix $K(t) := (ab^\top) \odot M(t)$. Then π_t admits the scaling form $\pi_t = \mathrm{diag}(u)\, K(t)\, \mathrm{diag}(v)$, where $u \in \mathbb{R}^n_+$ and $v \in \mathbb{R}^m_+$ are obtained by the Sinkhorn updates $u \leftarrow a \oslash (K(t)v)$ and $v \leftarrow b \oslash (K(t)^\top u)$.

Exactness w.r.t. ER-WD. Assume that for every $t \in \mathcal{T}$ the Sinkhorn/IPFP iterate converges to the exact KL projection, i.e., $\pi_t^{\mathrm{SK}} = \pi_t^\star \in \arg\min_{\pi\in\Pi_t(\nu,\mu)} \mathrm{KL}(\pi\|\nu\otimes\mu)$. If, in addition, $\mathcal{T}$ contains a minimizer of the outer problem, i.e., $\arg\min_{t\geq 0}\{t + \hat{\varepsilon}\inf_{\pi\in\Pi_t(\nu,\mu)} \mathrm{KL}(\pi\|\nu\otimes\mu)\} \cap \mathcal{T} \neq \emptyset$, then

$$W_\infty^{(\hat{\varepsilon}),\mathrm{SK}}(\nu,\mu;\mathcal{T}) = W_\infty^{(\hat{\varepsilon})}(\nu,\mu).$$

Relation to W_∞. For any $\hat{\varepsilon} > 0$ and any $\mathcal{T} \subseteq [0,\infty)$, one has

$$W_\infty^{(\hat{\varepsilon}),\mathrm{SK}}(\nu,\mu;\mathcal{T}) \geq \min\{t \in \mathcal{T} \;:\; \Pi_t(\nu,\mu) \neq \emptyset\}.$$

Moreover, if $\hat{\varepsilon} \downarrow 0$ and the inner KL problem is solved exactly (so that $\pi_t^{\mathrm{SK}} = \pi_t^\star$), then

$$\lim_{\hat{\varepsilon}\downarrow 0} W_\infty^{(\hat{\varepsilon})}(\nu,\mu) = W_\infty(\nu,\mu),$$

and, in the discrete setting, taking $\mathcal{T}$ to be the set of distinct distances $\{\|x_i - y_j\|\}_{i,j}$ yields $\min\{t \in \mathcal{T} : \Pi_t(\nu,\mu) \neq \emptyset\} = W_\infty(\nu,\mu)$.

In the present ∞-WD formulation, setting $\hat{\varepsilon} = 0$ removes the KL term and reduces the inner problem at a fixed threshold t to the feasibility question $\Pi_t(\nu,\mu) \neq \emptyset$, i.e., the existence of a coupling supported on $J_t = \{(x,y) : \|x-y\| \leq t\}$. In the discrete setting $\nu = \sum_i a_i \delta_{x_i}$, $\mu = \sum_j b_j \delta_{y_j}$, this feasibility problem is equivalent to the existence of a nonnegative transport matrix with prescribed marginals and support constrained to $\{(i,j) : \|x_i - y_j\| \leq t\}$, and can be decided via standard max-flow/feasible-circulation methods (equivalently, Hall-type conditions on the bipartite support graph). Consequently, the computational advantage of Sinkhorn/IPFP is tied to $\hat{\varepsilon} > 0$: the entropic regularization yields a strictly convex projection problem with a multiplicative scaling structure, whereas the unregularized case $\hat{\varepsilon} = 0$ reverts to a (potentially large-scale) feasibility/flow problem and does not admit the same Sinkhorn scaling iterations.

A.5.3 Limitations of Sinkhorn-Wasserstein Mechanism

For any query function $f : \mathcal{X} \to \mathbb{R}^d$ and any $\mathcal{T} \subseteq [0,\infty)$, define the ∞-*Sinkhorn-Wasserstein sensitivity (Sinkhorn-W sensitivity)* as

$$\Delta_\infty^{\mathrm{SK},\hat{\varepsilon}}(\mathcal{T}) := \max_{(s_i,s_j)\in Q,\theta\in\Theta} W_\infty^{(\hat{\varepsilon}),\mathrm{SK}}\left(P_\theta^{f,s_i}, P_\theta^{f,s_j}; \mathcal{T}\right). \tag{25}$$

Define the full-dimensional Rényi envelope:

$$R_\alpha(\zeta, z) = \sup_{\|a\|<z} \mathtt{D}_\alpha(\zeta_{-a}, \zeta),$$

where $\|\cdot\|$ is the same norm as in the definition of W_∞ in Definition 2.3.

The Sinkhorn-Wasserstein (Sinkhorn-W) mechanism replaces the unsliced W_∞ sensitivity by an entropy-regularized surrogate, Sinkhorn-W sensitivity, computed numerically. While this can reduce optimization difficulty, it introduces additional approximation layers that complicate privacy certification. In particular, if the sensitivity used for calibration is under-estimated (due to regularization bias, grid omission, or premature solver termination), then the injected noise may be insufficient for the intended worst-case guarantee, and the resulting privacy statement may no longer apply to the true W_∞-based notion.

Regularization Changes the Certified Notion. For $\hat{\varepsilon} > 0$, the entropy-regularized objective $W_\infty^{(\hat{\varepsilon})}$ generally does not equal W_∞. Consequently, calibrating noise using $\Delta_\infty^{\mathrm{SK},\hat{\varepsilon}}(\mathcal{T})$ yields a certificate tied to the *regularized* sensitivity. Interpreting this certificate as a guarantee for the true W_∞ sensitivity requires an additional approximation argument as $\hat{\varepsilon} \downarrow 0$ (and, in the discrete case, a sufficiently rich threshold set $\mathcal{T}$). Without such an argument, the mechanism may only be provably private with respect to the surrogate objective rather than the original worst-case coupling radius W_∞.

Two Additional Approximation Layers Beyond Sampling. Even for fixed $\hat{\varepsilon} > 0$, the computed value $W_\infty^{(\hat{\varepsilon}),\mathrm{SK}}(\nu,\mu;\mathcal{T})$ can differ from $W_\infty^{(\hat{\varepsilon})}(\nu,\mu)$ for two independent reasons. First, the outer minimization is discretized by restricting to $\min_{t\in\mathcal{T}}$ instead of $\inf_{t\geq 0}$; if $\mathcal{T}$ omits near-minimizers, the reported value may deviate from the true optimum. Second, the inner KL projection is solved only approximately by terminating Sinkhorn/IPFP after finitely many iterations. Without a certified primal–dual gap (or other a posteriori bound), the resulting coupling may violate the marginal constraints beyond the desired tolerance, and the reported objective value can be inaccurate. From the standpoint of privacy calibration, the critical failure mode is *under-estimation*: if the computed sensitivity proxy is smaller than the true one, the selected noise scale may be insufficient for the intended worst-case guarantee.

Conservativeness Can Be Large and Difficult to Control. At the optimal t, we pay t plus an entropic penalty that is generally strictly positive for the feasible set Π_t. Hence, $W_\infty^{(\hat{\varepsilon})}$ unless a KL-minimizing feasible coupling has zero KL (which is rare). This inflates the sensitivity proxy and can force the mechanism to inject unnecessary noise, degrading utility. Decreasing $\hat{\varepsilon}$ reduces this regularization bias, but it also exacerbates numerical issues (see the next paragraph). Thus, there is no monotone regime in which the method is simultaneously computationally benign and tightly aligned with W_∞.

Ill-Conditioning as $\hat{\varepsilon} \downarrow 0$. The computational appeal of Sinkhorn/IPFP relies on $\hat{\varepsilon} > 0$, which yields a strictly convex KL projection with a multiplicative scaling structure. As $\hat{\varepsilon} \downarrow 0$, the optimal t approaches the feasibility threshold. At (or near) this threshold, the mask $M(t)$ becomes sparse or near-disconnected, making scaling interations unstable or slow (and sometimes infeasible without enlargint t).

Runtime and Memory Remain Coupling-Scale in General. In the discrete setting with n samples per conditional law, each Sinkhorn/IPFP update with a dense pairwise structure costs $O(n^2)$ arithmetic (and typically $\Omega(n^2)$ memory if the pairwise distance/kernel is materialized). Moreover, the required number of iterations depends on $\hat{\varepsilon}$ and on the chosen marginal-violation tolerance. Thus, although Sinkhorn avoids generic coupling optimization, its worst-case scaling remains coupling-sized and is, in general, less favorable than sliced calibration pipelines that reduce computation to sorting-based one-dimensional operations.

Computational Properties. Fix an instance with empirical measures supported on n points in $\mathbb{R}^d$. Let $\mathcal{T}$ be a finite set of distance thresholds and let K be the number of Sinkhorn/IPFP iterations run per $t \in \mathcal{T}$. Forming the full pairwise distance table $\{\|x_i - y_j\|_2\}_{i,j\in[n]}$ costs $\Theta(dn^2)$ arithmetic operations. For each $t \in \mathcal{T}$, one Sinkhorn/IPFP update (two dense matrix–vector products plus entrywise scaling) costs $O(n^2)$ time, hence K updates cost $O(Kn^2)$. Therefore, evaluating the grid objective over $\mathcal{T}$ costs

$$T_{\mathrm{SK}}(1 \text{ instance}) = O\big(dn^2 + |\mathcal{T}|\,K\,n^2\big),$$

and over M instances,

$$T_{\mathrm{SK}} = O\Big(\mathsf{M}\big(dn^2 + |\mathcal{T}|\,K\,n^2\big)\Big),$$

where $\mathsf{M} := |\widehat{\Theta}||\widehat{Q}|$, with $\Omega(n^2)$ memory if the distance table or kernel is materialized.

Practical implication. A Sinkhorn-W mechanism can be attractive when a moderate $\hat{\varepsilon}$ and a moderate solver tolerance

yield acceptable utility. However, it introduces additional approximation knobs ($\hat{\varepsilon}, \mathcal{T}$, termination tolerance), and the resulting privacy calibration can become either overly conservative (when $\hat{\varepsilon}$ is large) or numerically burdensome (when $\hat{\varepsilon}$ is small). In contrast, sliced mechanisms avoid full-dimensional couplings entirely and admit exact one-dimensional subroutines, yielding simpler calibration logic and more predictable computational behavior.

B Properties of SRPP

In this section, we provide detailed discussions of Ave-SRPP and Joint-SRPP. Ave-SRPP and Joint-SRPP use the same ingredients (per-slice Rényi divergences along 1-D projections) but aggregate them in different ways, leading to different strengths. Ave-SRPP averages the per-slice Rényi divergences over directions $u \sim \omega$, so it controls the mean distinguishability of secrets across slices (good for mean geometry across slices). Joint-SRPP instead can be interpreted as bounding the Rényi divergence of the joint observation of slice index $V \sim \omega$ and projection $\Psi_{\#}^{V}(\mathcal{M}(X))$, which captures and penalizes rare, high-risk directions that Ave-SRPP alone can hide in the average. Joint-SRPP is generally a stronger requirement than Ave-SRPP: satisfying Joint-SRPP implies the corresponding averaged bound, but not conversely.

Sec. B.1 shows the ordering of Ave-SRD, Joint-SRD, and the standard Rényi divergence. In Sec. B.2, we characterize the post-processing immunity property of SRPP notions. Sec. B.3 further compares Ave-SRPP and Joint-SRPP. Finally, we show that our SRPP frameworks imply their corresponding sliced Pufferfish Privacy frameworks in Sec. B.4.

B.1 Ordering

Our two SRPP frameworks sit between each other and the standard RPP.

Proposition B.1. *Fix a slice profile* $\{\mathcal{U}, \omega\}$. *For* $d \geq 1$, $\alpha > 1$, *every secret pair* $(s_i, s_j) \in Q$ *and every prior* $\theta \in \Theta$,

$$\texttt{AveSD}_{\alpha}^{\omega}\big(\mathcal{M}_{s_i}^{\theta} \| \mathcal{M}_{s_j}^{\theta}\big) \leq \texttt{JSD}_{\alpha}^{\omega}\big(\mathcal{M}_{s_i}^{\theta} \| \mathcal{M}_{s_j}^{\theta}\big) \leq \mathtt{D}_{\alpha}\big(\mathcal{M}_{s_i}^{\theta} \| \mathcal{M}_{s_j}^{\theta}\big).$$

The ordering in Proposition B.1 only compares the *numerical tightness of upper bounds* and does *not* induce a straightforward ranking of privacy strength between the three notions. For example, an (α, ε)-RPP guarantee implies the corresponding $(\alpha, \varepsilon, \omega)$-Joint/Ave-SRPP guarantees by Proposition B.1, but the converse need not hold. Moreover, different choices of ω and different calibration strategies for the noise can lead to mechanisms with the same bound on $\texttt{AveSD}_{\alpha}^{\omega}$ or $\texttt{JSD}_{\alpha}^{\omega}$ but very different protections at the level of secrets and priors. For this reason, the inequalities in Proposition B.1 should be interpreted as relations between divergence-based certificates, rather than as an absolute ranking of which framework is "more private".

Proof of Proposition B.1. Fix a slice profile $\{\mathcal{U}, \omega\}$, a prior $\theta \in \Theta$, and a secret pair $(s_i, s_j) \in Q$ with $P_{\theta}^{S}(s_i), P_{\theta}^{S}(s_j) > 0$. Write

$$\mathrm{P} := \mathcal{M}_{s_i}^{\theta}, \text{ and } \mathrm{Q} := \mathcal{M}_{s_j}^{\theta}$$

for the output probability measures of the mechanism under s_i and s_j, respectively. Let $\{\mathcal{U}, \omega\}$ with $\mathcal{U} = \{u_\ell\}_{\ell=1}^{m}$ and $\omega = \{\omega_\ell\}_{\ell=1}^{m}$ be the discrete slicing profile, and denote

$$D(u) := \mathtt{D}_{\alpha}\big(\Psi_{\#}^{u_\ell}\mathrm{P} \| \Psi_{\#}^{u_\ell}\mathrm{Q}\big), \qquad \ell = 1, \ldots, m.$$

By Definition 4.1, the average sliced Rényi divergence is

$$\texttt{AveSD}_{\alpha}^{\omega}(\mathrm{P}\|\mathrm{Q}) = \int_{\mathbb{S}^{d-1}} D(u)\, d\omega(u)$$

By Definition 4.3, the joint sliced Rényi divergence is

$$\texttt{JSD}_{\alpha}^{\omega}(\mathrm{P}\|\mathrm{Q}) = \frac{1}{\alpha - 1} \log\Big(\int_{\mathbb{S}^{d-1}} \exp\big((\alpha-1)D(u)\, d\omega(u)\big)\Big).$$

Part 1: $\texttt{AveSD}_{\alpha}^{\omega}(\mathrm{P}\|\mathrm{Q}) \leq \texttt{JSD}_{\alpha}^{\omega}(\mathrm{P}\|\mathrm{Q})$**.**

Let Z be the random variable $Z = D(U)$ when $U \sim \omega$. Then,

$$\mathbb{E}[Z] = \int_{\mathbb{S}^{d-1}} D(u)\, d\omega(u) = \texttt{AveSD}_{\alpha}^{\omega}(\mathrm{P}\|\mathrm{Q}),$$

and

$$\texttt{JSD}_{\alpha}^{\omega}(\mathrm{P}\|\mathrm{Q}) = \frac{1}{\alpha - 1} \log \mathbb{E}\big[\exp\big((\alpha-1)Z\big)\big]$$

For $\alpha > 1$, $\exp((\alpha-1)z)$ is convex. Then, Jensen's inequality gives

$$\mathbb{E}\big[\exp\big((\alpha-1)Z\big)\big] \geq \exp\big((\alpha-1)\mathbb{E}[Z]\big).$$

Thus, by taking log and dividing by $(\alpha-1)$, we get

$$\texttt{JSD}_{\alpha}^{\omega}(\mathrm{P}\|\mathrm{Q}) \geq \mathbb{E}[Z] = \texttt{AveSD}_{\alpha}^{\omega}(\mathrm{P}\|\mathrm{Q}).$$

Part 2: $\texttt{JSD}_{\alpha}^{\omega}(\mathrm{P}\|\mathrm{Q}) \leq \mathtt{D}_{\alpha}(\mathrm{P}\|\mathrm{Q})$**.**

Define the randomized post-processing channel Slice as follows: given $Y \in \mathbb{R}^d$, draw $U \sim \omega$ independently of Y, then output $\mathsf{Slice}(Y) := (U, \langle Y, U\rangle)$. Let $\mathrm{P}^{\mathsf{Slice}}$ and $\mathrm{Q}^{\mathsf{Slice}}$ denote the laws of $\mathsf{Slice}(Y)$ when $Y \sim \mathrm{P}$ and $Y \sim \mathrm{Q}$, respectively.

For each direction u, the conditional law of $\langle Y, u\rangle$ under $Y \sim \mathrm{P}$ is $\Psi_{\#}^{u}\mathrm{P}$, and similarly under Q it is $\Psi_{\#}^{u}\mathrm{Q}$. By the definition of Rényi divergence for product/mixture measures with the same mixing law ω, we have

$$\begin{aligned}
&\mathtt{D}_{\alpha}(\mathrm{P}^{\mathsf{Slice}}\|\mathrm{Q}^{\mathsf{Slice}}) \\
&= \frac{1}{\alpha-1} \log \int_{\mathbb{S}^{d-1}} \exp\Big((\alpha-1)\mathtt{D}_{\alpha}(\Psi_{\#}^{u}\mathrm{P}\,\|\,\Psi_{\#}^{u}\mathrm{Q})\Big)\, d\omega(u) \\
&= \texttt{JSD}_{\alpha}^{\omega}(\mathrm{P}\|\mathrm{Q}).
\end{aligned}$$

Since Slice is a (data-independent) post-processing channel applied to Y, the data-processing inequality for Rényi divergence (an f-divergence for $\alpha > 1$) yields

$$\mathtt{D}_{\alpha}(\mathrm{P}^{\mathsf{Slice}}\|\mathrm{Q}^{\mathsf{Slice}}) \leq \mathtt{D}_{\alpha}(\mathrm{P}\|\mathrm{Q}),$$

and hence $\mathtt{JSD}_\alpha^\omega(\mathrm{P}\|\mathrm{Q}) \le \mathtt{D}_\alpha(\mathrm{P}\|\mathrm{Q})$.

By combining **Part 1** and **Part 2**, we can claim

$$\mathtt{AveSD}_\alpha^\omega\big(\mathcal{M}_{s_i}^\theta \| \mathcal{M}_{s_j}^\theta\big) \le \mathtt{JSD}_\alpha^\omega\big(\mathcal{M}_{s_i}^\theta \| \mathcal{M}_{s_j}^\theta\big) \le \mathtt{D}_\alpha\big(\mathcal{M}_{s_i}^\theta \| \mathcal{M}_{s_j}^\theta\big),$$

□

B.2 Post-Processing Immunity

A fundamental closure property of DP and RPP is post-processing immunity (PPI): the privacy loss of a mechanism cannot be increased by arbitrary data-independent post-processing [13, 34]. Formally, for any Markov kernel (randomized mapping) K that is independent of the dataset, post-processing acts on the *endpoint distributions* of a release. Since the Rényi divergence $\mathtt{D}_\alpha(\cdot\|\cdot)$ satisfies the data-processing inequality (DPI),

$$\mathtt{D}_\alpha(PK \,\|\, QK) \;\le\; \mathtt{D}_\alpha(P \,\|\, Q),$$

any privacy certificate expressed in terms of (upper bounds on) $\mathtt{D}_\alpha$ is automatically closed under post-processing.

Post-processing immunity has two aspects. (i) *True privacy loss never increases:* for any data-independent K, $\mathtt{D}_\alpha(PK \,\|\, QK) \le \mathtt{D}_\alpha(P \,\|\, Q)$ by DPI. (ii) *Safe certificate reuse for a downstream pipeline:* one may upper-bound the privacy loss of a post-processing pipeline using a certificate for the initial mechanism. For SRPP, this reuse must be interpreted through the relationship between sliced divergences and the full-dimensional Rényi divergence (Prop. B.1): SRPP/SRD is a quantification language for the full release, and changing the representation along the pipeline can change the numerical SRPP value even though the underlying distinguishability (and hence privacy loss in the sense of (i)) cannot increase. Consequently, reusing SRPP parameters for an end-to-end pipeline can be valid but may be overly conservative when the quantification geometry matters.

In privacy, PPI should be understood primarily as a *semantic closure* property: for any data-independent post-processing K applied to the true release, the post-processed pipeline cannot be *more* distinguishable than the original release in the underlying divergence sense (DPI). This is stronger than (and logically prior to) any particular *parameterization rule* such as "reusing the same numeric privacy parameters", which is merely a convenient upper-bounding convention and can be very conservative (or not directly meaningful) when the quantification geometry changes (as in SRPP).

SRPP's post-processing discussion is ultimately rooted in the data-processing inequality (DPI) for the standard Rényi divergence $\mathtt{D}_\alpha$. Concretely, SRPP quantifies privacy by applying a fixed (profile-determined) slicing/observation channel to the *endpoint distribution* of the (full-dimensional) released output, and then evaluating $\mathtt{D}_\alpha$ on the induced sliced laws: Ave-SRPP takes the ω-average of the resulting per-direction divergences, while Joint-SRPP applies a log-moment aggregation that is equivalent to a single $\mathtt{D}_\alpha$ on an augmented observation. Accordingly, for any data-independent post-processing kernel K applied to the true release $Y = \mathcal{M}(X)$, one first obtains the post-processed endpoint laws PK and QK, and SRPP/SRD is then evaluated on PK versus QK under the same slice profile. Any nonconservation of the SRPP numerical value under post-processing reflects this change in endpoint distributions (and the geometry induced by the chosen profile), rather than a failure of post-processing immunity: the underlying distinguishability cannot increase due to DPI.

Let $P = \Pr[\mathcal{M}(X) \mid s_i, \theta]$ and $Q = \Pr[\mathcal{M}(X) \mid s_j, \theta]$ be the endpoint distributions on $\mathbb{R}^d$, and let K be any data-independent post-processing kernel on $\mathbb{R}^d$. Then DPI gives $\mathtt{D}_\alpha(PK\|QK) \le \mathtt{D}_\alpha(P\|Q)$ (true privacy loss never increases). Moreover, by Prop. B.1,

$$\mathtt{AveSD}_\alpha^\omega(PK\|QK) \le \mathtt{JSD}_\alpha^\omega(PK\|QK) \le \mathtt{D}_\alpha(PK\|QK),$$

so any full-dimensional Rényi/RPP certificate for $\mathcal{M}$ remains a valid upper bound on the SRPP/SRD risk of the post-processed pipeline $K \circ \mathcal{M}$. In general, however, the SRPP values $\mathtt{AveSD}_\alpha^\omega(\cdot\|\cdot)$ and $\mathtt{JSD}_\alpha^\omega(\cdot\|\cdot)$ need not be numerically conserved under arbitrary K, because they quantify endpoint distinguishability through a fixed slice profile.

B.3 Average vs. Joint SRPP

Average SRPP (Ave-SRPP) and Joint SRPP (Joint-SRPP) are built from the same ingredients, Rényi divergences of one-dimensional projections, but aggregate these per-direction costs in different ways, leading to distinct strengths and use cases.

Recall that for any pair of probability measures P, Q on $\mathbb{R}^d$ and a slice profile $\{\mathcal{U}, \omega\}$, the average sliced Rényi divergence of order $\alpha > 1$ is

$$\mathtt{AveSD}_\alpha^\omega(\mathrm{P}\|\mathrm{Q}) := \int_{\mathbb{S}^{d-1}} \mathtt{D}_\alpha\big(\Psi_\#^u P \,\big\|\, \Psi_\#^u Q\big)\, d\omega(u),$$

while the joint sliced Rényi divergence is

$$\mathtt{JSD}_\alpha^\omega(\mathrm{P}\|\mathrm{Q}) := \frac{1}{\alpha-1} \log \mathbb{E}_{U\sim\omega}\Big[\exp\big((\alpha-1)\mathtt{D}_\alpha\big(\Psi_\#^U \mathrm{P} \big\| \Psi_\#^U \mathrm{Q}\big)\big)\Big].$$

Ave-SRPP requires that, for every prior $\theta \in \Theta$ and every secret pair $(s_0, s_1) \in \mathcal{Q}$, the *average* sliced Rényi divergence between the corresponding output probability distribution is bounded by ε. Joint-SRPP instead bounds the log-moment aggregation of per-slice divergences, which can be interpreted as the Rényi divergence between the joint distributions $(\widetilde{\mathrm{P}}(\cdot,\cdot))$ of the *slice (index) and the projected output*.

In particular, define the randomized sliced observation channel $\mathsf{Slice} : \mathbb{R}^d \to \mathcal{U} \times \mathbb{R}$ by

$$\mathsf{Slice}(y) := (U, \Psi^U(y)), \qquad U \sim \omega \text{ independent of } y.$$

Let $\widetilde{\mathrm{P}} := \mathsf{Slice}_{\#}\mathrm{P}$ and $\widetilde{\mathrm{Q}} := \mathsf{Slice}_{\#}\mathrm{Q}$ denote the endpoint laws of $(U, \Psi^U(Z_{\mathrm{P}}))$ and $(U, \Psi^U(Z_{\mathrm{Q}}))$ when $Z_{\mathrm{P}} \sim \mathrm{P}$ and $Z_{\mathrm{Q}} \sim \mathrm{Q}$, respectively. Then

$$\mathtt{JSD}_\alpha^\omega(\mathrm{P}\|\mathrm{Q}) = \mathtt{D}_\alpha\big(\widetilde{\mathrm{P}} \,\big\|\, \widetilde{\mathrm{Q}}\big).$$

In this sense, Joint-SRPP is exactly RPP applied to the augmented observation $(U, \Psi^U(\mathcal{M}(X)))$, whereas Ave-SRPP only controls the mean per-direction divergence.

These two aggregation schemes are ordered between each other and the full Rényi divergence. For any P, Q on $\mathbb{R}^d$ and any slice distribution ω,

$$\mathtt{AveSD}_\alpha^\omega(\mathrm{P}\|\mathrm{Q}) \le \mathtt{JSD}_\alpha^\omega(\mathrm{P}\|\mathrm{Q}) \le \mathtt{D}_\alpha(\mathrm{P}\|\mathrm{Q}),$$

and the same inequalities hold for the corresponding output distributions under any fixed Pufferfish model $(\mathcal{S}, \mathcal{Q}, \Theta)$ ((Proposition B.1)). Thus, at the level of numerical certificates, Joint-SRPP is always at least as "conservative" as Ave-SRPP for the same $(\alpha, \varepsilon, \omega)$, and both are upper-bounded by the unsliced RPP divergence.

Although Ave-SRPP and Joint-SRPP are closely related, they serve complementary purposes.

Ave-SRPP for Geometry-Aware *Average* Protection. Ave-SRPP directly reflects the geometry induced by the slicing distribution ω and is naturally aligned with sliced Wasserstein mechanism privatization, where one averages per-direction sensitivities. Formally, it controls the *expected* sliced Rényi divergence when the projection direction is drawn from ω. Equivalently, it upper-bounds the expected per-direction divergence for a direction drawn from ω; this is an analytic interpretation only (the mechanism still releases the full d-dimensional output). This "single-slice observer" is purely an analytic device: the actual mechanism always outputs the full d-dimensional vector, and the sliced view is used only to define a geometry-aware privacy budget. In practice, this makes Ave-SRPP a natural *calibration target*, often yielding tighter, more utility-friendly noise levels than enforcing a worst-case bound simultaneously in every direction.

Joint-SRPP for Robust, Slice-Aware Guarantees. Joint-SRPP instead can be interpreted as if it treats the slice index as part of the observation channel and bounds the Rényi divergence of the joint pair $(U, \Psi^U(\mathcal{M}(X)))$. In terms of risk quantification, this corresponds to accounting for an observer that can exploit knowledge of *which* direction was used, not just the projected value. Thus Joint-SRPP protects against worst-case sliced distinguishability over the randomized slicing procedure, rather than its average, and provides a more conservative, slice-aware notion of privacy than Ave-SRPP.

B.4 From SRPP to Sliced Pufferfish Privacy

This appendix records standard Rényi-to-approximate guarantees for our sliced notions, with notation consistent with Sec. 3–4.2. Fix a Pufferfish scenario $(\mathcal{S}, \mathcal{Q}, \Theta)$ and a slice profile $(\mathcal{U}, \omega)$, where ω is a probability measure on $\mathbb{S}^{d-1}$ supported on $\mathcal{U}$. For a mechanism $\mathcal{M}$ and $\theta \in \Theta$, $s \in \mathcal{S}$, write

$$\mathcal{M}_s^\theta \;:=\; \Pr\big(\mathcal{M}(X) \mid s, \theta\big),$$

for the conditional output distribution. For each $u \in \mathbb{S}^{d-1}$, recall the projection $\Psi^u(a) = \langle a, u\rangle$ and its pushforward $\Psi^u_{\#}$ as in (3).

Approximate Max-Divergence. For distributions P, Q on a common measurable space and $\delta \in (0,1)$, define

$$\mathtt{D}_\infty^\delta(P\|Q) := \inf\Big\{\varepsilon \ge 0 :\ \forall A,\ P(A) \le e^\varepsilon Q(A) + \delta\Big\}. \quad (26)$$

These are the (ε, δ) analogues of Definitions 4.2 and 4.4, obtained by replacing per-slice Rényi divergences with approximate max-divergences.

Definition B.1 ($(\varepsilon,\delta,\omega)$-Ave-SPP)**.** *A mechanism $\mathcal{M}$ satisfies $(\varepsilon,\delta,\omega)$-Ave-SPP in $(\mathcal{S},\mathcal{Q},\Theta)$ if for all $\theta \in \Theta$ and $(s_i, s_j) \in \mathcal{Q}$ with $P_\theta^S(s_i), P_\theta^S(s_j) > 0$,*

$$\int_{\mathbb{S}^{d-1}} \mathtt{D}_\infty^\delta\Big(\Psi^u_{\#}\mathcal{M}_{s_i}^\theta \,\Big\|\, \Psi^u_{\#}\mathcal{M}_{s_j}^\theta\Big)\, \mathrm{d}\omega(u) \;\le\; \varepsilon.$$

Definition B.2 ($(\varepsilon,\delta,\omega)$-Joint-SPP)**.** *Let $V \sim \omega$ be independent of X and of the internal randomness of $\mathcal{M}$, and define the direction-revealing projected mechanism*

$$\widetilde{\mathcal{M}}(X) \;:=\; \big(V,\ \Psi^V(\mathcal{M}(X))\big).$$

We say that $\mathcal{M}$ satisfies $(\varepsilon,\delta,\omega)$-Joint-SPP in $(\mathcal{S},\mathcal{Q},\Theta)$ if $\widetilde{\mathcal{M}}$ satisfies (ε,δ)-PP in $(\mathcal{S},\mathcal{Q},\Theta)$, i.e., for all $\theta \in \Theta$ and $(s_i,s_j) \in \mathcal{Q}$ with $P_\theta^S(s_i), P_\theta^S(s_j) > 0$ and all measurable A,

$$\Pr\big(\widetilde{\mathcal{M}}(X) \in A \mid s_i, \theta\big) \;\le\; e^\varepsilon \Pr\big(\widetilde{\mathcal{M}}(X) \in A \mid s_j, \theta\big) \;+\; \delta.$$

Recall the Rényi divergence $\mathtt{D}_\alpha(\cdot\|\cdot)$ given by (1). Lemma B.1 shows conversion from Rényi divergenece to the approximate max-divegrence $\mathtt{D}_\infty^\delta$ in (26).

Lemma B.1 (Rényi to approximate max-divergence)**.** *Let $\alpha > 1$ and $\delta \in (0,1)$. If $\mathtt{D}_\alpha(P\|Q) \le \varepsilon_\alpha$, then*

$$\mathtt{D}_\infty^\delta(P\|Q) \;\le\; \varepsilon_\alpha \;+\; \frac{\log(1/\delta)}{\alpha - 1}.$$

Proof. Let $r := \frac{dP}{dQ}$ on $\{dQ > 0\}$ (and $r := +\infty$ on $\{dQ = 0, dP > 0\}$). For any measurable A and any $\tau > 0$,

$$\begin{aligned} P(A) = \int_A r\,dQ &= \int_{A\cap\{r\le\tau\}} r\,dQ + \int_{A\cap\{r>\tau\}} r\,dQ \\ &\le \tau Q(A) + P(r > \tau). \end{aligned}$$

Let $\tau := e^{\varepsilon}$. It suffices to bound $P(r > e^{\varepsilon})$. Applying Markov's inequality to $r^{\alpha-1}$ under P gives

$$P(r > e^{\varepsilon}) = P\big(r^{\alpha-1} > e^{(\alpha-1)\varepsilon}\big) \le \frac{\mathbb{E}_P[r^{\alpha-1}]}{e^{(\alpha-1)\varepsilon}} = \frac{\int r^{\alpha}\, dQ}{e^{(\alpha-1)\varepsilon}}.$$

By definition of Rényi divergence,

$$\int r^{\alpha}\, dQ = \exp\big((\alpha-1)\mathtt{D}_{\alpha}(P\|Q)\big) \le \exp\big((\alpha-1)\varepsilon_{\alpha}\big).$$

Hence

$$P(r > e^{\varepsilon}) \le \exp\big((\alpha-1)\varepsilon_{\alpha} - (\alpha-1)\varepsilon\big).$$

Setting $\varepsilon := \varepsilon_{\alpha} + \frac{\log(1/\delta)}{\alpha-1}$ yields $P(r > e^{\varepsilon}) \le \delta$, and therefore $P(A) \le e^{\varepsilon} Q(A) + \delta$ for all measurable A, which is equivalent to $\mathtt{D}_{\infty}^{\delta}(P\|Q) \le \varepsilon$. □

Applying Lemma B.1 to each sliced pair $(\Psi^u_{\#}\mathcal{M}^{\theta}_{s_i}, \Psi^u_{\#}\mathcal{M}^{\theta}_{s_j})$ yields the pointwise bound

$$\mathtt{D}_{\infty}^{\delta}\Big(\Psi^u_{\#}\mathcal{M}^{\theta}_{s_i} \,\Big\|\, \Psi^u_{\#}\mathcal{M}^{\theta}_{s_j}\Big) \le \mathtt{D}_{\alpha}\Big(\Psi^u_{\#}\mathcal{M}^{\theta}_{s_i} \,\Big\|\, \Psi^u_{\#}\mathcal{M}^{\theta}_{s_j}\Big) + \frac{\log(1/\delta)}{\alpha-1}.$$

Integrating over $u \sim \omega$ converts $(\alpha,\varepsilon,\omega)$-Ave-SRPP into an $(\varepsilon,\delta,\omega)$-Ave-SPP guarantee with $\varepsilon = \varepsilon + \frac{\log(1/\delta)}{\alpha-1}$; the same conversion applies to Joint-SRPP by applying Lemma B.1 to the joint endpoint laws of $(V, \Psi^V(\mathcal{M}(X)))$.

B.4.1 Ave-SRPP Implies Ave-SPP

Theorem B.3 (Ave-SRPP ⇒ Ave-SPP)**.** *Fix $\alpha > 1$. If $\mathcal{M}$ satisfies $(\alpha,\varepsilon,\omega)$-Ave-SRPP in $(\mathcal{S},\mathcal{Q},\Theta)$, then for any $\delta \in (0,1)$ it satisfies $(\varepsilon',\delta,\omega)$-Ave-SPP with*

$$\varepsilon' := \varepsilon + \frac{\log(1/\delta)}{\alpha-1}.$$

Proof. Fix $\theta \in \Theta$ and $(s_i,s_j) \in \mathcal{Q}$ with $P^S_{\theta}(s_i), P^S_{\theta}(s_j) > 0$. For each $u \in \mathcal{U}$, apply Lemma B.1 to $P^u := \Psi^u_{\#}\mathcal{M}^{\theta}_{s_i}$ and $Q^u := \Psi^u_{\#}\mathcal{M}^{\theta}_{s_j}$ to obtain

$$\mathtt{D}_{\infty}^{\delta}(P^u\|Q^u) \le \mathtt{D}_{\alpha}(P^u\|Q^u) + \frac{\log(1/\delta)}{\alpha-1}.$$

Integrating both sides with respect to ω yields

$$\int \mathtt{D}_{\infty}^{\delta}(P^u\|Q^u)\,\mathrm{d}\omega(u) \le \int \mathtt{D}_{\alpha}(P^u\|Q^u)\,\mathrm{d}\omega(u) + \frac{\log(1/\delta)}{\alpha-1}.$$

By $(\alpha,\varepsilon,\omega)$-Ave-SRPP (Definition 4.2), the second integral satisfies $\int \mathtt{D}_{\alpha}(P^u\|Q^u)\,d\omega(u) \le \varepsilon$. Therefore, $\int \mathtt{D}_{\infty}^{\delta}(P^u\|Q^u)\,d\omega(u) \le \varepsilon + \frac{\log(1/\delta)}{\alpha-1}$. □

B.4.2 Joint-SRPP Implies Joint-SPP

Theorem B.4 (Joint-SRPP ⇒ Joint-SPP)**.** *Fix $\alpha > 1$. If $\mathcal{M}$ satisfies $(\alpha,\varepsilon,\omega)$-Joint-SRPP in $(\mathcal{S},Q,\Theta)$, then for any $\delta \in (0,1)$ it satisfies $(\varepsilon',\delta,\omega)$-Joint-SPP with*[5]

$$\varepsilon' := \varepsilon + \frac{\log(1/\delta)}{\alpha-1}.$$

Proof. Fix $\theta \in \Theta$ and $(s_i,s_j) \in Q$ with $P^S_{\theta}(s_i), P^S_{\theta}(s_j) > 0$. Let $P := \mathcal{M}^{\theta}_{s_i}$ and $Q := \mathcal{M}^{\theta}_{s_j}$ be distributions on $\mathbb{R}^d$. Let $\widetilde{P}, \widetilde{Q}$ be the laws of $\widetilde{\mathcal{M}}(X) = (V, \Psi^V(\mathcal{M}(X)))$ under secrets s_i and s_j.

Step 1. We show that Joint-SRD is the Rényi divergence of the direction-revealing output. For each u, write $P^u := \Psi^u_{\#}P$ and $Q^u := \Psi^u_{\#}Q$. Since $V \sim \omega$ is the same under both secrets and is independent of $\mathcal{M}(X)$, the disintegrations of $\widetilde{P}$ and $\widetilde{Q}$ satisfy

$$\widetilde{P}(du,dy) = \omega(du)\,P^u(dy), \quad \widetilde{Q}(du,dy) = \omega(du)\,Q^u(dy).$$

Assuming $\widetilde{P} \ll \widetilde{Q}$ (otherwise the divergence is $+\infty$), the Radon-Nikodym derivative factorizes:

$$\frac{d\widetilde{P}}{d\widetilde{Q}}(u,y) = \frac{dP^u}{dQ^u}(y)$$

for $\widetilde{Q}$-a.e. (u,y). Therefore,

$$\begin{aligned}
\mathtt{D}_{\alpha}(\widetilde{P}\|\widetilde{Q}) &= \frac{1}{\alpha-1}\log\int\left(\frac{d\widetilde{P}}{d\widetilde{Q}}\right)^{\alpha} d\widetilde{Q} \\
&= \frac{1}{\alpha-1}\log\int_{\mathbb{S}^{d-1}} \omega(du)\int\left(\frac{dP^u}{dQ^u}(y)\right)^{\alpha} dQ^u(y) \\
&= \frac{1}{\alpha-1}\log\int_{\mathbb{S}^{d-1}} \exp\Big((\alpha-1)\mathtt{D}_{\alpha}(P^u\|Q^u)\Big)\,\mathrm{d}\omega(u) \\
&= \mathtt{JSD}^{\omega}_{\alpha}(P\|Q),
\end{aligned}$$

which matches Definition 4.3.

Step 2 (Apply the standard conversion). By $(\alpha,\varepsilon,\omega)$-Joint-SRPP (Definition 4.4),

$$\mathtt{D}_{\alpha}(\widetilde{P}\|\widetilde{Q}) = \mathtt{JSD}^{\omega}_{\alpha}(P\|Q) \le \varepsilon.$$

Applying Lemma B.1 to $(\widetilde{P},\widetilde{Q})$ gives

$$\mathtt{D}_{\infty}^{\delta}(\widetilde{P}\|\widetilde{Q}) \le \varepsilon + \frac{\log(1/\delta)}{\alpha-1} = \varepsilon'.$$

By the definition of $\mathtt{D}_{\infty}^{\delta}$, this inequality is equivalent to the (ε',δ)-PP condition for $\widetilde{\mathcal{M}}$, i.e., $(\varepsilon',\delta,\omega)$-Joint-SPP for $\mathcal{M}$. □

[5] Equivalently, $\widetilde{\mathcal{M}}$ is (ε',δ)-PP.

B.4.3 Optimizing Over Rényi Orders

If a mechanism admits Ave/Joint-SRPP bounds ε_α for multiple $\alpha > 1$, then we may optimize the conversion as

$$\varepsilon(\delta) := \inf_{\alpha>1} \left\{ \varepsilon_\alpha + \frac{\log(1/\delta)}{\alpha - 1} \right\}.$$

C Examples of L_t-Lipschitz Update Map

In this section, we provide two common cases of L_t-Lipschitz update maps described in Sec. 5.1.

(i) Preconditioned SGD. Suppose we use a (possibly history- and randomness-dependent) preconditioner $A_t : \mathcal{Y}_{<t} \times \mathcal{R} \to \mathbb{R}^{d\times d}$, so that

$$T_t(z; y_{<t}) = \xi_{t-1} - A_t(y_{<t}, R_t)\, z.$$

Then T_t is L_t-Lipschitz in its first argument with

$$L_t := \sup_{y_{<t}, r} \big\| A_t(y_{<t}, r) \big\|_{\mathrm{op}}, \qquad \|A\|_{\mathrm{op}} := \sup_{\|v\|_2 = 1} \|Av\|_2.$$

In standard SGD, $A_t(y_{<t}, r) = \kappa I_d$ (step size κ), hence $L_t = \kappa$.

(ii) Proximal or projected updates. Let

$$\mathrm{prox}(v) := \arg\min_{w\in\mathbb{R}^d} \left\{ \tfrac{1}{2}\|w - v\|_2^2 + \bar{\lambda}\mathrm{Reg}(w) \right\},$$

where $\bar{\lambda} > 0$ and $\mathrm{Reg} : \mathbb{R}^d \to (-\infty, +\infty]$ is a proper, lower-semicontinuous, convex regularizer. Let $\overline{\Pi}_{\mathcal{C}}$ be the Euclidean projection onto a closed convex set $\mathcal{C} \subset \mathbb{R}^d$. Since both prox and $\overline{\Pi}_{\mathcal{C}}$ are 1-Lipschitz, the same constant L_t serves as a Lipschitz bound when

$$T_t(z; y_{<t}) = \mathrm{prox}\big(\xi_{t-1} - A_t(y_{<t}, R_t)z\big),$$

or

$$T_t(z; y_{<t}) = \overline{\Pi}_{\mathcal{C}}\big(\xi_{t-1} - A_t(y_{<t}, R_t)z\big).$$

D Discrepancy Cap Estimation

This appendix describes principled approaches to instantiating the *discrepancy cap* terms used by HUC/sa-HUC accounting. Fix a Pufferfish scenario $(\mathcal{S}, \mathcal{Q}, \Theta)$ and an iteration index t. Let η_t denote the (possibly randomized) minibatch selection rule (distribution) at iteration t with batch size B_t. For each $\theta \in \Theta$ and each secret $s \in \mathcal{S}$ with $P_\theta^S(s) > 0$, let $P_\theta(X \mid S = s)$ denote the conditional data distribution.

D.1 Overview and Estimation Intuition

The HUC and sa-HUC accountants require per-iteration discrepancy caps that quantify how much a minibatch update can change when the protected secret is switched within the specified Pufferfish scenario. The key point is that this quantity is not a gradient norm or a Wasserstein sensitivity. Instead, it is a discrepancy count: under a coupling of two secret-conditioned data distributions, we compare the two coupled datasets and count how many sampled training units in the minibatch differ. This count determines the maximal shift that gradient clipping can induce in the clipped minibatch update.

This appendix formalizes and estimates these caps. For deterministic HUC, the cap controls a worst-case or high-probability upper bound on the discrepancy count. For sa-HUC, the cap controls its second moment under minibatch subsampling, which captures the fact that a randomly sampled minibatch may contain only a small portion of the secret-induced discrepancies. Thus, estimating HUC/sa-HUC reduces the history-dependent Pufferfish sensitivity problem to an estimable discrepancy-count problem determined by three ingredients: the Pufferfish scenario $(\mathcal{S}, \mathcal{Q}, \Theta)$, a coupling between the corresponding secret-conditioned data distributions, and the minibatch sampling rule.

We provide two types of instantiations. First, when a simulator for the Pufferfish scenario is available, Monte Carlo sampling yields finite-sample upper confidence bounds for both HUC and sa-HUC caps. Second, when the scenario has additional record-level structure, closed-form bounds can be obtained from total-variation or localized-influence assumptions. The resulting caps can be plugged directly into the HUC/sa-HUC accounting theorems in the main text. This error-aware estimation layer is important because practical Pufferfish-family mechanisms depend on prior-driven, distributional quantities that are rarely available exactly; the HUC/sa-HUC formulation makes this estimation step explicit and provides finite-sample calibration guarantees for the resulting caps.

Notation and Correspondence to the Main Text. The main body defines the discrepancy count $K_t(x, x'; r) = \sum_{j \in \mathsf{I}_t(r)} \mathbf{1}\{x_j \neq x'_j\}$ and its coupling-based cap $K_t(\gamma) = \operatorname{ess\,sup}_{((X,X'),R_t)\sim\gamma\times\mathbb{P}_{\eta,\rho}} K_t(X, X'; R_t)$ (see (14)–(15)). In this appendix we denote the same random discrepancy count by $\Delta_t = \Delta_t(X^{(i)}, X^{(j)}, R_t)$ to emphasize its role as a random variable under the coupled experiment. Concretely, under the identification $(X^{(i)}, X^{(j)}) \equiv (X, X')$ and with the convention that $\mathsf{I}_t(r)$ is interpreted as a set (WOR) or a length-B_t sequence (WR), we have $\Delta_t \equiv K_t(X, X'; R_t)$. Accordingly, the deterministic HUC cap in (29) is a uniform upper bound on $\operatorname{ess\,sup} \Delta_t$, and the sa-HUC cap in (30) uniformly bounds $\mathbb{E}[\Delta_t^2]$ over the subsampling randomness.

D.2 Discrepancy Random Variable and Caps

Caps are defined relative to a *coupling* between two conditional data distributions. Fix $\theta \in \Theta$ and $(s_i, s_j) \in \mathcal{Q}$ with $P_\theta^S(s_i) > 0$ and $P_\theta^S(s_j) > 0$. Let $\Gamma_{\theta,ij} \in \Pi\big(P_\theta(X \mid S = s_i), P_\theta(X \mid S = s_j)\big)$ be a coupling, and draw $(X^{(i)}, X^{(j)}) \sim \Gamma_{\theta,ij}$. Draw a minibatch index multiset/vector $R_t =$

$(I_1,\dots,I_{B_t}) \sim \eta_t$. Define the batch discrepancy count

$$\Delta_t := \Delta_t\big(X^{(i)},X^{(j)},R_t\big) = \#\Big\{b \in \{1,\dots,B_t\} : x^{(i)}_{I_b} \neq x^{(j)}_{I_b}\Big\}, \tag{27}$$

where the indices $(I_1,\dots,I_{B_t})$ may be sampled with or without replacement according to η_t. For structured data, the relation "$\neq$" is interpreted as inequality of the atomic data unit that determines the per-example gradient contribution (e.g., a record, a user trajectory, or another unit of influence).

By construction,

$$0 \le \Delta_t \le B_t, \qquad \text{and hence} \qquad 0 \le \Delta_t^2 \le B_t^2. \tag{28}$$

Deterministic (HUC) Caps. For a fixed choice of coupling family $\{\Gamma_{\theta,ij}\}$, a deterministic HUC cap is any $K_t \in [0,B_t]$ such that

$$K_t \ge \sup_{\theta\in\Theta} \sup_{(s_i,s_j)\in\mathcal{Q}} \operatorname*{ess\,sup}_{(X^{(i)},X^{(j)},R_t)\sim\Gamma_{\theta,ij}\times\eta_t} \Delta_t\big(X^{(i)},X^{(j)},R_t\big). \tag{29}$$

Equivalently, (29) enforces $\Delta_t \le K_t$ almost surely, uniformly over all $\theta \in \Theta$ and $(s_i,s_j) \in \mathcal{Q}$.

Second-Moment (sa-HUC) Caps. A subsampling-aware (sa-HUC) cap is any scalar $\overline{K}_t^2 \in [0,B_t^2]$ such that

$$\overline{K}_t^2 \ge \sup_{\theta\in\Theta} \sup_{(s_i,s_j)\in\mathcal{Q}} \operatorname*{ess\,sup}_{(X^{(i)},X^{(j)})\sim\Gamma_{\theta,ij}} \mathbb{E}_{R_t\sim\eta_t}\big[\Delta_t\big(X^{(i)},X^{(j)},R_t\big)^2\big]. \tag{30}$$

Tail vs. Moment. In practice, rather than enforcing the almost-sure cap (29), one may specify a tail level $\delta_t \in (0,1)$ and seek a *high-probability* cap $K_t(\delta_t)$ such that

$$\Pr\big(\Delta_t \le K_t(\delta_t)\big) \ge 1-\delta_t \tag{31}$$

uniformly over $\theta \in \Theta$ and $(s_i,s_j) \in \mathcal{Q}$, where the probability is taken over $(X^{(i)},X^{(j)}) \sim \Gamma_{\theta,ij}$ and $R_t \sim \eta_t$. For sa-HUC, one instead seeks a uniform second-moment cap $\overline{K}_t^2$ satisfying (30).

D.3 Monte Carlo Estimation via a Simulator

Assume we have (i) sampling access to $P_\theta(X \mid S=s)$ for any $\theta \in \Theta$ and $s \in \mathcal{S}$, and (ii) the ability to sample a minibatch draw $R_t \sim \eta_t$. Fix $\theta \in \Theta$ and $(s_i,s_j) \in \mathcal{Q}$ with $P_\theta^S(s_i) > 0$ and $P_\theta^S(s_j) > 0$, and fix a coupling $\Gamma_{\theta,ij} \in \Pi\big(P_\theta(X \mid S=s_i), P_\theta(X \mid S=s_j)\big)$. The goal is to estimate the distribution of the discrepancy count Δ_t defined in (27), under the joint randomness

$$(X^{(i)},X^{(j)}) \sim \Gamma_{\theta,ij}, \quad R_t \sim \eta_t.$$

Algorithm 3 Monte Carlo estimation of discrepancy caps at iteration t (fixed θ, (s_i,s_j) and coupling $\Gamma_{\theta,ij}$)

Require: Iteration t; batch size B_t; minibatch subsampling rule η_t; coupling sampler for $\Gamma_{\theta,ij}$; Monte Carlo size M; tail level $\delta_t \in (0,1)$; calibration failure $\gamma_t \in (0,1)$.

for $m = 1,\dots,M$ **do**
  Sample $(X^{(i,m)},X^{(j,m)}) \sim \Gamma_{\theta,ij}$.
  Sample $R_t^{(m)} = (I_1^{(m)},\dots,I_{B_t}^{(m)}) \sim \eta_t$.
  Compute $\Delta_t^{(m)} = \#\{b \in \{1,\dots,B_t\} : x^{(i,m)}_{I_b^{(m)}} \neq x^{(j,m)}_{I_b^{(m)}}\}$.
end for
Set $\varepsilon \leftarrow \sqrt{\frac{1}{2M}\log\frac{2}{\gamma_t}}$.
sa-HUC output: $\widehat{\mu}_{2,t} \leftarrow \frac{1}{M}\sum_{m=1}^M (\Delta_t^{(m)})^2$ and $\widehat{K}_t^2 \leftarrow \widehat{\mu}_{2,t} + B_t^2\varepsilon$.
HUC output: if $\delta_t > \varepsilon$, let $\widehat{K}_t(\delta_t)$ be the empirical $(1-(\delta_t-\varepsilon))$-quantile of $\{\Delta_t^{(m)}\}_{m=1}^M$; otherwise set $\widehat{K}_t(\delta_t) \leftarrow B_t$.
return $(\widehat{K}_t(\delta_t), \widehat{K}_t^2)$.

Coupling Constructions. Any coupling $\Gamma_{\theta,ij}$ consistent with the scenario may be used. Typical constructions include:

- **Independent coupling:** draw $X^{(i)} \sim P_\theta(X \mid S=s_i)$ and $X^{(j)} \sim P_\theta(X \mid S=s_j)$ independently.
- **Shared-latent coupling:** when the conditional distribution admits a representation $X = g(Z,S)$ for some latent Z, draw $Z \sim P_\theta(Z)$ and set $X^{(i)} = g(Z,s_i)$, $X^{(j)} = g(Z,s_j)$.
- **Recordwise maximal coupling:** for i.i.d. record models, couple records coordinatewise using maximal couplings of the per-record conditionals; see Section D.4.

All bounds below are stated for an arbitrary fixed coupling $\Gamma_{\theta,ij}$; tighter couplings can yield smaller caps.

Monte Carlo Procedures. Draw i.i.d. replicates

$$(X^{(i,m)},X^{(j,m)}) \sim \Gamma_{\theta,ij}, \qquad R_t^{(m)} \sim \eta_t, \qquad m = 1,\dots,M,$$

compute $\Delta_t^{(m)}$ via (27), and construct either (i) a second-moment cap for sa-HUC or (ii) a tail cap for HUC. The parameter $\gamma_t \in (0,1)$ denotes the *calibration failure probability* associated with the Monte Carlo estimation, whereas $\delta_t \in (0,1)$ denotes the *tail level* of the desired HUC cap.

D.3.1 Finite-Sample Bounds for sa-HUC

We give a direct concentration bound for $\mathbb{E}[\Delta_t^2]$, using the bounded range $0 \le \Delta_t^2 \le B_t^2$ from (28).

Lemma D.1 (Monte Carlo sa-HUC cap via Hoeffding)**.** *Fix an iteration t, a prior $\theta \in \Theta$, a secret pair $(s_i,s_j) \in \mathcal{Q}$ with*

$P^S_\theta(s_i) > 0$ *and* $P^S_\theta(s_j) > 0$*, a coupling* $\Gamma_{\theta,ij}$*, and a minibatch subsampling rule* η_t*. Let* $\Delta_t^{(1)}, \dots, \Delta_t^{(M)}$ *be i.i.d. draws of the discrepancy count* Δ_t *under the joint randomness*

$$(X^{(i)}, X^{(j)}) \sim \Gamma_{\theta,ij}, \qquad R_t \sim \eta_t,$$

and define the empirical second moment

$$\widehat{\mu}_{2,t}(\theta, ij) \;:=\; \frac{1}{M} \sum_{m=1}^{M} \big(\Delta_t^{(m)}\big)^2.$$

Then for any $\gamma_t \in (0,1)$*, with probability at least* $1-\gamma_t$ *over the Monte Carlo sampling,*

$$\mathbb{E}\big[\Delta_t^2\big] \;\le\; \widehat{\mu}_{2,t}(\theta, ij) \;+\; B_t^2 \sqrt{\frac{1}{2M} \log \frac{2}{\gamma_t}}. \tag{32}$$

Proof. Let $Z_m := (\Delta_t^{(m)})^2$ for $m = 1, \dots, M$. By construction, $\Delta_t \in \{0, 1, \dots, B_t\}$ ((28)), hence

$$0 \le Z_m \le B_t^2 \qquad \text{a.s.}$$

Moreover, $\{Z_m\}_{m=1}^M$ are i.i.d. because the pairs $(X^{(i,m)}, X^{(j,m)}) \sim \Gamma_{\theta,ij}$ and minibatch draws $R_t^{(m)} \sim \eta_t$ are sampled i.i.d. across m.

Denote $\mu := \mathbb{E}[Z_1] = \mathbb{E}[\Delta_t^2]$ and $\bar{Z} := \frac{1}{M}\sum_{m=1}^M Z_m = \widehat{\mu}_{2,t}(\theta, ij)$. By Hoeffding's inequality for bounded i.i.d. random variables, for any $\varepsilon > 0$,

$$\Pr(\mu - \bar{Z} \ge \varepsilon) \le \exp\left(-\frac{2M\varepsilon^2}{(B_t^2 - 0)^2}\right) = \exp\left(-\frac{2M\varepsilon^2}{B_t^4}\right).$$

Set $\varepsilon := B_t^2 \sqrt{\frac{1}{2M} \log \frac{2}{\gamma_t}}$. Then

$$\Pr\left(\mu - \bar{Z} \ge B_t^2 \sqrt{\frac{1}{2M} \log \frac{2}{\gamma_t}}\right) \le \exp\left(-\log \frac{2}{\gamma_t}\right) = \frac{\gamma_t}{2} \le \gamma_t.$$

Equivalently, with probability at least $1-\gamma_t$,

$$\mathbb{E}[\Delta_t^2] = \mu \le \bar{Z} + B_t^2 \sqrt{\frac{1}{2M} \log \frac{2}{\gamma_t}} = \widehat{\mu}_{2,t}(\theta, ij) + B_t^2 \sqrt{\frac{1}{2M} \log \frac{2}{\gamma_t}},$$

which is exactly (32). □

Uniform sa-Cap Over a Finite Instantiated Family. Let $\widehat{\Theta} \subseteq \Theta$ be a finite set of instantiated priors, and let $\widehat{\mathcal{Q}} \subseteq \mathcal{Q}$ be a finite set of secret pairs. Define the index set $I_t := \widehat{\Theta} \times \widehat{\mathcal{Q}}$ and let $N_t := |I_t|$.

Corollary D.1 (Uniform sa-HUC cap over $\widehat{\Theta} \times \widehat{\mathcal{Q}}$)**.** *For each* $(\theta, (s_i, s_j)) \in I_t$*, compute*

$$\widehat{K}^2_{t,\theta,ij} \;:=\; \widehat{\mu}_{2,t}(\theta, ij) \;+\; B_t^2 \sqrt{\frac{1}{2M} \log \frac{2N_t}{\gamma_t}},$$

i.e., apply Lemma D.1 with calibration failure γ_t/N_t*. Then, with probability at least* $1-\gamma_t$*, simultaneously for all* $(\theta, (s_i, s_j)) \in I_t$*,*

$$\mathbb{E}[\Delta_t^2] \;\le\; \widehat{K}^2_{t,\theta,ij}.$$

Consequently, the uniform sa-HUC cap

$$\widehat{K}_t^2 \;:=\; \max_{(\theta,(s_i,s_j)) \in I_t} \widehat{K}^2_{t,\theta,ij}$$

satisfies $\mathbb{E}[\Delta_t^2] \le \widehat{K}_t^2$ *for all* $(\theta, (s_i, s_j)) \in I_t$ *on the same event.*

Proof. For each index $k = (\theta, (s_i, s_j)) \in I_t$, apply Lemma D.1 with failure probability γ_t/N_t. This yields

$$\Pr\Big(\mathbb{E}[\Delta_t^2] \le \widehat{K}^2_{t,\theta,ij}\Big) \;\ge\; 1 - \frac{\gamma_t}{N_t}.$$

By the union bound over the N_t indices,

$$\Pr\Big(\forall (\theta, (s_i, s_j)) \in I_t :\; \mathbb{E}[\Delta_t^2] \le \widehat{K}^2_{t,\theta,ij}\Big) \;\ge\; 1 - \sum_{k \in I_t} \frac{\gamma_t}{N_t} \;=\; 1 - \gamma_t.$$

On this same event, taking $\widehat{K}_t^2 := \max_{(\theta,(s_i,s_j)) \in I_t} \widehat{K}^2_{t,\theta,ij}$ gives $\mathbb{E}[\Delta_t^2] \le \widehat{K}_t^2$ simultaneously for all $(\theta, (s_i, s_j)) \in I_t$. □

D.3.2 Finite-Sample Bounds for HUC

We estimate a high-probability HUC cap $K_t(\delta_t)$ satisfying (31) via an empirical quantile bound based on the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality.

Fix $\theta \in \Theta$, $(s_i, s_j) \in \mathcal{Q}$ with $P^S_\theta(s_i), P^S_\theta(s_j) > 0$, a coupling $\Gamma_{\theta,ij} \in \Pi(P_\theta(X \mid S = s_i), P_\theta(X \mid S = s_j))$, and a subsampling rule η_t. Let Δ_t be the discrepancy count induced by $(X^{(i)}, X^{(j)}) \sim \Gamma_{\theta,ij}$ and $R_t \sim \eta_t$. Let $F(u) := \Pr(\Delta_t \le u)$ and $\widehat{F}_M$ be the empirical CDF from i.i.d. samples $\Delta_t^{(1)}, \dots, \Delta_t^{(M)}$.

Lemma D.2. *Let* $\Delta_t^{(1)}, \dots, \Delta_t^{(M)}$ *be i.i.d. samples of* Δ_t *with true CDF* F *and empirical CDF* $\widehat{F}_M$*. For any calibration failure* $\gamma_t \in (0,1)$*, define*

$$\varepsilon \;:=\; \sqrt{\frac{1}{2M} \log \frac{2}{\gamma_t}}.$$

Then, with probability at least $1-\gamma_t$*,*

$$\sup_{u \in \mathbb{R}} \big|\widehat{F}_M(u) - F(u)\big| \;\le\; \varepsilon.$$

Consequently, for any $\delta_t \in (0,1)$ *with* $\delta_t > \varepsilon$*, let* $\widehat{q}_{1-(\delta_t-\varepsilon)}$ *be an empirical* $(1-(\delta_t-\varepsilon))$*-quantile of* $\{\Delta_t^{(m)}\}_{m=1}^M$ *(e.g., the* $\lceil (1-(\delta_t-\varepsilon))M \rceil$*-th order statistic). Then on the same event,*

$$\Pr\big(\Delta_t \le \widehat{q}_{1-(\delta_t-\varepsilon)}\big) \;\ge\; 1 - \delta_t.$$

In particular, $K_t(\delta_t) := \widehat{q}_{1-(\delta_t-\varepsilon)}$ *is a valid tail cap at level* δ_t *for this fixed* $(\theta, (s_i, s_j))$ *and coupling* $\Gamma_{\theta,ij}$*.*

Proof. By the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality [25], for any $\varepsilon > 0$,

$$\Pr\left(\sup_{u\in\mathbb{R}} \left|\widehat{F}_M(u) - F(u)\right| > \varepsilon\right) \;\le\; 2e^{-2M\varepsilon^2}.$$

With $\varepsilon := \sqrt{\frac{1}{2M}\log\frac{2}{\gamma_t}}$, the right-hand side equals γ_t. Hence, with probability at least $1-\gamma_t$,

$$\sup_{u\in\mathbb{R}} \left|\widehat{F}_M(u) - F(u)\right| \;\le\; \varepsilon. \tag{33}$$

Assume (33) holds and fix any $\delta_t \in (0,1)$ with $\delta_t > \varepsilon$. Let $\widehat{q} := \widehat{q}_{1-(\delta_t-\varepsilon)}$ be an empirical $(1-(\delta_t-\varepsilon))$-quantile, i.e., $\widehat{F}_M(\widehat{q}) \ge 1-(\delta_t-\varepsilon)$. Then, using (33),

$$F(\widehat{q}) \;\ge\; \widehat{F}_M(\widehat{q}) - \varepsilon \;\ge\; \big(1-(\delta_t-\varepsilon)\big) - \varepsilon \;=\; 1-\delta_t.$$

Since $F(\widehat{q}) = \Pr(\Delta_t \le \widehat{q})$, we conclude that on the same event,

$$\Pr(\Delta_t \le \widehat{q}) \;\ge\; 1-\delta_t.$$

Therefore, with probability at least $1-\gamma_t$ over the Monte Carlo sampling, the choice $K_t(\delta_t) := \widehat{q}_{1-(\delta_t-\varepsilon)}$ satisfies $\Pr(\Delta_t \le K_t(\delta_t)) \ge 1-\delta_t$, as claimed. □

Uniform Tail Cap over a Finite Instantiated Family. Let $\widehat{\Theta} \subseteq \Theta$ and $\widehat{\mathcal{Q}} \subseteq \mathcal{Q}$ be finite sets, and let $I_t := \widehat{\Theta} \times \widehat{\mathcal{Q}}$ with $N_t := |I_t|$.

Corollary D.2. *For each $(\theta,(s_i,s_j)) \in I_t$, apply Lemma D.2 with calibration failure γ_t/N_t, i.e., with*

$$\varepsilon \;:=\; \sqrt{\frac{1}{2M}\log\frac{2N_t}{\gamma_t}}.$$

Assume $\delta_t > \varepsilon$. Define

$$\widehat{K}_{t,\theta,ij}(\delta_t) \;:=\; \widehat{q}^{(\theta,ij)}_{1-(\delta_t-\varepsilon)}, \qquad \widehat{K}_t(\delta_t) \;:=\; \max_{(\theta,(s_i,s_j))\in I_t} \widehat{K}_{t,\theta,ij}(\delta_t),$$

where $\widehat{q}^{(\theta,ij)}_{1-(\delta_t-\varepsilon)}$ is the empirical $(1-(\delta_t-\varepsilon))$-quantile computed from the Monte Carlo samples generated for that $(\theta,(s_i,s_j))$. Then, with probability at least $1-\gamma_t$ over the Monte Carlo sampling (simultaneously across all indices in I_t),

$$\Pr\big(\Delta_t \le \widehat{K}_{t,\theta,ij}(\delta_t)\big) \;\ge\; 1-\delta_t \quad \forall(\theta,(s_i,s_j)) \in I_t,$$

and hence $\Pr(\Delta_t \le \widehat{K}_t(\delta_t)) \ge 1-\delta_t$ holds uniformly over $(\theta,(s_i,s_j)) \in I_t$ on the same event.

Proof. Fix t and the finite index set $I_t = \widehat{\Theta}\times\widehat{\mathcal{Q}}$ with $N_t := |I_t|$. For each index $a = (\theta,(s_i,s_j)) \in I_t$, let $\{\Delta_t^{(m)}(a)\}_{m=1}^M$ denote the M i.i.d. Monte Carlo draws generated under the joint randomness $(X^{(i)},X^{(j)}) \sim \Gamma_{\theta,ij}$ and $R_t \sim \eta_t$, and let $\widehat{K}_{t,a}(\delta_t)$ be the resulting empirical tail cap defined in the corollary.

For each $a \in I_t$, define the "calibration-success" event

$$E_a \;:=\; \Big\{\Pr\big(\Delta_t \le \widehat{K}_{t,a}(\delta_t)\big) \;\ge\; 1-\delta_t\Big\},$$

where the probability is with respect to the true joint law of Δ_t for that fixed a (i.e., over $(X^{(i)},X^{(j)})$ and R_t), and the randomness in E_a is only the Monte Carlo sampling used to construct $\widehat{K}_{t,a}(\delta_t)$.

Apply Lemma D.2 to each fixed $a \in I_t$ with calibration failure γ_t/N_t. With

$$\varepsilon = \sqrt{\frac{1}{2M}\log\frac{2N_t}{\gamma_t}},$$

and assuming $\delta_t > \varepsilon$, the lemma yields

$$\Pr_{\mathrm{MC}}(E_a) \;\ge\; 1-\frac{\gamma_t}{N_t}, \qquad \forall a \in I_t,$$

where $\Pr_{\mathrm{MC}}$ denotes probability over the Monte Carlo sampling for index a.

By a union bound,

$$\Pr_{\mathrm{MC}}\Big(\bigcap_{a\in I_t} E_a\Big) \;\ge\; 1-\sum_{a\in I_t}\Pr_{\mathrm{MC}}(E_a^c) \;\ge\; 1-\sum_{a\in I_t}\frac{\gamma_t}{N_t} \;=\; 1-\gamma_t.$$

Therefore, with probability at least $1-\gamma_t$ (simultaneously over all indices $a \in I_t$), we have $\Pr(\Delta_t \le \widehat{K}_{t,a}(\delta_t)) \ge 1-\delta_t$ for every $a = (\theta,(s_i,s_j)) \in I_t$, which proves the first claim.

Finally, define $\widehat{K}_t(\delta_t) := \max_{a\in I_t}\widehat{K}_{t,a}(\delta_t)$. On the event $\bigcap_{a\in I_t} E_a$, for any fixed a we have the set inclusion $\{\Delta_t \le \widehat{K}_{t,a}(\delta_t)\} \subseteq \{\Delta_t \le \widehat{K}_t(\delta_t)\}$, hence

$$\Pr\big(\Delta_t \le \widehat{K}_t(\delta_t)\big) \;\ge\; \Pr\big(\Delta_t \le \widehat{K}_{t,a}(\delta_t)\big) \;\ge\; 1-\delta_t.$$

This holds for every $a \in I_t$ on the same event, completing the proof. □

Suppose a tail cap $K_t(\delta_t)$ is used in HUC-style accounting at each round $t = 1,\dots,T$. Let E_t denote the runtime cap-validity event for the coupled draw and minibatch draw at round t,

$$E_t \;:=\; \{\Delta_t \le K_t(\delta_t)\},$$

where the probability is over $(X^{(i)},X^{(j)})$ and $R_t \sim \eta_t$. If $\Pr(E_t) \ge 1-\delta_t$ holds for each round, then by a union bound,

$$\Pr\Big(\bigcap_{t=1}^T E_t\Big) \;\ge\; 1-\sum_{t=1}^T \delta_t.$$

If, additionally, each $K_t(\delta_t)$ is obtained from Monte Carlo estimation with calibration failure probability γ_t (e.g., via Corollary D.2), then the event that *all* per-round tail guarantees hold across all rounds has probability at least

$$1-\sum_{t=1}^T \delta_t \;-\; \sum_{t=1}^T \gamma_t. \tag{34}$$

For sa-HUC, there is no tail parameter δ_t; only the calibration failures γ_t appear, and the analogous across-round union bound yields probability at least $1-\sum_{t=1}^T \gamma_t$.

D.4 Closed-Form Bounds under Record-Level Structure

This section provides analytic bounds for $\mathbb{E}[\Delta_t^2]$ and for the tail behavior of Δ_t under record-level i.i.d. structure. Assume the dataset consists of N records $X = (X_1, \dots, X_N)$ and, conditional on $(S = s, \theta)$, records are i.i.d.:

$$X_k \mid (S = s, \theta) \;\sim\; P^{\mathrm{rec}}_{\theta,s}, \qquad k = 1, \dots, N. \tag{35}$$

Recordwise Maximal Coupling and Mismatch Probability. Fix $\theta \in \Theta$ and $(s_i, s_j) \in \mathcal{Q}$. Let $p := P^{\mathrm{rec}}_{\theta,s_i}$ and $q := P^{\mathrm{rec}}_{\theta,s_j}$. There exists a *maximal coupling* of a single record $(X_k^{(i)}, X_k^{(j)})$ with marginals p and q such that

$$\Pr\big[X_k^{(i)} \neq X_k^{(j)}\big] \;=\; \tau_{\theta,ij} \;:=\; \mathrm{TV}(p, q), \tag{36}$$

where $\mathrm{TV}(\cdot,\cdot)$ denotes total variation distance. Coupling records independently across k yields mismatch indicators $M_k := \mathbf{1}\{X_k^{(i)} \neq X_k^{(j)}\}$ that are i.i.d. Bernoulli$(\tau_{\theta,ij})$.

Minibatch Discrepancy: WR vs. WOR. Let η_t select a minibatch $\mathsf{I}_t \subseteq \{1, \dots, N\}$ of size B_t, and recall $\Delta_t = \sum_{k \in \mathsf{I}_t} M_k$.

WR (or i.i.d. index draws). If η_t draws indices $I_1, \dots, I_{B_t}$ independently (with replacement), then

$$\Delta_t \;\sim\; \mathrm{Binomial}\big(B_t, \tau_{\theta,ij}\big), \tag{37}$$

and hence

$$\mathbb{E}[\Delta_t] = B_t \tau_{\theta,ij}, \tag{38}$$

$$\mathbb{E}[\Delta_t^2] = B_t \tau_{\theta,ij}(1 - \tau_{\theta,ij}) + (B_t \tau_{\theta,ij})^2. \tag{39}$$

WOR. If η_t samples I_t uniformly without replacement, then conditional on the total mismatch count $S_N := \sum_{k=1}^N M_k$, the discrepancy Δ_t has a hypergeometric distribution:

$$\Delta_t \,\big|\, S_N \;\sim\; \mathrm{Hypergeometric}\big(N, S_N, B_t\big).$$

In particular, $\mathbb{E}[\Delta_t] = B_t \tau_{\theta,ij}$ still holds, and standard concentration inequalities for sampling without replacement (e.g., Serfling/Hoeffding-type bounds) yield binomial-style tails. Consequently, the tail cap derived below for the with-replacement case remains a valid (conservative) choice under without-replacement sampling.

Resulting sa-Cap from Total-Variation Control. In the with-replacement case, (38) implies $\mathbb{E}[\Delta_t^2] = G(B_t, \tau_{\theta,ij})$ with

$$G(B_t, \tau) \;:=\; B_t \tau (1 - \tau) + (B_t \tau)^2.$$

Thus one may take the sa-HUC cap

$$\overline{K}_t^2 \;:=\; \sup_{\theta \in \Theta} \sup_{(s_i,s_j) \in \mathcal{Q}} G\big(B_t, \tau_{\theta,ij}\big), \tag{40}$$

whenever $\tau_{\theta,ij}$ (or an upper bound on it) is available uniformly over θ and (s_i, s_j).

Tail Bounds and a High-Probability HUC. Under (37), Hoeffding's inequality gives, for any $\lambda > 0$,

$$\Pr\big(\Delta_t \geq \mathbb{E}[\Delta_t] + \lambda\big) \;\leq\; \exp\Big(-\frac{2\lambda^2}{B_t}\Big).$$

Therefore, for any tail level $\delta_t \in (0, 1)$, the choice

$$K_t(\delta_t) \;:=\; \sup_{\theta \in \Theta} \sup_{(s_i,s_j) \in \mathcal{Q}} \Big(B_t \tau_{\theta,ij} + \sqrt{\tfrac{B_t}{2} \log \tfrac{1}{\delta_t}}\Big) \tag{41}$$

ensures $\Pr(\Delta_t \leq K_t(\delta_t)) \geq 1 - \delta_t$ for the with-replacement minibatch rule. Moreover, the same cap remains valid (conservatively) for common without-replacement rules by standard concentration bounds for sampling without replacement.

D.5 Localized Influence of Secrets

In some scenarios, the secret affects only a limited subset of records. Fix $\theta \in \Theta$ and $(s_i, s_j) \in \mathcal{Q}$, and suppose there exists a (possibly random) index set $A \subseteq \{1, \dots, N\}$ and a coupling $\Gamma_{\theta,ij}$ such that, for $(X^{(i)}, X^{(j)}) \sim \Gamma_{\theta,ij}$,

$$X_k^{(i)} = X_k^{(j)} \quad \text{for all } k \notin A, \qquad \text{and} \qquad |A| \leq D_{\max} \text{ a.s.}$$

Then, for any minibatch rule selecting B_t indices, the discrepancy count satisfies the deterministic bound

$$\Delta_t \;\leq\; \min\{B_t, D_{\max}\}.$$

Accordingly, one may take the HUC and sa-HUC caps

$$K_t \;:=\; \min\{B_t, D_{\max}\}, \qquad \overline{K}_t^2 \;:=\; \min\{B_t^2, D_{\max}^2\}. \tag{42}$$

If the distribution of $|A|$ is available (analytically or via simulation), then sharper moment or tail caps can be obtained by conditioning on $|A|$ and applying the corresponding sampling-with/without-replacement formulas.

D.6 Instantiation of HUC and sa-HUC

To instantiate the HUC/sa-HUC accountants, it suffices to specify (a) a coupling family $\{\Gamma_{\theta,ij}\}_{\theta \in \Theta, (s_i,s_j) \in \mathcal{Q}}$ consistent with the scenario assumptions and (b) the minibatch selection rule η_t. Given these, one obtains admissible caps either by analytic control of the induced discrepancy distribution (e.g., Section D.4 or Section D.5) or by Monte Carlo estimation (Section D.3), with finite-sample calibration guarantees given by Lemma D.1 and Corollary D.1 for sa-HUC, and by Lemma D.2 and Corollary D.2 for HUC tail caps, and across-round bookkeeping in (34). The resulting caps $K_t(\delta_t)$ and/or $\overline{K}_t^2$ can then be used directly in the corresponding HUC/sa-HUC accounting theorems.

The sa-HUC accounting accepts any cap $\overline{K}_t^2$ such that $\mathbb{E}[\Delta_t^2] \leq \overline{K}_t^2$ uniformly over $\theta \in \Theta$ and $(s_i, s_j) \in \mathcal{Q}$ (or over

Algorithm 4 Monte Carlo discrepancy-cap estimation at iteration t

Require: Prior $\theta \in \Theta$; secret pair $(s_i, s_j) \in \mathcal{Q}$ with $P_\theta^S(s_i), P_\theta^S(s_j) > 0$; coupling sampler $\mathsf{Couple}_\theta(s_i, s_j)$ returning $(X^{(i)}, X^{(j)}) \sim \Gamma_{\theta,ij}$ with marginals $X^{(i)} \sim P_\theta(\cdot \mid S = s_i)$ and $X^{(j)} \sim P_\theta(\cdot \mid S = s_j)$; minibatch subsampling rule η_t selecting B_t indices; Monte Carlo size M; HUC tail level $\delta_t \in (0,1)$; calibration failure $\gamma_t \in (0,1)$.

Ensure: A tail cap $\widehat{K}_t(\delta_t)$ for HUC and a second-moment cap $\widehat{K}_t^2$ for sa-HUC.

1: **for** $m = 1, \dots, M$ **do**

2: Sample $(X^{(i,m)}, X^{(j,m)}) \leftarrow \mathsf{Couple}_\theta(s_i, s_j)$.

3: Sample minibatch indices $R_t^{(m)} = (I_1^{(m)}, \dots, I_{B_t}^{(m)}) \sim \eta_t$.

4: Compute the discrepancy count

$$\Delta_t^{(m)} \leftarrow \#\{b \in \{1, \dots, B_t\} : x^{(i,m)}_{I_b^{(m)}} \neq x^{(j,m)}_{I_b^{(m)}}\},\ 0 \le \Delta_t^{(m)} \le B_t.$$

5: **end for**

6: Set $\varepsilon \leftarrow \sqrt{\frac{1}{2M} \log \frac{2}{\gamma_t}}$.

7: **(sa-HUC cap).** Compute $\widehat{\mu}_{2,t} \leftarrow \frac{1}{M} \sum_{m=1}^M (\Delta_t^{(m)})^2$ and set

$$\widehat{K}_t^2 \leftarrow \widehat{\mu}_{2,t} + B_t^2 \varepsilon.$$

8: **(HUC tail cap).** If $\delta_t > \varepsilon$, set $\widehat{K}_t(\delta_t)$ to the empirical $(1 - (\delta_t - \varepsilon))$-quantile of $\{\Delta_t^{(m)}\}_{m=1}^M$; otherwise set $\widehat{K}_t(\delta_t) \leftarrow B_t$.

9: **return** $(\widehat{K}_t(\delta_t), \widehat{K}_t^2)$.

a finite instantiated family as in Corollary D.1). If $\overline{K}_t^2$ is obtained via Monte Carlo (Algorithm 4), then Lemma D.1 (and Corollary D.1 with a union bound) implies that $\mathbb{E}[\Delta_t^2] \le \widehat{K}_t^2$ holds with probability at least $1 - \gamma_t$ over the cap-estimation randomness.

For HUC accounting, we use tail caps $K_t(\delta_t)$ such that $\Pr(\Delta_t \le K_t(\delta_t)) \ge 1 - \delta_t$ uniformly over $\theta \in \Theta$ and $(s_i, s_j) \in \mathcal{Q}$. If $K_t(\delta_t)$ is obtained via Monte Carlo (Algorithm 4), then Lemma D.2 (and Corollary D.2 with a union bound) implies that, with probability at least $1 - \gamma_t$ over the cap-estimation randomness, the tail guarantee holds. Across rounds, a union bound yields overall probability at least $1 - \sum_{t=1}^T \delta_t - \sum_{t=1}^T \gamma_t$ as in (34).

D.7 Example Visualization of HUC and sa-HUC

Figure 4 visualizes the difference between HUC and sa-HUC. For each slice direction u_i, the plotted radius is the (one-round) per-slice envelope value $R_i = R_\alpha(\zeta_\sigma, z_i)$ induced by the corresponding directional shift cap z_i. HUC uses a worst-case cap over minibatch randomness, while sa-HUC incorporates subsampling by bounding the envelope *on average* over minibatch selection; consequently the sa-HUC profile is uniformly smaller and typically strictly so when the per-round discrepancy varies with the minibatch. The rings visualize the two SRPP aggregations applied to the same discrete profile: Ave-SRPP averages $\{R_i\}$ over ω, whereas Joint-SRPP applies a log-moment aggregation that emphasizes larger directions (and is always at least the Ave value). Increasing m refines the discrete approximation of the directional profile.

E SRPP-SGD with Poisson Subsampling

Our definitions allow $\rho \in \{\mathsf{WR}, \mathsf{WOR}, \mathsf{Poisson}\}$, where under $\rho = \mathsf{Poisson}$ the minibatch size $B_t(R_t) = |\mathsf{I}_t(R_t)|$ is random. The SRPP-SGD accounting theorems in Sec. 5.1–5.2 apply to $\rho = \mathsf{Poisson}$ as well, provided the per-round HUC/sa-HUC caps are formed using a *normalized discrepancy* rather than a fixed $1/B$ factor. Concretely, define

$$\kappa_t(X, X'; R_t) := \frac{K_t(X, X'; R_t)}{B_t(R_t)} \mathbf{1}\{B_t(R_t) \ge 1\},$$

and assume either (i) the implementation enforces $B_t(R_t) \ge 1$ a.s. (e.g., resampling when $B_t = 0$), or (ii) the update is defined deterministically when $B_t = 0$. Then Proposition 5.1 and Proposition 5.2 remain valid after replacing the deterministic bound K_t/B_t by a uniform cap on κ_t (for HUC) or by a uniform bound on $\mathbb{E}_{\eta_t}[\kappa_t^2]$ (for sa-HUC). All subsequent envelope and composition results are unchanged.

F Per-Layer Gradient Clipping

DP-SGD with *per-layer gradient clipping* [26] is a structured variant of DP-SGD, where gradient clipping and noise addition are applied independently to disjoint parameter blocks, rather than globally across the entire model.

F.1 DP-SGD with Per-Layer Gradient Clipping

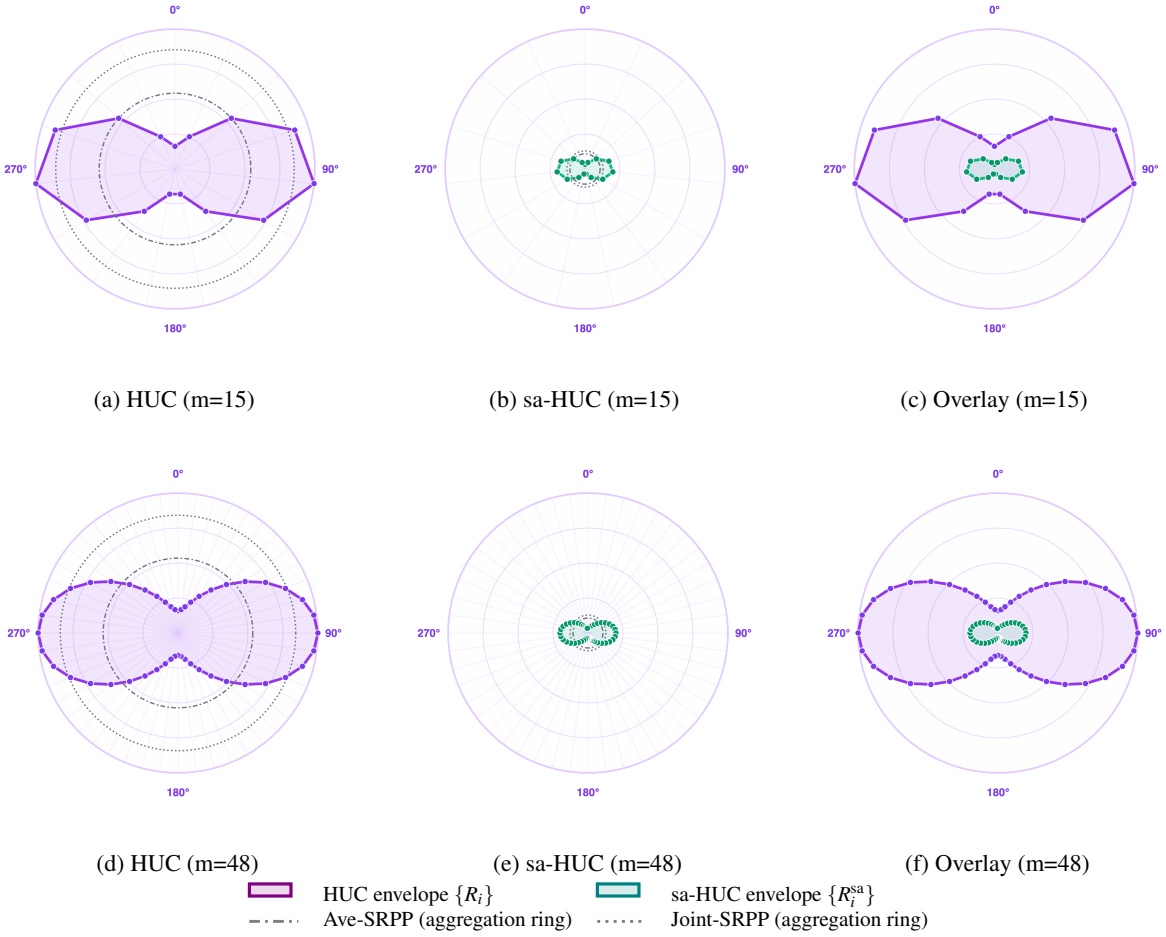
We briefly recall the standard DP-SGD [1] with per-example gradient clipping. Given a minibatch B of size b, DP-SGD computes per-example gradients $g_i = \nabla_\zeta \ell(\zeta; x_i)$, clips each to an ℓ_2 bound C, averages the clipped gradients, and adds Gaussian noise calibrated to C:

$$\tilde{g}_i = g_i \cdot \min\Big\{1, \tfrac{C}{\|g_i\|_2}\Big\}, \quad \hat{g} = \frac{1}{b} \sum_{i \in B} \tilde{g}_i + \mathcal{N}\Big(0, \tfrac{\sigma^2 C^2}{b^2} I\Big),$$

followed by the SGD update:

$$\zeta \leftarrow \zeta - \eta \cdot \hat{g}.$$

DP-SGD with per-layer (gradient) clipping partitions parameters into E disjoint layers $\zeta = (\zeta^{(1)}, \dots, \zeta^{(E)})$, and each



(a) HUC (m=15) (b) sa-HUC (m=15) (c) Overlay (m=15)

(d) HUC (m=48) (e) sa-HUC (m=48) (f) Overlay (m=48)

Figure 4: Directional envelope profiles for HUC vs. sa-HUC under a uniform discrete slice profile ω. Each panel plots the per-direction Rényi envelope values $\{R_i\}_{i=1}^m$ on a polar axis (radius = envelope magnitude, angle = slice direction). Panels (a,d) show the worst-case HUC envelope; (b,e) show the subsampling-aware envelope (sa-HUC); (c,f) overlay the two profiles on the same radial scale. The gray rings indicate the corresponding aggregated SRPP costs (Ave-SRPP and Joint-SRPP) computed from the discrete profile. We use $\alpha = 2.5$ and subsampling rate $q = 0.20$, and compare $m = 15$ versus $m = 48$ slice directions.

per-example block-gradient is clipped using a block-specific bound C_e:

$$\tilde{g}_i^{(e)} = g_i^{(e)} \cdot \min\left\{1, \frac{C_e}{\|g_i^{(e)}\|_2}\right\}, \; \hat{g}^{(e)} = \frac{1}{b}\sum_{i\in B} \tilde{g}_i^{(e)} + \mathcal{N}\left(0, \frac{\sigma^2 C_e^2}{b^2} I\right),$$

and then $\hat{g} = (\hat{g}^{(1)}, \ldots, \hat{g}^{(E)})$ is used in the update.

F.2 Layer-wise Lipschitz Assumption

We view $\zeta = (\zeta^{(1)}, \ldots, \zeta^{(E)})$ as the concatenation of layer parameters, which induces the orthogonal direct-sum decomposition $\mathbb{R}^d = \bigoplus_{e=1}^E \mathbb{R}^{d_e}$. Let $\{\mathsf{P}_e\}_{e=1}^E$ denote the canonical orthogonal projector onto the e-th block. Consider a generic (possibly preconditioned) SGD-type update written in the form

$$f_t(x, y_{<t}; R_t) \;=\; \xi_{t-1} \;-\; A_t\, \bar{g}_t(x; R_t),$$

where $\bar{g}_t(\cdot)$ is the (clipped) minibatch gradient estimator, built from per-layer clipped per-example gradients, and A_t is a deterministic linear map.

Assumption F.1 is a layer-wise analogue of Assumption 5.2.

Assumption F.1 (Layer-wise Lipschitz Updates)**.** *For each iteration t and each block $e \in \{1, \ldots, E\}$, there exists a finite constant $L_{t,e} \geq 0$ such that*

$$\|\mathsf{P}_e A_t^\top u\|_2 \leq L_{t,e} \|\mathsf{P}_e u\|_2, \qquad \forall u \in \mathcal{U}.$$

For vanilla SGD with step size η_t, we have $A_t = \eta_t I$, so the Lipschitz constant is $L_{t,e} = \eta_t$ for all $e = 1, \ldots, E$. For SGD with per-layer clipping with $\eta_{t,e}$ for each e, we have $A_t = \mathsf{blkdiag}(\eta_{t,1} I, \ldots, \eta_{t,E} I)$, so the Lipschitz constant for each block e is $L_{t,e} = \eta_{t,e}$.

F.3 HUC and sa-HUC under per-layer clipping

Fix time t, a slice profile $U = \{u_i\}_{i=1}^m$, and let B_t be the mini-batch size and K_t the discrepancy cap used in our main HUC construction.

A convenient way to express the per-slice cap in the per-layer clipped setting is via the same blockwise template as Corollary H.1 (linear, blockwise-clipped case). Specialized to "layers as blocks," this yields the slice-dependent form

$$h_{t,i}^{\text{layer}} \;=\; \left(\tfrac{2K_t}{B_t}\right)^2 \left(\sum_{\ell=1}^{E} C_\ell \left\|P_\ell A_t^\top u_i\right\|_2\right)^2, \qquad i \in [m].$$

Using the layer-wise Lipschitz bound $\|P_\ell A_t^\top u_i\|_2 \le L_{t,\ell}\|P_\ell u_i\|_2$, we obtain the fully explicit and easy-to-evaluate envelope

$$h_{t,i}^{\text{layer}} \;\le\; \left(\tfrac{2K_t}{B_t}\right)^2 \left(\sum_{\ell=1}^{E} C_\ell L_{t,\ell} \|P_\ell u_i\|_2\right)^2.$$

The corresponding subsampling-aware case (sa-HUC), which underpins sa-SRPP-SGD in Section 5.2, is obtained by replacing the worst-case $\overline{K}_t^2$ factor by its subsampling-aware counterpart:

$$\begin{aligned} h_{t,i}^{\text{sa,layer}} \;&=\; \left(\tfrac{2}{B_t}\right)^2 \overline{K}_t^2 \left(\sum_{\ell=1}^{E} C_\ell \left\|\mathsf{P}_\ell A_t^\top u_i\right\|_2\right)^2 \\ &\le\; \left(\tfrac{2}{B_t}\right)^2 \overline{K}_t^2 \left(\sum_{\ell=1}^{E} C_\ell L_{t,\ell} \|\mathsf{P}_\ell u_i\|_2\right)^2. \end{aligned}$$

Unlike the "global Lipschitz + global clipping" special case (where the cap can collapse to a slice-independent expression), per-layer clipping makes $\|P_\ell u_i\|_2$ appear explicitly, so Ave- vs. Joint-type aggregation across slices can differ in SRPP/sa-SRPP-SGD, even though the clipping itself is performed on the true gradients and is not directional.

F.4 Numerical Example

We repeat the main-body SGD privatization experiment under the same Pufferfish scenario, training protocol, and subsampling rate ($\eta = 0.02$), but replace global clipping by *parameter-tensor blockwise* clipping ("per-layer"). Concretely, we treat each trainable parameter tensor in `model.named_parameters()` as one block, yielding $L = 23$ blocks for `gep_resnet22` (convolution weights plus the final linear weight and bias). We use non-uniform clipping thresholds with the $\sqrt{\text{dim}}$ rule:

$$C_\ell \;=\; C_{\text{base}} \sqrt{\frac{d_\ell}{\sum_j d_j}}, \quad C_{\text{base}} = 5.0,$$

where d_ℓ is the number of parameters in block ℓ.

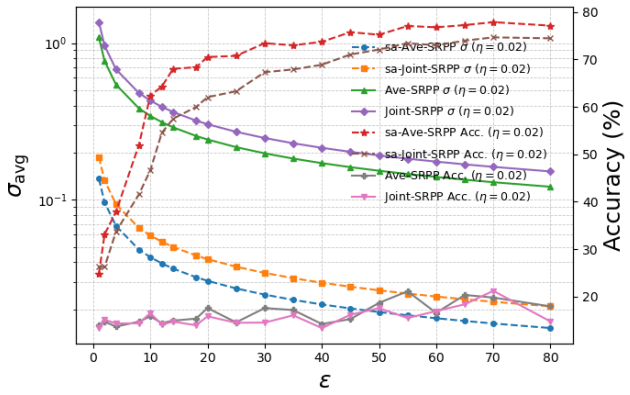


Figure 5: Per-block ("per-layer") clipping experiment under the same setup as the main-body SGD study ($\eta = 0.02$). Left axis: the calibrated reference noise scale σ_{avg} (at C_{base}), where per-block noise obeys $\sigma_\ell / C_\ell = \sigma_{\text{avg}} / C_{\text{base}}$. Right axis: test accuracy.

For each target SRPP/sa-SRPP budget ε, we first compute the aggregate accountant value H_{total} (Ave or Joint) and calibrate a *single* reference noise scale

$$\sigma_{\text{avg}}^2(\varepsilon) \;=\; \frac{\alpha}{2\varepsilon} H_{\text{total}}.$$

This σ_{avg} is *not* an average over per-block standard deviations; it is the standard deviation of the Gaussian noise added to the *batch-averaged clipped gradient*, expressed at the reference clipping scale C_{base}. Per-block noise is then set by

$$\sigma_\ell \;=\; \sigma_{\text{avg}} \cdot \frac{C_\ell}{C_{\text{base}}}, \quad \text{so that} \quad \frac{\sigma_\ell}{C_\ell} = \frac{\sigma_{\text{avg}}}{C_{\text{base}}} \;\text{ for all } \ell.$$

Figure 5 reports σ_{avg} and test accuracy versus ε. Subsampling-aware accounting (sa-Ave/sa-Joint) certifies the same target budget with smaller σ_{avg} than the corresponding worst-case calibration (Ave/Joint), translating into improved accuracy at matched ε.

In addition, Joint-type accounting is more conservative than Ave-type accounting: because it controls the *joint* leakage across slices (equivalently, a log-moment / log-sum-exp aggregation over directions), it yields a larger aggregate H_{total} and therefore requires a larger calibrated σ_{avg} at the same target budget ε.

G Relationship to Group DP-SGD

In this section, we characterize the relationship and the difference between our SRPP/sa-SRPP and the group DP-SGD. Please note that the group DP here is different from the *worst-case group privacy* described in App. A.4.

The dependence on the discrepancy cap K_t (worst-case) and its subsampling-aware analogue $\mathbb{E}_\eta[K^2]$ in our SRPP/sa-SRPP privatization of SGD naturally raises the question: *can*

group DP-SGD, with an appropriate group size, serve as a reliable proxy for SRPP-SGD when the protected information is a specific Pufferfish secret? In this appendix, we clarify the relationship between SRPP-/sa-SRPP-SGD and group DP-SGD, and explain why the latter cannot generally substitute for the former when one cares about secret-level Pufferfish guarantees.

Our SRPP- and sa-SRPP-SGD mechanisms are calibrated directly from the HUC and sa-HUC envelopes developed in Sections 5.1 and 5.2, respectively. For a fixed Rényi order $\alpha > 1$ and a target sliced privacy budget ε, the corresponding theorems give an upper bound on the per-run SRPP or sa-SRPP cost in terms of the HUC/sa-HUC sequence $\{h_t\}$ and the per-step noise covariances $\{\Sigma_t\}$. In our experiments we specialize to isotropic Gaussian noise $\Sigma_t = \sigma^2 I$ and, for each desired ε, solve the resulting bound for a noise scale $\sigma(\varepsilon)$ that guarantees the prescribed SRPP or sa-SRPP budget. This calibration is explicitly secret-aware: H_{total} depends on the Pufferfish scenario $(\mathcal{S}, Q, \Theta)$ through the discrepancy caps K_t and subsampling-aware discrepancy caps $\mathbb{E}_\eta[K_t^2]$, which, for each iteration t, bound the ℓ_2–distance between the clipped, averaged gradient updates under any admissible secret pair $(s_i, s_j) \in Q$ and any coupling of their corresponding data-generating distributions.

In contrast, standard group DP-SGD is calibrated with respect to a *dataset-level* adjacency relation

$$X \sim_k X' \iff d_H(X, X') \le k,$$

where d_H is Hamming distance and $k \in \mathbb{N}$ is a chosen group size. A mechanism M is $(\varepsilon_g, \delta_g)$-*group-DP* for group size k if

$$\Pr[\mathcal{M}(X) \in E] \le e^{\varepsilon_g} \Pr[\mathcal{M}(X') \in E] + \delta_g, \ \forall E \subseteq \mathcal{Y}, \ \forall X \sim_k X'. \tag{43}$$

In DP-SGD, one typically fixes k and uses a moments accountant to obtain $(\varepsilon_g, \delta_g)$ as a function of k, the sampling probability $q = B_t/N$, the number of steps T, and the per-step noise scale σ_{DP}.

A natural question is whether group DP-SGD can be used as a proxy for SRPP-/sa-SRPP-SGD by choosing a group size k that reflects the secret-induced discrepancy at iteration t. However, k in group DP is a *dataset-level* Hamming distance (an integer), whereas our sa-HUC uses a *minibatch-level* subsampling-aware discrepancy cap $\overline{K}_t^2$ satisfying $\mathbb{E}_{\eta_t}[\Delta_t^2] \le \overline{K}_t^2$. A dimensionally compatible proxy would be $k \approx \sqrt{\overline{K}_t^2}$ (or more conservatively $k \approx \lceil \sqrt{\overline{K}_t^2} \rceil$), but even with such a choice, group DP-SGD does not generally yield a certified SRPP/sa-SRPP guarantee for an arbitrary Pufferfish scenario.

In this section, we clarify two points:

- There is no general result that, given a group-DP guarantee (43) for some group size k, produces an (α, ε)-SRPP or sa-SRPP guarantee for an arbitrary Pufferfish scenario $(\mathcal{S}, Q, \Theta)$. In particular, even if one chooses a *proxy* group size $k \approx \lceil \sqrt{\overline{K}_t^2} \rceil$ (or uses a tail cap $k = K_t(\delta_t)$) motivated by our HUC/sa-HUC analysis, there is no closed-form or universal mapping from $(\varepsilon_g, \delta_g, k)$ to an SRPP/sa-SRPP budget ε.
- Even when we fix a specific Pufferfish scenario $(\mathcal{S}, Q, \Theta)$ and choose a proxy group size $k \approx \lceil \sqrt{\overline{K}_t^2} \rceil$ (or $k = K_t(\delta_t)$), the resulting $(\varepsilon_g, \delta_g)$ group-DP-SGD guarantee is in general not a certified conservative bound for our (α, ε) SRPP/sa-SRPP guarantees, and there is no theorem ensuring that this group-DP guarantee upper-bounds the leakage measured in the SRPP/sa-SRPP sense.

G.1 No Universal Mapping

We first formalize the absence of a universal conversion from group-DP guarantees to SRPP/sa-SRPP guarantees.

Definition G.1 (Hypothetical DP→SRPP conversion)**.** *A DP-to-SRPP conversion map is a function*

$$F : (0, \infty) \times [0, 1] \times \mathbb{N} \to (0, \infty]$$

that, for given group-DP parameters $(\varepsilon_g, \delta_g)$ and group size k, assigns an SRPP budget $F(\varepsilon_g, \delta_g, k)$ with the following property: for every mechanism $\mathcal{M}$ and every Pufferfish scenario $(\mathcal{S}, Q, \Theta)$, if $\mathcal{M}$ is $(\varepsilon_g, \delta_g)$-group-DP with group size k, then for all $\theta \in \Theta$ and all $(s_i, s_j) \in Q$ the SRPP/sa-SRPP divergence between the distributions of $\mathcal{M}$ under $(S = s_i, \theta)$ and $(S = s_j, \theta)$ is bounded by $F(\varepsilon_g, \delta_g, k)$.

Our first result is that such a universal conversion map cannot exist.

Proposition G.1. *There is no function F satisfying Definition G.1, even if we restrict to pure group-DP ($\delta_g = 0$) and to finite secret sets $\mathcal{S}$.*

Proof of Proposition G.1. Fix any candidate function F. We construct a counterexample.

Let the dataset space be $\mathcal{X} = \{X_0, X_1\}$, and choose $k \in \mathbb{N}$ so that $d_H(X_0, X_1) > k$, where d_H is the Hamming distance. Under the group adjacency relation restricted to $\mathcal{X} = \{X_0, X_1\}$, there are no distinct adjacent pairs: the only pairs with $d_H(X, X') \le k$ are (X_0, X_0) and (X_1, X_1). Hence, for any mechanism $\mathcal{M}$ and any $(\varepsilon_g, \delta_g)$, the group-DP inequality (43) is trivially satisfied, so $\mathcal{M}$ is $(\varepsilon_g, \delta_g)$-group-DP for all choices of $(\varepsilon_g, \delta_g)$ and k.

Now define a Pufferfish scenario with two secrets $\mathcal{S} = \{0, 1\}$, a single prior $\Theta = \{\theta\}$, and a deterministic data-generating process

$$P_\theta(X = X_0 \mid S = 0) = 1, \quad P_\theta(X = X_1 \mid S = 1) = 1.$$

In other words, the secret value deterministically selects which of the two datasets is realized.

Consider the mechanism $\mathcal{M}$ that simply reveals the secret:

$$\mathcal{M}(X_0) = 0, \quad \mathcal{M}(X_1) = 1.$$

As argued above, $\mathcal{M}$ is $(\varepsilon_g, \delta_g)$-group-DP for *all* $(\varepsilon_g, \delta_g)$ and all k, because there are no nontrivial adjacent dataset pairs.

However, the SRPP (and RPP/PP) divergence between the output distributions under $S = 0$ and $S = 1$ is infinite: for any measurable set E containing 0 but not 1 we have

$$\Pr[\mathcal{M}(X) \in E \mid S = 0] = 1, \qquad \Pr[\mathcal{M}(X) \in E \mid S = 1] = 0.$$

In particular, the Rényi divergence of any order $\alpha > 1$ between the distributions of $\mathcal{M}(X)$ given $S = 0$ and $S = 1$ is $+\infty$ (i.e., the two output laws are mutually singular), so no finite SRPP budget can hold.

This contradicts the existence of a finite-valued $F(\varepsilon_g, \delta_g, k)$ that would upper-bound SRPP leakage whenever $\mathcal{M}$ is $(\varepsilon_g, \delta_g)$-group-DP with group size k. Since $(\varepsilon_g, \delta_g, k)$ were arbitrary, no such universal conversion function F can exist. □

Proposition G.1 remains true if, instead of allowing arbitrary mechanisms, we restrict attention to mechanisms obtained by running DP-SGD and calibrating noise for a fixed group size k (including proxy choices such as $k = K_t(\delta_t)$ or $k \approx \lceil \sqrt{K_t^2} \rceil$ motivated by our HUC/sa-HUC analysis). We can always construct a Pufferfish scenario whose secrets are encoded in dataset differences at Hamming distance strictly greater than k and hence outside the scope of the group-DP guarantee. Therefore, even when the group-DP parameters $(\varepsilon_g, \delta_g, k)$ are known from a moments-accountant analysis of DP-SGD, there is no closed-form or universal transformation that converts them into the SRPP/sa-SRPP budget used in our HUC/sa-HUC–based SRPP-SGD calibration.

G.2 No Safe Bounds

We now specialize to the prevalence-based Pufferfish scenarios used in our SGD privatization experiments, where the secret changes only the label configuration on a fixed feature set (e.g., the `cat` label in CIFAR-10). Concretely, we consider two secrets s_0, s_1 corresponding to two datasets $X^{(0)}, X^{(1)} \in \bar{\mathcal{X}}^N$ with different prevalences $p_{\text{low}} \neq p_{\text{high}}$ of the protected class. Let

$$\tilde{\Delta} := d_H\big(X^{(0)}, X^{(1)}\big)$$

be the number of records whose labels differ between the two secrets. In our HUC/sa-HUC analysis, $\tilde{\Delta}$ influences the secret-induced update discrepancy through the (minibatch-level) discrepancy cap sequence, e.g., via K_t (or $\mathbb{E}_{\eta_t}[K_t^2]$), which determines the SRPP/sa-SRPP noise calibration. A natural question is whether one can instead treat such quantities as an *effective group size* and calibrate standard group DP-SGD with some k derived from K_t or $\mathbb{E}_{\eta_t}[K_t^2]$. The next proposition shows that, even in prevalence-based scenarios, a group-DP guarantee with any fixed k does not yield a general *conservative* (i.e., universally safe) bound on secret-level SRPP/sa-SRPP leakage.

Proposition G.2 (Group-DP does not upper-bound secret leakage in general)**.** *Fix any group size $k \in \mathbb{N}$ and any approximate group-DP parameters $(\varepsilon_g, \delta_g)$ with $\delta_g > 0$. There exist*

(i) a prevalence-based Pufferfish scenario $(\mathcal{S}, \mathcal{Q}, \Theta)$ with two secrets $s_0, s_1 \in \mathcal{S}$ whose realized datasets $X^{(0)}, X^{(1)}$ satisfy $d_H(X^{(0)}, X^{(1)}) = \tilde{\Delta} > k$, and

(ii) a Gaussian mechanism $\mathcal{M}$ that is $(\varepsilon_g, \delta_g)$-group-DP with group size k,

such that, for every $\alpha > 1$, the secret-level Rényi divergence between the output laws under $(S = s_0, \theta)$ and $(S = s_1, \theta)$ can be made arbitrarily large by increasing the dataset size. In particular, there is no finite function of $(\varepsilon_g, \delta_g, k)$ alone that universally upper-bounds this secret-level leakage.

Proof. Fix $k \in \mathbb{N}$ and $(\varepsilon_g, \delta_g)$ with $\delta_g > 0$. Choose a dataset size $N > k$ and an integer $\tilde{\Delta} \in \{k+1, \ldots, N\}$.

Prevalence-based Pufferfish scenario. Let the record domain be $\bar{\mathcal{X}} = \mathcal{A} \times \{0, 1\}$, where $\mathcal{A}$ denotes non-sensitive features and the second coordinate is the protected-class label. Let the dataset space be $\mathcal{X} := \bar{\mathcal{X}}^N$. Fix feature vectors $(a_1, \ldots, a_N) \in \mathcal{A}^N$. Define two label vectors by taking $b_i^{(0)} = 0$ for all i, and taking $b_i^{(1)} = 1$ for exactly $\tilde{\Delta}$ indices and $b_i^{(1)} = 0$ for the remaining indices. Let

$$X^{(0)} := ((a_i, b_i^{(0)}))_{i=1}^N, \qquad X^{(1)} := ((a_i, b_i^{(1)}))_{i=1}^N.$$

Then $d_H(X^{(0)}, X^{(1)}) = \tilde{\Delta} > k$, and the two datasets have different protected-class prevalences.

Define a Pufferfish scenario by

$$\mathcal{S} = \{s_0, s_1\}, \qquad \mathcal{Q} = \{(s_0, s_1), (s_1, s_0)\}, \qquad \Theta = \{\theta\},$$

with deterministic conditional data laws

$$P_\theta(X = X^{(0)} \mid S = s_0) = 1, \qquad P_\theta(X = X^{(1)} \mid S = s_1) = 1.$$

A group-DP mechanism with large secret-level leakage. Consider the statistic $T(X) = \sum_{i=1}^N b_i$, which counts the number of protected-class labels in the dataset. Under group adjacency

$$X \sim_k X' \iff d_H(X, X') \leq k,$$

the k-group sensitivity of T is

$$\Delta_k(T) := \sup_{X \sim_k X'} |T(X) - T(X')| = k.$$

Let

$$\mathcal{M}(X) := T(X) + Z, \qquad Z \sim \mathcal{N}(0, \sigma^2),$$

where σ is calibrated to sensitivity $\Delta_k(T) = k$ so that $\mathcal{M}$ satisfies $(\varepsilon_g, \delta_g)$-group-DP for group size k. For fixed $(\varepsilon_g, \delta_g)$, this calibration has the form $\sigma = c(\varepsilon_g, \delta_g)\, k$, where $c(\varepsilon_g, \delta_g) > 0$ does not depend on N or $\tilde{\Delta}$.

Under the Pufferfish scenario above,

$$\mathcal{M}(X) \mid (S = s_0, \theta) \sim \mathcal{N}(T(X^{(0)}), \sigma^2),$$

$$\mathcal{M}(X) \mid (S = s_1, \theta) \sim \mathcal{N}(T(X^{(1)}), \sigma^2).$$

Moreover,

$$|T(X^{(1)}) - T(X^{(0)})| = \tilde{\Delta}.$$

Therefore, for every $\alpha > 1$,

$$\begin{aligned} &\mathtt{D}_\alpha\Big(\mathcal{N}(T(X^{(0)}), \sigma^2) \,\Big\|\, \mathcal{N}(T(X^{(1)}), \sigma^2)\Big) \\ &= \frac{\alpha}{2}\frac{\tilde{\Delta}^2}{\sigma^2} = \frac{\alpha}{2}\frac{\tilde{\Delta}^2}{c(\varepsilon_g, \delta_g)^2 k^2}. \end{aligned}$$

Since k and $(\varepsilon_g, \delta_g)$ are fixed, we can increase N and choose $\tilde{\Delta}$ arbitrarily large, for example $\tilde{\Delta} = N$. The secret-level Rényi divergence then grows as $(\tilde{\Delta}/k)^2$ and can be made arbitrarily large.

Thus, a group-DP guarantee with fixed group size k does not provide a universal upper bound, as a function only of $(\varepsilon_g, \delta_g, k)$, on the secret-level leakage induced by an arbitrary prevalence-based Pufferfish scenario. □

Proposition G.2 shows that even for prevalence-based secrets as in our CIFAR-10 experiments, calibrating DP-SGD using a group size chosen from HUC/sa-HUC-derived quantities (e.g., $k = \lceil\sqrt{\overline{K}_t^2}\rceil$ or $k = K_{\text{cap}}$) does *not* yield a certified conservative bound in the SRPP/sa-SRPP sense. Group-DP is indexed by Hamming-adjacent dataset pairs, whereas SRPP/sa-SRPP quantify divergence between the secret-conditioned output laws in the specified Pufferfish scenario. Our HUC/sa-HUC-based SRPP-SGD calibration is therefore not a reparameterization of group DP-SGD, but a distinct, secret-aware analysis.

H Minimality and Attainability of HUC

In this section, we characterize the *minimality* and *attainability* of HUC defined by Definition 5.1 in Sec. 5.1.

Proposition H.1. *Fix t, a slicing set $\mathcal{U} = \{u_i\}_{i=1}^m \subset \mathbb{S}^{d-1}$, a history $y_{<t} \in \mathcal{Y}_{<t}$, and an admissible coupling γ. Define the* shift

$$\Delta_t(X, X'; R_t) := f_t(X, y_{<t}; R_t) - f_t(X', y_{<t}; R_t). \tag{44}$$

For each direction u_i, define

$$\phi_t(u_i) := \operatorname*{ess\,sup}_{(X,X',R_t)\sim\gamma\times\mathbb{P}_{\eta,\rho}} \big|\langle \Delta_t(X, X'; R_t), u_i\rangle\big| \text{ and } h^\star_{t,i} := \phi_t(u_i)^2.$$

In addition, let $\mathsf{P} := \gamma \times \mathbb{P}_{\eta,\rho}$ and $\mu_t := (\Delta_t)_\# \mathsf{P}$ be the push-forward measure of Δ_t, and define

$$\mathtt{ELS}_t := \operatorname{supp}(\mu_t) = \big\{z \in \mathbb{R}^d : \mu_t(\mathcal{B}_{\bar{\varepsilon}}(z)) > 0 \text{ for all } \bar{\varepsilon} > 0\big\},$$

where $\mathcal{B}_{\bar{\varepsilon}}(z)$ is a ball of radius $\bar{\varepsilon}$ centered at z.

Then, the following holds.

(i) **Minimality.** *Any HUC $h_t = (h_{t,1}, \ldots, h_{t,m})$ for $(y_{<t}, \gamma)$ satisfies $h_{t,i} \geq h^\star_{t,i}$ for all $i \in [m]$.*

(ii) **Attainability.** *If $\mathtt{ELS}_t$ is compact, then there exists $\Delta^\star_i \in \mathtt{ELS}_t$ such that $|\langle \Delta^\star_i, u_i\rangle| = \sup_{\Delta\in\mathtt{ELS}_t} |\langle \Delta, u_i\rangle|$. Moreover, by definition of $\phi_t(u_i)$ as an essential supremum, for very $\varepsilon > 0$, $\mathsf{P}(|\langle \Delta_t(X, X'; R_t), u_i\rangle| \geq \phi_t(u_i) - \varepsilon) > 0$.*

However, the HUC constructed in Proposition 5.1 is generally not tight; it is a computable conservative bound. Proposition H.1 characterizes the *minimal* HUC and specifies when it is attained. In particular, any valid HUC h_t must satisfy $h_{t,i} \geq h^\star_{t,i}$ for all $i \in [m]$, and the minimal cap $h^\star_t$ is attained whenever the essential shift set $\mathtt{ELS}_t$ is compact.

In our setting, compactness of $\mathtt{ELS}_t$ follows from three ingredients: (i) per-example clipping, which uniformly bounds each individual gradient contribution; (ii) a finite discrepancy cap together with a deterministic (or bounded) batch size; and (iii) an L_t-Lipschitz post-update map. These yield a uniform almost-sure bound $\|\Delta_t\|_2 \leq C_t$ for some finite C_t, and hence $\mathtt{ELS}_t \subset \overline{B(0, C_t)}$, which is closed and bounded in $\mathbb{R}^d$ and therefore compact.

The following corollary follows Proposition H.1.

Corollary H.1. *Consider the linear, blockwise-clipped case where*

$$f_t(x, y_{<t}; R_t) = \xi_{t-1} - A_t\, \bar{g}_t(x; R_t),$$

with A_t independent of $((X, X'), R_t)$, and per-block clipping thresholds $\{C_b\}_{b=1}^{B_{\text{blk}}}$ over an orthogonal block decomposition with projectors $\{\mathsf{P}_b\}_{b=1}^{B_{\text{blk}}}$. Assume that, for each block b, the blockwise gradient differences can attain the clipping radius in directions aligned with the vectors $\mathsf{P}_b A_t^\top u_i$ (for the slices of interest). Then the minimal per-slice HUC is

$$h^\star_{t,i} = \Big(\frac{2K_t}{B_t}\Big)^2 \Big(\sum_{b=1}^{B_{\text{blk}}} C_b \big\|\mathsf{P}_b A_t^\top u_i\big\|_2\Big)^2, \qquad i \in [m]. \tag{45}$$

H.1 Proof of Proposition H.1

Let $\mathsf{P} := \gamma \times \mathbb{P}_{\eta,\rho}$ and write

$$Z_i(X, X', R_t) := \big|\langle \Delta_t(X, X'; R_t), u_i\rangle\big|, \qquad i \in [m].$$

Recall that $\phi_t(u_i) = \operatorname{ess\,sup}_{\mathsf{P}} Z_i$ and $h^\star_{t,i} = \phi_t(u_i)^2$.

We use the following standard property: if $Z \geq 0$ is measurable and $Z \leq c$ holds P-a.s., then $\operatorname{ess\,sup}_{\mathsf{P}} Z \leq c$. Moreover, by definition of essential supremum, for every $\varepsilon > 0$,

$$\mathsf{P}\big(Z \geq \operatorname*{ess\,sup}_{\mathsf{P}} Z - \varepsilon\big) > 0. \tag{46}$$

(i) Minimality. Let $h_t = (h_{t,1}, \dots, h_{t,m})$ be any HUC for $(y_{<t}, \gamma)$. By Definition 5.1, for each $i \in [m]$ we have

$$Z_i(X, X', R_t) = \big|\langle \Delta_t(X, X'; R_t), u_i \rangle\big| \le \sqrt{h_{t,i}} \quad \mathsf{P}\text{-a.s.}$$

Applying the essential-supremum property gives

$$\phi_t(u_i) = \operatorname*{ess\,sup}_{\mathsf{P}} Z_i \le \sqrt{h_{t,i}}.$$

Squaring yields $h^\star_{t,i} = \phi_t(u_i)^2 \le h_{t,i}$ for all $i \in [m]$, proving minimality.

(ii) Attainability. Define $\mu_t := (\Delta_t)_\# \mathsf{P}$ and $\mathtt{ELS}_t := \operatorname{supp}(\mu_t)$. Assume $\mathtt{ELS}_t$ is compact. The map $\Delta \mapsto |\langle \Delta, u_i \rangle|$ is continuous, hence by the Weierstrass Extreme Value Theorem there exists $\Delta_i^\star \in \mathtt{ELS}_t$ such that

$$|\langle \Delta_i^\star, u_i \rangle| = \sup_{\Delta \in \mathtt{ELS}_t} |\langle \Delta, u_i \rangle|.$$

This establishes attainment of the *set supremum* over $\mathtt{ELS}_t$.

Finally, the *essential* attainability statement follows directly from (46) applied to Z_i: for every $\varepsilon > 0$,

$$\mathsf{P}\big(Z_i \ge \phi_t(u_i) - \varepsilon\big) > 0.$$

Equivalently, in pushforward form,

$$\mu_t\Big(\big\{\Delta : \ |\langle \Delta, u_i \rangle| \ge \phi_t(u_i) - \varepsilon\big\}\Big) > 0.$$

□

H.2 Proof of Corollary H.1

Fix iteration t and a slice $u_i \in \mathcal{U}$. Let $\mathsf{P} := \gamma \times \mathbb{P}_{\eta,\rho}$. Write

$$f_t(x, y_{<t}; R_t) = \xi_{t-1} - A_t\, \bar{g}_t(x; R_t),$$
$$\Delta_t(X, X'; R_t) = -A_t\big(\bar{g}_t(X; R_t) - \bar{g}_t(X'; R_t)\big).$$

Define

$$v := \bar{g}_t(X; R_t) - \bar{g}_t(X'; R_t), \quad w := A_t^\top u_i.$$

Then $|\langle \Delta_t, u_i \rangle| = |\langle v, w \rangle|$.

Step 1: a.s. feasible set for v. Under blockwise clipping with thresholds $\{C_b\}_{b=1}^{B_{\mathrm{blk}}}$ and a discrepancy cap K_t, at most K_t per-example contributions in the minibatch can differ between X and X'. Since each per-example block-gradient is clipped to ℓ_2-norm at most C_b in block b, the blockwise difference in the averaged clipped gradients satisfies, P-a.s.,

$$\|\mathsf{P}_b v\|_2 \ \le \ \frac{2C_b}{B_t} K_t \ =: \ \alpha_b, \quad b = 1, \dots, B_{\mathrm{blk}}.$$

Equivalently, $v \in \mathcal{V}$ P-a.s., where

$$\mathcal{V} := \Big\{ v \in \mathbb{R}^d : \ \|\mathsf{P}_b v\|_2 \le \alpha_b, \ \ b = 1, \dots, B_{\mathrm{blk}} \Big\}.$$

Step 2: upper bound on the minimal cap. By Proposition H.1 (minimality with $\phi_t(u_i) = \operatorname{ess\,sup}_{\mathsf{P}} |\langle \Delta_t, u_i \rangle|$),

$$\phi_t(u_i) = \operatorname*{ess\,sup}_{\mathsf{P}} |\langle v, w \rangle| \ \le \ \sup_{v \in \mathcal{V}} |\langle v, w \rangle|.$$

Now decompose $v = \sum_b v_b$ and $w = \sum_b w_b$ with $v_b := \mathsf{P}_b v$, $w_b := \mathsf{P}_b w$. Then, for any $v \in \mathcal{V}$,

$$|\langle v, w \rangle| = \Big| \sum_{b=1}^{B_{\mathrm{blk}}} \langle v_b, w_b \rangle \Big| \le \sum_{b=1}^{B_{\mathrm{blk}}} \|v_b\|_2 \, \|w_b\|_2 \le \sum_{b=1}^{B_{\mathrm{blk}}} \alpha_b \|w_b\|_2.$$

Hence

$$\sup_{v \in \mathcal{V}} |\langle v, w \rangle| = \sum_{b=1}^{B_{\mathrm{blk}}} \alpha_b \|w_b\|_2 = \frac{2K_t}{B_t} \sum_{b=1}^{B_{\mathrm{blk}}} C_b \, \|\mathsf{P}_b A_t^\top u_i\|_2.$$

Therefore,

$$\phi_t(u_i) \ \le \ \frac{2K_t}{B_t} \sum_{b=1}^{B_{\mathrm{blk}}} C_b \, \|\mathsf{P}_b A_t^\top u_i\|_2. \tag{47}$$

Step 3: matching lower bound (attainability in essential-sup sense). Consider the optimizer $v^\star \in \mathcal{V}$ defined blockwise by

$$v_b^\star = \begin{cases} \alpha_b \dfrac{w_b}{\|w_b\|_2}, & w_b \ne 0, \\ 0, & w_b = 0, \end{cases} \qquad b = 1, \dots, B_{\mathrm{blk}}.$$

Then $v^\star \in \mathcal{V}$ and

$$\langle v^\star, w \rangle = \sum_{b=1}^{B_{\mathrm{blk}}} \alpha_b \|w_b\|_2 = \sup_{v \in \mathcal{V}} \langle v, w \rangle,$$

so $\sup_{v \in \mathcal{V}} |\langle v, w \rangle|$ equals the right-hand side of (47).

By the corollary's assumption ("the blockwise gradient differences can attain the clipping radius in directions aligned with $\mathsf{P}_b A_t^\top u_i$"), together with the requirement that the corresponding discrepancy pattern of size K_t occurs with nonzero probability under the sampling rule, we have that for every $\varepsilon > 0$,

$$\mathsf{P}\Big(|\langle v, w \rangle| \ge |\langle v^\star, w \rangle| - \varepsilon \Big) > 0.$$

By the defining property of essential supremum, this implies

$$\phi_t(u_i) = \operatorname*{ess\,sup}_{\mathsf{P}} |\langle v, w \rangle| = |\langle v^\star, w \rangle| = \frac{2K_t}{B_t} \sum_{b=1}^{B_{\mathrm{blk}}} C_b \, \|\mathsf{P}_b A_t^\top u_i\|_2.$$

Squaring yields

$$h^\star_{t,i} = \phi_t(u_i)^2 = \Big(\frac{2K_t}{B_t}\Big)^2 \Big(\sum_{b=1}^{B_{\mathrm{blk}}} C_b \, \|\mathsf{P}_b A_t^\top u_i\|_2 \Big)^2,$$

which is exactly (45). □

I Minibatch Subsampling models: WR vs. WOR

Our HUC/sa-HUC analysis is compatible with standard minibatch subsampling variants. The proofs follow the same skeleton. Only the sampling model for the round-t randomness R_t and the interpretation of the index collection I_t change (set vs. sequence with possible repeats). In this section, we characterize the comparison between sampling With Replacement (WR) and Without Replacement (WOR).

Convention for WOR. Unless stated otherwise, we adopt the following standard model.

Assumption I.1 (Minibatch Subsampling WOR)**.** *At each iteration t, the minibatch indices $\mathsf{I}_t \subseteq [n]$ are sampled uniformly without replacement with $|\mathsf{I}_t| = B_t$. All algorithmic randomness at round t (including subsampling and any additional randomness) is included in $R_t \sim \mathbb{P}_{\eta,\rho}$.*

Under Assumption I.1, $\mathsf{I}_t(r)$ is a *set*. The discrepancy in (14) is unambiguous:

$$K_t(x,x';r) \;:=\; \sum_{j\in\mathsf{I}_t(r)} \mathbf{1}\{x_j \neq x'_j\},$$

where $\mathsf{I}_t(r)$ is a set. In particular, if x and x' differ in at most $\overline{K}$ coordinates, then $K_t(x,x';r) \le \min\{\overline{K}, B_t\}$ for every draw r.

Assumption I.2 (Minibatch Subsampling WR)**.** *At each iteration t, the minibatch indices are formed by B_t i.i.d. draws from $[n]$. Equivalently, $\mathsf{I}_t(r) = (I_{t,1}(r), \dots, I_{t,B_t}(r))$ is a length-B sequence in $[n]$ (allowing repetitions). All round-t randomness is included in $R_t \sim \mathbb{P}_{\eta,\rho}$.*

We keep the same discrepancy definition as in (14), interpreted *per draw*:

$$K_t(x,x';r) \;:=\; \sum_{b=1}^{B_t} \mathbf{1}\{x_{I_{t,b}(r)} \neq x'_{I_{t,b}(r)}\},$$

where $\mathsf{I}_t(r)$ is a sequence. Thus, repeated indices contribute repeatedly; this matches the potential influence entering through the sampled minibatch.

Proposition I.1 (Worst-case discrepancy caps under WR vs. WOR)**.** *Fix any $x,x' \in \bar{\mathcal{X}}^n$.*

WR. Under Assumption I.2, for any draw r one has $0 \le K_t(x,x';r) \le B_t$, where

$$K_t(x,x';r) := \sum_{b=1}^{B_t} \mathbf{1}\{x_{I_{t,b}(r)} \neq x'_{I_{t,b}(r)}\}.$$

If $d_H(x,x') = k$, then $K_t(x,x';R_t) \sim \mathrm{Binomial}(B_t, k/n)$ and $\mathbb{E}[K_t] = B_t(k/n)$.

Moreover, for any coupling γ,

$$K_t(\gamma) = \operatorname*{ess\,sup}_{((X,X'),R_t)\sim\gamma\times\mathbb{P}_{\eta,\rho}} K_t(X,X';R_t) = \begin{cases} 0, & \text{if } X = X' \ \gamma\text{-a.s.}, \\ B_t, & \text{otherwise.} \end{cases}$$

WOR. Under Assumption I.1, $\mathsf{I}_t(r)$ is a set and

$$K_t(x,x';r) := \sum_{j\in\mathsf{I}_t(r)} \mathbf{1}\{x_j \neq x'_j\}.$$

If $d_H(x,x') \le \overline{K}$, then deterministically $K_t(x,x';r) \le \min\{\overline{K}, B_t\}$ for every draw r, and hence $K_t(\gamma) \le \min\{\overline{K}, B_t\}$ whenever $d_H(X,X') \le \overline{K}$ γ-a.s.

Proof. (WR upper bound.) Under Assumption I.2, $\mathsf{I}_t(r) = (I_{t,1}(r), \dots, I_{t,B_t}(r))$ is a length-B_t sequence. Hence

$$K_t(x,x';r) := \sum_{b=1}^{B_t} \mathbf{1}\{x_{I_{t,b}(r)} \neq x'_{I_{t,b}(r)}\}$$

is a sum of B_t indicators, so $0 \le K_t(x,x';r) \le B_t$.

(WR binomial law.) If x and x' differ in exactly k coordinates, each draw hits a differing index with probability k/n, independently across draws, so $K_t(x,x';R_t) \sim \mathrm{Binomial}(B_t, k/n)$.

(WR worst-case cap.) Fix a coupling γ and consider

$$K_t(\gamma) = \operatorname*{ess\,sup}_{((X,X'),R_t)\sim\gamma\times\mathbb{P}_{\eta,\rho}} K_t(X,X';R_t).$$

If $X = X'$ γ-a.s., then $K_t(X,X';R_t) = 0$ a.s., so $K_t(\gamma) = 0$. Otherwise, $\Pr_\gamma(X \neq X') > 0$. On the event $\{X \neq X'\}$, let $D := \{j \in [n] : X_j \neq X'_j\}$, so $D \neq \emptyset$. Under WR, for any fixed $j \in D$, the event $\{I_{t,1} = \cdots = I_{t,B_t} = j\}$ has probability $(1/n)^{B_t} > 0$, and on this event $K_t(X,X';R_t) = B_t$. Therefore $\Pr\big(K_t(X,X';R_t) = B_t\big) > 0$ on $\{X \neq X'\}$, which implies $K_t(\gamma) = B_t$.

(WOR comparison.) Under Assumption I.1, $\mathsf{I}_t(r)$ is a set and $K_t(x,x';r) = \sum_{j\in\mathsf{I}_t(r)} \mathbf{1}\{x_j \neq x'_j\}$. If x and x' differ in at most $\overline{K}$ coordinates then deterministically $K_t(x,x';r) \le \min\{\overline{K}, B_t\}$ for every draw r, and the same holds γ-a.s. when $d_H(X,X') \le \overline{K}$ γ-a.s. □

J Utility of Subsampling-Aware SRPP Accounting

This appendix formalizes why incorporating minibatch randomness into the accounting can improve utility. We compare two *sufficient calibrations* for the *same* SRPP guarantee for the *same* SGD mechanism: a worst-case (HUC) envelope bound versus a subsampling-aware (sa-HUC) envelope bound.

J.1 A Rényi bound for shared-mixtures

Let $\alpha > 1$. Let $(\Omega, \mathcal{F}, \nu)$ be a probability space and $\eta \sim \nu$. For each $\eta \in \Omega$, let P_η and Q_η be distributions on a common measurable space $(\mathcal{Y}, \mathcal{A})$, with $P_\eta \ll Q_\eta$ for ν-a.e. η. Define the shared-mixtures $P := \int P_\eta \, d\nu(\eta)$ and $Q := \int Q_\eta \, d\nu(\eta)$.

Lemma J.1 (Rényi divergence of shared-mixtures)**.** *Assume $P_\eta \ll Q_\eta$ for ν-a.e. η. Assume further that $\eta \mapsto D_\alpha(P_\eta \| Q_\eta)$ is measurable and that $\mathbb{E}_{\eta\sim\nu}\big[\exp\big((\alpha-1)D_\alpha(P_\eta\|Q_\eta)\big)\big] < \infty$ (otherwise the right-hand side is $+\infty$ and the inequality below is trivial). Then*

$$\mathtt{D}_\alpha(P\|Q) \;\le\; \frac{1}{\alpha-1}\log \mathbb{E}_{\eta\sim\nu}\big[\exp\big((\alpha-1)\mathtt{D}_\alpha(P_\eta\|Q_\eta)\big)\big].$$

Moreover, the right-hand side is strictly smaller than $\operatorname{ess\,sup}_\eta \mathtt{D}_\alpha(P_\eta\|Q_\eta)$ *whenever* $\mathtt{D}_\alpha(P_\eta\|Q_\eta)$ *is not* ν*-a.s. constant.*

Proof. Fix a σ-finite dominating measure μ on $(\mathcal{Y}, \mathcal{A})$ such that $P_\eta, Q_\eta \ll \mu$ for ν-a.e. η (e.g. take $\mu = Q + \sum_{j\ge1} 2^{-j} P_{\eta_j}$ for a dense countable subfamily, or simply assume existence as standard in this setting). Write $p_\eta = \frac{dP_\eta}{d\mu}$ and $q_\eta = \frac{dQ_\eta}{d\mu}$. Then $p = \mathbb{E}_\eta[p_\eta]$ and $q = \mathbb{E}_\eta[q_\eta]$ are Radon–Nikodym derivatives of $P = \int P_\eta d\nu(\eta)$ and $Q = \int Q_\eta d\nu(\eta)$ with respect to μ.

By definition of Rényi divergence,

$$\exp\big((\alpha-1)\mathtt{D}_\alpha(P\|Q)\big) = \int \left(\frac{p}{q}\right)^\alpha q\, d\mu = \int p^\alpha q^{1-\alpha}\, d\mu.$$

For $\alpha > 1$, the map

$$f : (0,\infty)\times(0,\infty) \to (0,\infty), \quad f(a,b) := a^\alpha b^{1-\alpha} = b\Big(\frac{a}{b}\Big)^\alpha$$

is jointly convex (it is the perspective of the convex function $x \mapsto x^\alpha$). Therefore, for each $y \in \mathcal{Y}$, applying Jensen's inequality to the random pair $(p_\eta(y), q_\eta(y))$ gives

$$\begin{aligned}&(\mathbb{E}_\eta p_\eta(y))^\alpha (\mathbb{E}_\eta q_\eta(y))^{1-\alpha} = f\big(\mathbb{E}_\eta p_\eta(y), \mathbb{E}_\eta q_\eta(y)\big)\\ &\le \mathbb{E}_\eta f\big(p_\eta(y), q_\eta(y)\big) = \mathbb{E}_\eta\big[p_\eta(y)^\alpha q_\eta(y)^{1-\alpha}\big].\end{aligned}$$

Integrating over μ and using Tonelli/Fubini (the integrand is nonnegative) yields

$$\begin{aligned}&\exp\big((\alpha-1)D_\alpha(P\|Q)\big) = \int p^\alpha q^{1-\alpha}\, d\mu\\ &\le \mathbb{E}_\eta \int p_\eta^\alpha q_\eta^{1-\alpha}\, d\mu = \mathbb{E}_\eta \exp\big((\alpha-1)D_\alpha(P_\eta\|Q_\eta)\big).\end{aligned}$$

Taking $\frac{1}{\alpha-1}\log$ proves the stated bound.

For the strictness claim, note that for any random variable Z,

$$\frac{1}{\alpha-1}\log \mathbb{E}[e^{(\alpha-1)Z}] \le \operatorname{ess\,sup} Z,$$

with strict inequality whenever Z is not ν-a.s. constant (since then $\mathbb{P}(Z \le \operatorname{ess\,sup} Z - \varepsilon) > 0$ for some $\varepsilon > 0$ and thus the exponential moment is strictly smaller than $e^{(\alpha-1)\operatorname{ess\,sup} Z}$). Apply this to $Z = \mathtt{D}_\alpha(P_\eta\|Q_\eta)$. □

J.2 Subsampling-aware envelopes at one iteration

Fix an iteration t, a Pufferfish scenario $(\mathcal{S}, \mathcal{Q}, \Theta)$, and a slice profile $\{\mathcal{U}, \omega\}$. Let η_t denote the (random) minibatch selector at round t. Condition on a fixed history up to round t (as in the main accountant). Fix $(\theta, (s_i, s_j)) \in \Theta \times \mathcal{Q}$ and a coupling $\Gamma_{\theta,ij}$ for $(X^{(i)}, X^{(j)})$.

Let R_t denote all remaining internal randomness used by the round-t update, and view (η_t, R_t) as jointly generated by the algorithm at round t. (Equivalently, one may treat minibatch selection as part of the round randomness; the conditioning below is the same.)

For each slice $u \in \mathcal{U}$ and each minibatch outcome η in the support of η_t, let $\mathbb{P}(\cdot \mid \eta_t = \eta)$ denote a regular conditional distribution for $(X^{(i)}, X^{(j)}, R_t)$ given $\eta_t = \eta$ (under the coupled experiment). Define the conditional shift bound

$$z_t^u(\eta) \;:=\; \operatorname*{ess\,sup}_{\mathbb{P}(\cdot\mid\eta_t=\eta)} \Big|\big\langle f_t(X^{(i)}, y_{<t}; R_t) - f_t(X^{(j)}, y_{<t}; R_t),\, u\big\rangle\Big|. \tag{48}$$

Let $z_t^u(\eta_t)$ be the resulting random variable when η_t is drawn by the algorithm. Define the worst-case (HUC-style) per-slice cap

$$\bar{z}_t^u \;:=\; \operatorname*{ess\,sup}_{\eta_t} z_t^u(\eta_t).$$

Worst-case envelope bounds (HUC). Define the per-round worst-case averaged and joint envelopes by

$$\mathtt{AR}_t^{\mathrm{wc}}(\sigma) := \int R_\alpha\big(\zeta_\sigma, \bar{z}_t^u\big)\, d\omega(u), \tag{49}$$

$$\mathtt{JR}_t^{\mathrm{wc}}(\sigma) := \frac{1}{\alpha-1}\log\left(\int \exp\big((\alpha-1)R_\alpha(\zeta_\sigma, \bar{z}_t^u)\big)\, d\omega(u)\right), \tag{50}$$

where ζ_σ denotes the Gaussian noise distribution with variance σ^2.

Subsampling-aware envelope bounds (sa-HUC). Define the per-round subsampling-aware envelopes by

$$\begin{aligned}\mathtt{AR}_t^{\mathrm{sa}}(\sigma) := \int \frac{1}{\alpha-1}\log \mathbb{E}_{\eta_t}\big[&\\ \exp\big((\alpha-1)R_\alpha(\zeta_\sigma, z_t^u(\eta_t))\big)\big]\, d\omega(u),&\end{aligned} \tag{51}$$

$$\begin{aligned}\mathtt{JR}_t^{\mathrm{sa}}(\sigma) := \frac{1}{\alpha-1}\log \mathbb{E}_{\eta_t}\Big[&\\ \int \exp\big((\alpha-1)R_\alpha(\zeta_\sigma, z_t^u(\eta_t))\big)\, d\omega(u)\Big].&\end{aligned} \tag{52}$$

Lemma J.2 (Subsampling-aware envelopes are never looser)**.** *Fix $\alpha > 1$ and $\sigma > 0$. Assume that for ω-a.e. $u \in \mathcal{U}$, the map $\eta \mapsto z_t^u(\eta)$ is measurable and*

$$\mathbb{E}_{\eta_t}\big[\exp\big((\alpha-1)R_\alpha(\zeta_\sigma, z_t^u(\eta_t))\big)\big] < \infty,$$

(otherwise the corresponding $\mathtt{AR}_t^{\mathrm{sa}}(\sigma)$ term equals $+\infty$ and the inequality is trivial). Then

$$\mathtt{AR}_t^{\mathrm{sa}}(\sigma) \;\le\; \mathtt{AR}_t^{\mathrm{wc}}(\sigma), \qquad \mathtt{JR}_t^{\mathrm{sa}}(\sigma) \;\le\; \mathtt{JR}_t^{\mathrm{wc}}(\sigma).$$

Moreover, the inequality is strict for AR *if there exists a measurable set $U_0 \subseteq \mathcal{U}$ with $\omega(U_0) > 0$ such that for every $u \in U_0$,*

$$\Pr\Big(R_\alpha(\zeta_\sigma, z_t^u(\eta_t)) < R_\alpha(\zeta_\sigma, \bar{z}_t^u)\Big) > 0.$$

Similarly, the inequality is strict for JR *if*

$$\begin{aligned}\Pr\Big(\int \exp\big((\alpha-1)R_\alpha(\zeta_\sigma, z_t^u(\eta_t))\big)\, d\omega(u) \\ < \int \exp\big((\alpha-1)R_\alpha(\zeta_\sigma, \bar{z}_t^u)\big)\, d\omega(u)\Big) \\ > 0.\end{aligned}$$

Proof. Fix $u \in \mathcal{U}$. By definition, $z_t^u(\eta_t) \le \bar{z}_t^u$ almost surely. Since $R_\alpha(\zeta_\sigma, \cdot)$ is nondecreasing, we have

$$R_\alpha(\zeta_\sigma, z_t^u(\eta_t)) \le R_\alpha(\zeta_\sigma, \bar{z}_t^u) \quad \text{a.s.}$$

Therefore,

$$\mathbb{E}_{\eta_t} \exp\big((\alpha-1)R_\alpha(\zeta_\sigma, z_t^u(\eta_t))\big) \le \exp\big((\alpha-1)R_\alpha(\zeta_\sigma, \bar{z}_t^u)\big).$$

Taking $\frac{1}{\alpha-1}\log$ yields

$$\frac{1}{\alpha-1}\log \mathbb{E}_{\eta_t}\big[\exp\big((\alpha-1)R_\alpha(\zeta_\sigma, z_t^u(\eta_t))\big)\big] \le R_\alpha(\zeta_\sigma, \bar{z}_t^u).$$

Integrating both sides over $u \sim \omega$ gives $\mathtt{AR}_t^{\mathrm{sa}}(\sigma) \le \mathtt{AR}_t^{\mathrm{wc}}(\sigma)$.

For strictness of AR, fix any u such that $\Pr(R_\alpha(\zeta_\sigma, z_t^u(\eta_t)) < R_\alpha(\zeta_\sigma, \bar{z}_t^u)) > 0$. Then the random variable $D_u(\eta_t) := R_\alpha(\zeta_\sigma, z_t^u(\eta_t))$ is not ν-a.s. constant and satisfies $D_u(\eta_t) \le \operatorname{ess\,sup}_{\eta_t} D_u(\eta_t) = R_\alpha(\zeta_\sigma, \bar{z}_t^u)$. Hence,

$$\frac{1}{\alpha-1}\log \mathbb{E}_{\eta_t}\Big[e^{(\alpha-1)D_u(\eta_t)}\Big] < R_\alpha(\zeta_\sigma, \bar{z}_t^u).$$

Integrating this strict inequality over a set of u of positive ω-mass yields $\mathtt{AR}_t^{\mathrm{sa}}(\sigma) < \mathtt{AR}_t^{\mathrm{wc}}(\sigma)$.

For JR, the same pointwise monotonicity implies that almost surely in η_t,

$$\begin{aligned}&\int \exp\big((\alpha-1)R_\alpha(\zeta_\sigma, z_t^u(\eta_t))\big)\, d\omega(u) \\ &\le \int \exp\big((\alpha-1)R_\alpha(\zeta_\sigma, \bar{z}_t^u)\big)\, d\omega(u).\end{aligned}$$

Taking $\mathbb{E}_{\eta_t}$ and then $\frac{1}{\alpha-1}\log$ yields $\mathtt{JR}_t^{\mathrm{sa}}(\sigma) \le \mathtt{JR}_t^{\mathrm{wc}}(\sigma)$. Strictness follows immediately under the stated strict-inequality event (since log is strictly increasing). □

J.3 Full Version of Proposition 5.3

Let $\mathcal{M}_{\mathrm{SGD}}$ be the full T-round mechanism in Algorithm 1. Our accountant theorems (e.g., Theorem 5.6 for HUC and Theorem 5.7 for sa-HUC) aggregate valid per-round envelope bounds to obtain a total SRPP guarantee.

Proposition J.1 (Full Version of Proposition 5.3)**.** *Let $\mathcal{M}_{\mathrm{SGD}}$ be the T-round mechanism in Algorithm 1. Assume the hypotheses of Theorem 5.6 (HUC accountant) and Theorem 5.7 (sa-HUC accountant) hold. Assume additionally that for each fixed z, the envelope $R_\alpha(\zeta_\sigma, z)$ is nonincreasing in σ^2 (e.g., for Gaussian noise ζ_σ).*

For each round t and each $\sigma > 0$, suppose the per-round envelope contribution in the averaged case admits either a worst-case bound $\mathtt{AR}_t^{\mathrm{wc}}(\sigma)$ or a subsampling-aware bound $\mathtt{AR}_t^{\mathrm{sa}}(\sigma)$, and similarly in the joint case admits either $\mathtt{JR}_t^{\mathrm{wc}}(\sigma)$ or $\mathtt{JR}_t^{\mathrm{sa}}(\sigma)$, as defined in Section J.2.

Let $\mathsf{Agg} : \mathbb{R}_+^T \to \mathbb{R}_+$ denote the aggregation rule used by the corresponding accountant theorem, and assume Agg is coordinatewise nondecreasing, i.e., if $a_t \le b_t$ for all t then $\mathsf{Agg}(a_{1:T}) \le \mathsf{Agg}(b_{1:T})$.

Then for every $\sigma > 0$,

$$\mathsf{Agg}\big(\mathtt{AR}_{1:T}^{\mathrm{sa}}(\sigma)\big) \;\le\; \mathsf{Agg}\big(\mathtt{AR}_{1:T}^{\mathrm{wc}}(\sigma)\big),$$

and

$$\mathsf{Agg}\big(\mathtt{JR}_{1:T}^{\mathrm{sa}}(\sigma)\big) \;\le\; \mathsf{Agg}\big(\mathtt{JR}_{1:T}^{\mathrm{wc}}(\sigma)\big),$$

where $\mathtt{AR}_{1:T}^{\bullet}(\sigma) := (\mathtt{AR}_t^{\bullet}(\sigma))_{t=1}^T$ and $\mathtt{JR}_{1:T}^{\bullet}(\sigma) := (\mathtt{JR}_t^{\bullet}(\sigma))_{t=1}^T$, for $\bullet \in \{\mathrm{wc}, \mathrm{sa}\}$. Moreover, if Agg is strictly increasing in at least one coordinate (e.g. $\mathsf{Agg}(a_{1:T}) = \sum_{t=1}^T a_t$) and the strictness condition of Lemma J.2 holds for at least one round t (for the corresponding Ave/JR case), then the above inequality is strict.

Consequently, for any target SRPP (or sa-SRPP) budget level ε, the minimal noise variance

$$\sigma_{\mathrm{sa}}^2(\varepsilon) := \inf\{\sigma^2 :\; \mathsf{Agg}(\mathtt{AR}_{1:T}^{\mathrm{sa}}(\sigma)) \le \varepsilon\} \quad (\textit{or with } \mathtt{JR})$$

satisfies $\sigma_{\mathrm{sa}}^2(\varepsilon) \le \sigma_{\mathrm{wc}}^2(\varepsilon)$, and typically $\sigma_{\mathrm{sa}}^2(\varepsilon) < \sigma_{\mathrm{wc}}^2(\varepsilon)$.

Proof. Lemma J.2 gives $\mathtt{AR}_t^{\mathrm{sa}}(\sigma) \le \mathtt{AR}_t^{\mathrm{wc}}(\sigma)$ and $\mathtt{JR}_t^{\mathrm{sa}}(\sigma) \le \mathtt{JR}_t^{\mathrm{wc}}(\sigma)$ for each t and each σ. Applying the coordinatewise nondecreasing aggregation rule Agg yields the claimed inequalities for the full T-round mechanism.

If Agg is strictly increasing in some coordinate and at least one round admits a strict inequality (as characterized in Lemma J.2), then the aggregated inequality is strict.

Finally, since the aggregated envelope bound is nonincreasing in σ^2, the set of σ^2 values sufficient to certify a fixed target budget under the subsampling-aware bound contains the corresponding set under the worst-case bound. Thus the minimal sufficient σ^2 under the subsampling-aware bound cannot exceed that under the worst-case bound. □

K Proof of Lemma 4.1

Fix $\theta \in \Theta$, $(s_i, s_j) \in \mathcal{Q}$ with $P_\theta^S(s_i), P_\theta^S(s_j) > 0$, a direction $u \in \mathcal{U}$, and $\alpha > 1$. Let $P^{f,s} := P_\theta(f(X) \mid s)$ and define the one-dimensional projected pre-noise laws

$$P_u := \Psi^u_{\#} P_\theta^{f,s_i}, \quad Q_u := \Psi^u_{\#} P_\theta^{f,s_j}.$$

Let $z^u_\infty := W_\infty(P_u, Q_u)$.

By definition of W_∞, for every $\varepsilon > 0$ there exists a coupling $\pi_\varepsilon \in \Pi(P_u, Q_u)$ such that

$$|Z - Z'| \le z^u_\infty + \varepsilon \quad \text{for } \pi_\varepsilon\text{-a.e. } (Z, Z'). \tag{53}$$

Step 1 (Projected additive-noise form). Let $\widetilde{N} := \Psi^u(N)$. Then $\widetilde{N}$ is independent of $f(X)$ and has distribution $\widetilde{\zeta} := \Psi^u_{\#}\zeta$ on $\mathbb{R}$. For $z \in \mathbb{R}$, write $\widetilde{\zeta}_z$ for the law of $\widetilde{N} + z$. Then the projected endpoint laws are mixtures of shifted noises:

$$\Psi^u_{\#}\mathcal{M}^\theta_{s_i} = \int \widetilde{\zeta}_z \, P_u(dz), \quad \Psi^u_{\#}\mathcal{M}^\theta_{s_j} = \int \widetilde{\zeta}_{z'} \, Q_u(dz').$$

Using the coupling π_ε of (P_u, Q_u), we can write both mixtures over the common index $(z, z') \sim \pi_\varepsilon$:

$$\Psi^u_{\#}\mathcal{M}^\theta_{s_i} = \int \widetilde{\zeta}_z \, \pi_\varepsilon(dz, dz'), \quad \Psi^u_{\#}\mathcal{M}^\theta_{s_j} = \int \widetilde{\zeta}_{z'} \, \pi_\varepsilon(dz, dz').$$

Step 2 (Pointwise bound for each shift pair). For any $z, z' \in \mathbb{R}$, Rényi divergence is invariant under common translations, hence

$$\mathtt{D}_\alpha(\widetilde{\zeta}_z \| \widetilde{\zeta}_{z'}) = \mathtt{D}_\alpha(\widetilde{\zeta}_{z-z'} \| \widetilde{\zeta}).$$

By definition of the (1-D) shift-Rényi envelope,

$$R_\alpha(\widetilde{\zeta}, r) := \sup_{|a| \le r} \mathtt{D}_\alpha(\widetilde{\zeta}_a \| \widetilde{\zeta}),$$

we obtain

$$\mathtt{D}_\alpha(\widetilde{\zeta}_z \| \widetilde{\zeta}_{z'}) \le R_\alpha(\widetilde{\zeta}, |z - z'|).$$

Combining with (53) yields, for π_ε-a.e. (z, z'),

$$\mathtt{D}_\alpha(\widetilde{\zeta}_z \| \widetilde{\zeta}_{z'}) \le R_\alpha(\widetilde{\zeta}, z^u_\infty + \varepsilon).$$

Step 3 (From pointwise to mixtures via DPI). Define probability measures on $\mathbb{R}^2 \times \mathbb{R}$ by

$$\overline{P}_\varepsilon(dz, dz', dy) := \pi_\varepsilon(dz, dz') \, \widetilde{\zeta}_z(dy),$$
$$\overline{Q}_\varepsilon(dz, dz', dy) := \pi_\varepsilon(dz, dz') \, \widetilde{\zeta}_{z'}(dy).$$

Their y-marginals are exactly the desired mixtures:

$$(\overline{P}_\varepsilon)_Y = \Psi^u_{\#}\mathcal{M}^\theta_{s_i}, \quad (\overline{Q}_\varepsilon)_Y = \Psi^u_{\#}\mathcal{M}^\theta_{s_j}.$$

Since marginalization is a data-independent post-processing, DPI implies

$$\mathtt{D}_\alpha\Big(\Psi^u_{\#}\mathcal{M}^\theta_{s_i} \,\Big\|\, \Psi^u_{\#}\mathcal{M}^\theta_{s_j}\Big) \le \mathtt{D}_\alpha(\overline{P}_\varepsilon \| \overline{Q}_\varepsilon).$$

Moreover, conditioning on (z, z') and using the standard expression for Rényi divergence of a mixture with a common mixing law,

$$\mathtt{D}_\alpha(\overline{P}_\varepsilon \| \overline{Q}_\varepsilon) = \frac{1}{\alpha - 1} \log \int \exp\Big((\alpha - 1)\mathtt{D}_\alpha(\widetilde{\zeta}_z \| \widetilde{\zeta}_{z'})\Big) \, \pi_\varepsilon(dz, dz').$$

Using the pointwise bound from Step 2 and the monotonicity of $r \mapsto R_\alpha(\widetilde{\zeta}, r)$ gives

$$\mathtt{D}_\alpha(\overline{P}_\varepsilon \| \overline{Q}_\varepsilon) \le R_\alpha(\widetilde{\zeta}, z^u_\infty + \varepsilon).$$

Therefore,

$$\mathtt{D}_\alpha\Big(\Psi^u_{\#}\mathcal{M}^\theta_{s_i} \,\Big\|\, \Psi^u_{\#}\mathcal{M}^\theta_{s_j}\Big) \le R_\alpha(\widetilde{\zeta}, z^u_\infty + \varepsilon).$$

Finally, letting $\varepsilon \downarrow 0$ yields

$$\mathtt{D}_\alpha\Big(\Psi^u_{\#}\mathcal{M}^\theta_{s_i} \,\Big\|\, \Psi^u_{\#}\mathcal{M}^\theta_{s_j}\Big) \le R_\alpha(\widetilde{\zeta}, z^u_\infty).$$

Renaming $\widetilde{\zeta} = \Psi^u_{\#}\zeta$ (as in the lemma statement) proves the claim. □

L Proof of Theorem 4.5

Fix $\theta \in \Theta$ and $(s_i, s_j) \in \mathcal{Q}$ with $P_\theta^S(s_i), P_\theta^S(s_j) > 0$. Let $\zeta = \mathcal{N}(0, \sigma^2 I_d)$ and define

$$\mathrm{P} := \mathcal{M}^\theta_{s_i} = P_\theta^{f,s_i} * \zeta, \quad \mathrm{Q} := \mathcal{M}^\theta_{s_j} = P_\theta^{f,s_j} * \zeta.$$

Step 1 (per-slice envelope bound). Fix any direction $u \in \mathcal{U}$. Applying Lemma 4.1 to the 1-D projection $\Psi^u(y) = \langle y, u \rangle$ yields

$$\mathtt{D}_\alpha\big(\Psi^u_{\#}\mathrm{P} \,\big\|\, \Psi^u_{\#}\mathrm{Q}\big) \le R_\alpha\big(\zeta, \, z^u_\infty(\theta, s_i, s_j)\big),$$

where

$$z^u_\infty(\theta, s_i, s_j) := W_\infty\Big(\Psi^u_{\#}P_\theta^{f,s_i}, \, \Psi^u_{\#}P_\theta^{f,s_j}\Big).$$

By the definition of the per-slice ∞-Wasserstein sensitivity Δ^u_∞ in (6), for every $u \in \mathcal{U}$ we have

$$z^u_\infty(\theta, s_i, s_j) \le \Delta^u_\infty,$$

and since $R_\alpha(\zeta, \cdot)$ is non-decreasing,

$$\mathtt{D}_\alpha\big(\Psi^u_{\#}\mathrm{P} \,\big\|\, \Psi^u_{\#}\mathrm{Q}\big) \le R_\alpha\big(\zeta, \, \Delta^u_\infty\big). \tag{54}$$

Step 2 (Gaussian envelope). For $\zeta = \mathcal{N}(0, \sigma^2 I_d)$, the projected noise $\Psi^u(N)$ is $\mathcal{N}(0, \sigma^2)$. For any shift $a \in \mathbb{R}$,

$$\mathtt{D}_\alpha\big(\mathcal{N}(a, \sigma^2) \,\big\|\, \mathcal{N}(0, \sigma^2)\big) = \frac{\alpha a^2}{2\sigma^2}.$$

Hence the 1-D Gaussian shift envelope satisfies

$$R_\alpha(\zeta,z)=\sup_{|a|\le z}\mathtt{D}_\alpha\big(\mathcal{N}(a,\sigma^2)\,\|\,\mathcal{N}(0,\sigma^2)\big)=\frac{\alpha}{2\sigma^2}z^2.$$

Substituting this into (54) gives, for all $u\in\mathcal{U}$,

$$\mathtt{D}_\alpha\big(\Psi^u_\#\mathrm{P}\,\|\,\Psi^u_\#\mathrm{Q}\big)\le\frac{\alpha}{2\sigma^2}\,(\Delta^u_\infty)^2.$$

Step 3 (average over ω and choose σ^2). By Definition 4.1,

$$\mathtt{AveSD}^\omega_\alpha(\mathrm{P}\|\mathrm{Q})=\int_{\mathbb{S}^{d-1}}\mathtt{D}_\alpha\big(\Psi^u_\#\mathrm{P}\,\|\,\Psi^u_\#\mathrm{Q}\big)\,d\omega(u).$$

Therefore,

$$\mathtt{AveSD}^\omega_\alpha(\mathrm{P}\|\mathrm{Q})\le\int_{\mathbb{S}^{d-1}}\frac{\alpha}{2\sigma^2}(\Delta^u_\infty)^2\,d\omega(u)=\frac{\alpha}{2\sigma^2}\,\overline{\Delta}^2.$$

Choosing $\sigma^2=\frac{\alpha\overline{\Delta}^2}{2\varepsilon}$ yields $\mathtt{AveSD}^\omega_\alpha(\mathcal{M}^\theta_{s_i}\|\mathcal{M}^\theta_{s_j})\le\varepsilon$ for all θ and (s_i,s_j), i.e., $\mathcal{M}$ satisfies $(\alpha,\varepsilon,\omega)$-Ave-SRPP. □

M Proof of Theorem 4.6

Fix $\theta\in\Theta$ and $(s_i,s_j)\in Q$ with $P^S_\theta(s_i),P^S_\theta(s_j)>0$, and fix a slice profile $\{\mathcal{U},\omega\}$. Let

$$\mathrm{P}:=\mathcal{M}^\theta_{s_i}=P^{f,s_i}_\theta*\zeta,\quad\text{and}\quad\mathrm{Q}:=\mathcal{M}^\theta_{s_j}=P^{f,s_j}_\theta*\zeta.$$

Step 1. Fix a slice $u\in\mathcal{U}$. Lemma 4.1 applied to the 1-D projection $\Psi^u(y)=\langle y,u\rangle$ states that, for all $\alpha>1$,

$$\mathtt{D}_\alpha\big(\Psi^u_\#\mathrm{P}\,\|\,\Psi^u_\#Q\big)\le R_\alpha\big(\zeta,z^u_\infty(\theta,s_i,s_j)\big),$$

where $z^u_\infty(\theta,s_i,s_j)$ upper-bounds the 1-D shift (in W_∞) between the pre-noise probability measures $\Psi^u_\#P^{f,s_i}_\theta$ and $\Psi^u_\#P^{f,s_j}_\theta$. By definition,

$$z^u_\infty(\theta,s_i,s_j)\le\tilde\Delta^u_\infty(\theta,s_i,s_j):=W_\infty\big(\Psi^u_\#P^{f,s_i}_\theta\,\|\,\Psi^u_\#P^{f,s_j}_\theta\big),$$

and the envelope $R_\alpha(\zeta,\cdot)$ is non-decreasing in its second argument. Hence

$$\mathtt{D}_\alpha\big(\Psi^u_\#\mathrm{P}\,\|\,\Psi^u_\#\mathrm{Q}\big)\le R_\alpha\big(\zeta,\tilde\Delta^u_\infty(\theta,s_i,s_j)\big)\tag{55}$$

for all $u\in\mathcal{U}$.

Step 2. Now, we specialize to Gaussian noise $\zeta=\mathcal{N}(0,\sigma^2I_d)$. For any fixed slice $u\in\mathcal{U}$, the projected noise $\Psi^u(N)$ is 1-D Gaussian $\mathcal{N}(0,\sigma^2)$. If we shift by $a\in\mathbb{R}$, the noise becomes $\mathcal{N}(a,\sigma^2)$, and the order-α Rényi divergence between $\mathcal{N}(a,\sigma^2)$ and $\mathcal{N}(0,\sigma^2)$ is

$$\mathtt{D}_\alpha\big(\mathcal{N}(a,\sigma^2)\,\|\,\mathcal{N}(0,\sigma^2)\big)=\frac{\alpha a^2}{2\sigma^2}.$$

Therefore the 1-D Gaussian shift envelope is

$$R_\alpha(\zeta,z):=\sup_{|a|\le z}\mathtt{D}_\alpha\big(\zeta-a\,\|\,\zeta\big)=\sup_{|a|\le z}\frac{\alpha a^2}{2\sigma^2}=\frac{\alpha}{2\sigma^2}z^2.$$

Substituting $z=\tilde\Delta^u_\infty(\theta,s_i,s_j)$ into (55), we obtain

$$\mathtt{D}_\alpha\big(\Psi^u_\#\mathrm{P}\,\|\,\Psi^u_\#\mathrm{Q}\big)\le\frac{\alpha}{2\sigma^2}\,\big(\tilde\Delta^u_\infty(\theta,s_i,s_j)\big)^2.\tag{56}$$

By Definition 4.2, the joint sliced Rényi divergence of order $\alpha>1$ with slicing distribution ω is

$$\mathtt{JSD}^\omega_\alpha(\mathrm{P}\|\mathrm{Q})$$
$$:=\frac{1}{\alpha-1}\log\int_{\mathbb{S}^{d-1}}\exp\Big((\alpha-1)\,\mathtt{D}_\alpha\big(\Psi^u_\#\mathrm{P}\,\|\,\Psi^u_\#\mathrm{Q}\big)\Big)d\omega(u).$$

Using the bound (56), we get for each u

$$\mathtt{D}_\alpha\big(\Psi^u_\#\mathrm{P}\,\|\,\Psi^u_\#\mathrm{Q}\big)\le\frac{\alpha}{2\sigma^2}\,\big(\tilde\Delta^u_\infty(\theta,s_i,s_j)\big)^2\le\frac{\alpha}{2\sigma^2}\,\Delta^2_\star(\theta,s_i,s_j),$$

where

$$\Delta^2_\star(\theta,s_i,s_j):=\sup_{u\in\mathcal{U}}\big(\tilde\Delta^u_\infty(\theta,s_i,s_j)\big)^2.$$

Therefore, we have

$$\mathtt{JSD}^\omega_\alpha(\mathrm{P}\|\mathrm{Q})$$
$$=\frac{1}{\alpha-1}\log\int_{\mathbb{S}^{d-1}}\exp\Big((\alpha-1)\,\mathtt{D}_\alpha\big(\Psi^u_\#\mathrm{P}\,\|\,\Psi^u_\#\mathrm{Q}\big)\Big)d\omega(u)$$
$$\le\frac{1}{\alpha-1}\log\int_{\mathbb{S}^{d-1}}\exp\Big((\alpha-1)\,\tfrac{\alpha}{2\sigma^2}\,\Delta^2_\star(\theta,s_i,s_j)\Big)d\omega(u).$$

Here, the integrand is independent of u, so the integral is just the constant itself:

$$\int_{\mathbb{S}^{d-1}}\exp\Big((\alpha-1)\,\tfrac{\alpha}{2\sigma^2}\,\Delta^2_\star(\theta,s_i,s_j)\Big)d\omega(u)$$
$$=\exp\Big((\alpha-1)\,\tfrac{\alpha}{2\sigma^2}\,\Delta^2_\star(\theta,s_i,s_j)\Big).$$

Hence

$$\mathtt{JSD}^\omega_\alpha(\mathrm{P}\|\mathrm{Q})\le\frac{1}{\alpha-1}\log\exp\Big((\alpha-1)\,\tfrac{\alpha}{2\sigma^2}\Delta^2_\star(\theta,s_i,s_j)\Big)$$
$$=\frac{\alpha}{2\sigma^2}\Delta^2_\star(\theta,s_i,s_j).$$

Step 3. By definition of the global constant $\Delta^2_\star$, we have

$$\Delta^2_\star(\theta,s_i,s_j)\le\Delta^2_\star$$

for all $\theta\in\Theta$, $(s_i,s_j)\in Q$. Thus

$$\mathtt{JSD}^\omega_\alpha\big(\mathcal{M}^\theta_{s_i}\,\|\,\mathcal{M}^\theta_{s_j}\big)\le\frac{\alpha}{2\sigma^2}\,\Delta^2_\star.$$

With the choice

$$\sigma^2=\frac{\alpha\Delta^2_\star}{2\varepsilon},$$

we obtain

$$\mathtt{JSD}^\omega_\alpha\big(\mathcal{M}^\theta_{s_i}\,\|\,\mathcal{M}^\theta_{s_j}\big)\le\varepsilon$$

for every $\theta\in\Theta$ and $(s_i,s_j)\in Q$ with positive prior mass. This is exactly the $(\alpha,\varepsilon,\omega)$-Joint-SRPE guarantee in $(\mathcal{S},Q,\Theta)$.

N Proof of Theorem 4.7

Let $U_{1:m} \overset{\text{iid}}{\sim} \omega$. Define the random variables

$$X_\ell := \left(\Delta_\infty^{U_\ell}\right)^2 \in [0, \Delta_0^2], \qquad \ell = 1, \dots, m,$$

and recall

$$\overline{\Delta}^2(\omega) := \mathbb{E}_{U\sim\omega}\left[(\Delta_\infty^U)^2\right] = \mathbb{E}[X_1].$$

Step 1. Define the event

$$\mathcal{E}_{\text{dist}} := \left\{\Delta_\infty^{U_\ell} \le \widehat{\Delta}_\infty(U_\ell) \text{ for all } \ell \in [m]\right\}.$$

By assumption,

$$\Pr(\mathcal{E}_{\text{dist}}) \ge 1 - \gamma/2. \tag{57}$$

On $\mathcal{E}_{\text{dist}}$,

$$\frac{1}{m}\sum_{\ell=1}^m X_\ell = \frac{1}{m}\sum_{\ell=1}^m \left(\Delta_\infty^{U_\ell}\right)^2 \le \frac{1}{m}\sum_{\ell=1}^m \left(\widehat{\Delta}_\infty(U_\ell)\right)^2. \tag{58}$$

Step 2. Since $X_1, \dots, X_m$ are i.i.d. and bounded in $[0, \Delta_0^2]$, Hoeffding's inequality yields that for any $\eta \in (0,1)$,

$$\Pr\left(\mathbb{E}[X_1] \le \frac{1}{m}\sum_{\ell=1}^m X_\ell + \Delta_0^2\sqrt{\frac{\log(2/\eta)}{2m}}\right) \ge 1 - \eta.$$

Setting $\eta = \gamma/2$ and defining the event

$$\mathcal{E}_{\text{mc}} := \left\{\overline{\Delta}^2(\omega) \le \frac{1}{m}\sum_{\ell=1}^m X_\ell + \Delta_0^2\sqrt{\frac{\log(4/\gamma)}{2m}}\right\},$$

we obtain

$$\Pr(\mathcal{E}_{\text{mc}}) \ge 1 - \gamma/2. \tag{59}$$

Step 3. On the intersection event $\mathcal{E}_{\text{dist}} \cap \mathcal{E}_{\text{mc}}$, combining the defining inequality of $\mathcal{E}_{\text{mc}}$ with (58) gives

$$\overline{\Delta}^2(\omega) \le \frac{1}{m}\sum_{\ell=1}^m \left(\widehat{\Delta}_\infty(U_\ell)\right)^2 + \Delta_0^2\sqrt{\frac{\log(4/\gamma)}{2m}}. \tag{60}$$

By the union bound together with (57) and (59),

$$\Pr(\mathcal{E}_{\text{dist}} \cap \mathcal{E}_{\text{mc}}) \ge 1 - \gamma. \tag{61}$$

Step 4. Assume the mechanism uses $N \sim \mathcal{N}(0, \sigma^2 I_d)$ with

$$\sigma^2 \ge \frac{\alpha}{2\varepsilon}\left(\frac{1}{m}\sum_{\ell=1}^m \left(\widehat{\Delta}_\infty(U_\ell)\right)^2 + \Delta_0^2\sqrt{\frac{\log(4/\gamma)}{2m}}\right).$$

On $\mathcal{E}_{\text{dist}} \cap \mathcal{E}_{\text{mc}}$, inequality (60) implies

$$\sigma^2 \ge \frac{\alpha}{2\varepsilon}\overline{\Delta}^2(\omega).$$

Therefore, by Theorem 4.5 (Gaussian calibration for $(\alpha, \varepsilon, \omega)$-Ave-SRPP), the mechanism $\mathcal{M}(X) = f(X) + N$ satisfies $(\alpha, \varepsilon, \omega)$-*Ave-SRPP* on this event. Finally, (61) gives that this conclusion holds with probability at least $1 - \gamma$ over both the conditional sampling (used to form $\widehat{\Delta}_\infty$) and the draw of $U_{1:m}$. □

O Proof of Theorem 4.8

Fix $\sigma^2 > 0$ and abbreviate $c := c(\sigma) = \frac{\alpha(\alpha-1)}{2\sigma^2}$ and $b := b(\sigma) = \exp(c\Delta_0^2)$. Let $U_{1:m} \overset{\text{iid}}{\sim} \omega$ and define the random variables

$$Y_\ell := \exp\left(c\,(\Delta_\infty^{U_\ell})^2\right), \qquad \ell = 1, \dots, m.$$

Since $0 \le \Delta_\infty^u \le \Delta_0$ for all u, we have $Y_\ell \in [1, b]$ almost surely, and

$$\mu(\sigma) := \mathbb{E}_{U\sim\omega}\left[\exp\left(c\,(\Delta_\infty^U)^2\right)\right] = \mathbb{E}[Y_1].$$

Step 1. Define the event

$$\mathcal{E}_{\text{dist}} := \left\{\Delta_\infty^{U_\ell} \le \widehat{\Delta}_\infty(U_\ell) \text{ for all } \ell \in [m]\right\}.$$

By assumption, $\Pr(\mathcal{E}_{\text{dist}}) \ge 1 - \gamma/2$. On $\mathcal{E}_{\text{dist}}$, for each ℓ,

$$Y_\ell = \exp\left(c\,(\Delta_\infty^{U_\ell})^2\right) \le \exp\left(c\,(\widehat{\Delta}_\infty(U_\ell))^2\right),$$

and hence

$$\frac{1}{m}\sum_{\ell=1}^m Y_\ell \le \widehat{\mu}_m(\sigma) := \frac{1}{m}\sum_{\ell=1}^m \exp\left(c\left(\widehat{\Delta}_\infty(U_\ell)\right)^2\right). \tag{62}$$

Step 2. Since $Y_1, \dots, Y_m$ are i.i.d. and bounded in $[1, b]$, by Hoeffding's inequality, we have, for any $\eta \in (0,1)$,

$$\Pr\left(\mu(\sigma) \le \frac{1}{m}\sum_{\ell=1}^m Y_\ell + (b-1)\sqrt{\frac{\log(2/\eta)}{2m}}\right) \ge 1 - \eta.$$

Setting $\eta = \gamma/2$ gives the stated $\log(4/\gamma)$ term. Set $\eta = \gamma/2$ and define the event

$$\mathcal{E}_{\text{mc}} := \left\{\mu(\sigma) \le \frac{1}{m}\sum_{\ell=1}^m Y_\ell + (b-1)\sqrt{\frac{\log(4/\gamma)}{2m}}\right\}.$$

Then $\Pr(\mathcal{E}_{\text{mc}}) \ge 1 - \gamma/2$.

Step 3. On $\mathcal{E}_{\text{dist}} \cap \mathcal{E}_{\text{mc}}$, combining the defining inequality of $\mathcal{E}_{\text{mc}}$ with (62) yields

$$\mu(\sigma) \le \widehat{\mu}_m(\sigma) + (b-1)\sqrt{\frac{\log(4/\gamma)}{2m}}. \tag{63}$$

Moreover, by the union bound,

$$\Pr(\mathcal{E}_{\text{dist}} \cap \mathcal{E}_{\text{mc}}) \ge 1 - \gamma.$$

Step 4. For Gaussian SWM with noise $N \sim \mathcal{N}(0, \sigma^2 I_d)$, the joint envelope equals

$$\text{JR}_{\alpha,\omega}^{\infty} = \frac{1}{\alpha - 1}\log \mu(\sigma).$$

On $\mathcal{E}_{\text{dist}} \cap \mathcal{E}_{\text{mc}}$, inequality (63) implies

$$\mathrm{JR}^{\infty}_{\alpha,\omega} \le \frac{1}{\alpha-1}\log\left(\widehat{\mu}_m(\sigma) + (b-1)\sqrt{\frac{\log(4/\gamma)}{2m}}\right).$$

Hence, if σ^2 satisfies the stated condition

$$\frac{1}{\alpha-1}\log\left(\widehat{\mu}_m(\sigma) + (b-1)\sqrt{\frac{\log(4/\gamma)}{2m}}\right) \le \varepsilon,$$

then on $\mathcal{E}_{\text{dist}} \cap \mathcal{E}_{\text{mc}}$ we have $\mathrm{JR}^{\infty}_{\alpha,\omega} \le \varepsilon$. Therefore, by the Gaussian Joint-SRPP calibration result (Theorem 4.6 stated for SRPP), the mechanism $\mathcal{M}(X) = f(X) + \mathcal{N}(0, \sigma^2 I_d)$ satisfies $(\alpha, \varepsilon, \omega)$-*Joint-SRPP* on this event. Since $\Pr(\mathcal{E}_{\text{dist}} \cap \mathcal{E}_{\text{mc}}) \ge 1-\gamma$, the claim follows. □

P Formal Version and Proof of Proposition 4.1

We compare the cost of evaluating (i) sliced sensitivities Δ^u_∞ over finitely many directions versus (ii) the unsliced d-dimensional empirical W_∞ sensitivity.

Assume that for each (θ, s) we have n samples from $P^{f,s}_\theta$, and let $\mathsf{M} := |\widehat{\Theta}|\,|\widehat{\mathcal{Q}}|$ denote the number of $(\theta, (s_i, s_j))$ instances. Fix L directions $U_1, \dots, U_L \in \mathbb{S}^{d-1}$.

Lemma P.1 (Exact empirical W_∞ on $\mathbb{R}$)**.** *Let $a_1, \dots, a_n \in \mathbb{R}$ and $b_1, \dots, b_n \in \mathbb{R}$ and define $\widehat{P} := \frac{1}{n}\sum_{k=1}^n \delta_{a_k}$ and $\widehat{Q} := \frac{1}{n}\sum_{k=1}^n \delta_{b_k}$. Let $a_{(1)} \le \cdots \le a_{(n)}$ and $b_{(1)} \le \cdots \le b_{(n)}$ be the order statistics. Then*

$$W_\infty(\widehat{P}, \widehat{Q}) = \max_{1 \le k \le n} |a_{(k)} - b_{(k)}|.$$

Proof. Define the coupling $\pi := \frac{1}{n}\sum_{k=1}^n \delta_{(a_{(k)}, b_{(k)})} \in \Pi(\widehat{P}, \widehat{Q})$. Then $\operatorname{ess\,sup}_{(A,B)\sim\pi} |A - B| = \max_k |a_{(k)} - b_{(k)}|$, hence $W_\infty(\widehat{P}, \widehat{Q}) \le \max_k |a_{(k)} - b_{(k)}|$.

For the reverse inequality, let $t^\star := \max_k |a_{(k)} - b_{(k)}|$ and fix $t < t^\star$. Choose k such that $|a_{(k)} - b_{(k)}| > t$. Assume $a_{(k)} < b_{(k)} - t$ (the other case is symmetric) and set $A := (-\infty, a_{(k)}]$ and $B := [a_{(k)} + t, \infty)$. Then $\widehat{P}(A) = k/n$ and $\widehat{Q}(B) \ge (n - k + 1)/n$ since $b_{(k)} \in B$ implies $b_{(k)}, \dots, b_{(n)} \in B$. Therefore, under any coupling $\pi' \in \Pi(\widehat{P}, \widehat{Q})$, at least $1/n$ mass must be transported from A to B, so $\pi'(\{|x - y| > t\}) > 0$. Hence $\operatorname{ess\,sup}_{(X,Y)\sim\pi'} |X - Y| \ge t^\star$. Taking the infimum over π' yields $W_\infty(\widehat{P}, \widehat{Q}) \ge t^\star$. □

Lemma P.2 (Uniform empirical W_∞ reduces to a bottleneck matching problem)**.** *Let $\widehat{P} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\widehat{Q} = \frac{1}{n}\sum_{j=1}^n \delta_{y_j}$ on $\mathbb{R}^d$. For $t \ge 0$, define the bipartite graph G_t on $(\{1, \dots, n\}, \{1, \dots, n\})$ with an edge (i, j) iff $\|x_i - y_j\|_2 \le t$. Then*

$$W_\infty(\widehat{P}, \widehat{Q}) \le t \quad \Longleftrightarrow \quad G_t \text{ has a perfect matching.}$$

Consequently,

$$W_\infty(\widehat{P}, \widehat{Q}) = \min\{t \ge 0 : \ G_t \text{ has a perfect matching}\}.$$

Proof. ($\Leftarrow$) If G_t has a perfect matching, there exists a permutation σ such that $\|x_i - y_{\sigma(i)}\|_2 \le t$ for all i. Then $\pi := \frac{1}{n}\sum_{i=1}^n \delta_{(x_i, y_{\sigma(i)})} \in \Pi(\widehat{P}, \widehat{Q})$ is supported on $\{(x, y) : \|x - y\|_2 \le t\}$, so $W_\infty(\widehat{P}, \widehat{Q}) \le t$.

($\Rightarrow$) If $W_\infty(\widehat{P}, \widehat{Q}) \le t$, there exists $\pi \in \Pi(\widehat{P}, \widehat{Q})$ supported on $\{(x, y) : \|x - y\|_2 \le t\}$. Write $\pi = \frac{1}{n}\sum_{i,j} \pi_{ij}\, \delta_{(x_i, y_j)}$ where $\Pi := (\pi_{ij})$ is doubly stochastic. Support implies $\pi_{ij} = 0$ whenever (i, j) is not an edge of G_t. By the Birkhoff–von Neumann theorem, Π is a convex combination of permutation matrices, so there exists a permutation matrix supported on edges of G_t, i.e. G_t contains a perfect matching. □

Lemma P.3 (Elementary domination for asymptotic sums)**.** *For all $n \ge 3$, one has $n \le n \log n$ (with* log *the natural logarithm). Consequently, for any constants $c_1, c_2, c_3 \ge 0$ and all $n \ge 3$,*

$$c_1 dn + c_2 n \log n + c_3 n \ \le \ (c_1 + c_2 + c_3)\,(dn + n \log n).$$

Proof. For $n \ge 3$, $\log n \ge 1$, hence $n \le n \log n$. Thus $c_3 n \le c_3 n \log n$ and the inequality follows by grouping terms. □

Proposition P.1 (Formal Version of Proposition 4.1)**.** *Fix a Pufferfish scenario $(\mathcal{S}, \mathcal{Q}, \Theta)$.*

1. **Per-slice computation is exact in 1-D.** *For a fixed direction u, the empirical W_∞ between the one-dimensional projected samples $\{\langle x_k, u\rangle\}_{k=1}^n$ and $\{\langle y_k, u\rangle\}_{k=1}^n$ equals $\max_k |a_{(k)} - b_{(k)}|$ after sorting (Lemma P.1). Per instance, projection costs $O(dn)$ and sorting costs $O(n \log n)$, hence*
$$T_{\text{1D}}(u) = O\big(\mathsf{M}\,(dn + n \log n)\big).$$
2. **Finite slicing over L directions.** *Repeating (1) for $u = U_\ell$ over all $\ell \in [L]$ gives*
$$T_{\text{slice}} = O\big(\mathsf{M} L\,(dn + n \log n)\big).$$
3. **Unsliced ∞-Wasserstein.** *For each instance, forming all distances $\|x_i - y_j\|_2$ requires $\Theta(dn^2)$ arithmetic and $\Omega(n^2)$ memory to materialize the distance table. Moreover, exact evaluation of W_∞ can be performed by searching over thresholds t and checking perfect-matching feasibility in G_t (Lemma P.2). If $\mathcal{T} := \{\|x_i - y_j\|_2 : \ 1 \le i, j \le n\}$, then $|\mathcal{T}| \le n^2$ and binary search over $\mathcal{T}$ requires $O(\log n)$ feasibility checks. Using a standard bipartite matching algorithm with $O(n^3)$ time, the per-instance runtime is*
$$T_{\text{unsliced,exact}}(1 \text{ instance}) = O(dn^2 + n^3 \log n),$$
and over all M instances,
$$T_{\text{unsliced,exact}} = O\big(\mathsf{M}\,(dn^2 + n^3 \log n)\big),$$
with memory $\Omega(n^2)$ per instance if distances are materialized.

Proof. We count arithmetic operations and comparisons in the standard word-RAM model. Fix any instance $(\theta,(s_i,s_j)) \in \widehat{\Theta}\times\widehat{Q}$ and let $x_1,\dots,x_n \in \mathbb{R}^d$ and $y_1,\dots,y_n \in \mathbb{R}^d$ be the corresponding samples.

(1) Per-slice computation with one fixed direction u. Fix $u \in \mathbb{S}^{d-1}$. For each $k \in [n]$, compute $a_k := \langle x_k, u\rangle$ and $b_k := \langle y_k, u\rangle$. A dot product in $\mathbb{R}^d$ can be computed using exactly d multiplications and $d-1$ additions; thus there exists a constant $c_{\mathrm{dot}} > 0$ such that each dot product costs at most $c_{\mathrm{dot}}\, d$ primitive operations. Hence computing all $2n$ projections costs at most $2c_{\mathrm{dot}}\, dn$ operations.

Next, sort $\{a_k\}_{k=1}^n$ and $\{b_k\}_{k=1}^n$ to obtain order statistics $a_{(1)} \le \cdots \le a_{(n)}$ and $b_{(1)} \le \cdots \le b_{(n)}$. Using a standard comparison-based sorting algorithm (e.g., mergesort), sorting n reals costs at most $c_{\mathrm{sort}}\, n\log n$ comparisons for some constant $c_{\mathrm{sort}} > 0$ (for all $n \ge 2$). Thus sorting both lists costs at most $2c_{\mathrm{sort}}\, n\log n$ operations.

Define $\widehat{P}_u := \frac{1}{n}\sum_{k=1}^n \delta_{a_k}$ and $\widehat{Q}_u := \frac{1}{n}\sum_{k=1}^n \delta_{b_k}$. By Lemma P.1,

$$W_\infty(\widehat{P}_u, \widehat{Q}_u) = \max_{k\in[n]} |a_{(k)} - b_{(k)}|.$$

Computing the maximum can be done by one scan through $k = 1,\dots,n$, performing $O(1)$ operations per k; hence there exists $c_{\mathrm{scan}} > 0$ such that this step costs at most $c_{\mathrm{scan}}\, n$ operations.

Therefore, for this fixed instance and fixed u, the total cost is bounded by

$$T_{\mathrm{inst}}(u) \;\le\; 2c_{\mathrm{dot}}\, dn \;+\; 2c_{\mathrm{sort}}\, n\log n \;+\; c_{\mathrm{scan}}\, n.$$

For all $n \ge 3$, Lemma P.3 implies that

$$T_{\mathrm{inst}}(u) \le C_{\mathrm{inst}}\,(dn + n\log n), \quad C_{\mathrm{inst}} := 2c_{\mathrm{dot}} + 2c_{\mathrm{sort}} + c_{\mathrm{scan}}.$$

Hence $T_{\mathrm{inst}}(u) = O(dn + n\log n)$.

Now fix u and consider the M instances. Computing Δ_∞^u requires computing $W_\infty(\widehat{P}_u, \widehat{Q}_u)$ once per instance and taking their maximum. Maintaining a running maximum costs at most $c_{\max}$ operations per instance for some constant $c_{\max} > 0$, thus at most $c_{\max}\mathsf{M}$ operations in total. Consequently, for all $n \ge 3$,

$$T_{\mathrm{1D}}(u) \le \sum_{r=1}^{\mathsf{M}} T_{\mathrm{inst}}^{(r)}(u) \;+\; c_{\max}\mathsf{M} = O\big(\mathsf{M}(dn + n\log n)\big),$$

which proves (1).

(2) Finite slicing over L directions. For each $\ell \in [L]$, computing $\Delta_\infty^{U_\ell}$ is exactly the computation in (1) with $u = U_\ell$. Therefore,

$$T_{\mathrm{slice}} = \sum_{\ell=1}^{L} T_{\mathrm{1D}}(U_\ell) = O\big(\mathsf{M}L(dn + n\log n)\big),$$

which proves (2).

(3) Unsliced ∞-Wasserstein (exact computation). Fix one instance and define $\widehat{P} := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\widehat{Q} := \frac{1}{n}\sum_{j=1}^n \delta_{y_j}$. Compute the full distance table $D \in \mathbb{R}^{n\times n}$ with entries $D_{ij} := \|x_i - y_j\|_2$. Computing one Euclidean distance requires $\Theta(d)$ arithmetic operations; hence computing all n^2 distances requires $\Theta(dn^2)$ arithmetic operations. If the table is materialized, it stores n^2 reals and therefore requires $\Omega(n^2)$ memory.

Let $\mathcal{T} := \{D_{ij} : i, j \in [n]\}$ be the set of candidate thresholds, with $|\mathcal{T}| \le n^2$. By Lemma P.2,

$$W_\infty(\widehat{P}, \widehat{Q}) = \min\{t \in \mathcal{T} :\ G_t \text{ has a perfect matching}\}.$$

Sorting the multiset of at most n^2 values in $\mathcal{T}$ costs $O(n^2\log(n^2)) = O(n^2\log n)$, and binary search over the sorted list uses $O(\log|\mathcal{T}|) = O(\log n)$ feasibility checks. Given t, scanning D to determine edges of G_t costs $O(n^2)$ time, which is dominated by matching. A perfect matching in a bipartite graph on $2n$ vertices can be found in $O(n^3)$ time by standard algorithms. Thus each feasibility check costs $O(n^3)$ time, and the total time spent in feasibility checks is $O(n^3\log n)$.

Combining the distance-table computation, threshold sorting, and feasibility checks yields the per-instance bound

$$\begin{aligned} &T_{\mathrm{unsliced,exact}}(1 \text{ instance}) \\ &= O(dn^2) + O(n^2\log n) + O(n^3\log n) \\ &= O(dn^2 + n^3\log n). \end{aligned}$$

Over M instances this becomes $O(\mathsf{M}(dn^2 + n^3\log n))$, with $\Omega(n^2)$ memory per instance if D is materialized. This proves (3) and completes the proof. □

Q Proof of Proposition 5.1

Fix an iteration t and a slice profile $\{\mathcal{U}, \omega\}$ with $\mathcal{U} = \{u_i\}_{i=1}^m$. Fix arbitrarily a prior $\theta \in \Theta$, a secret pair $(s_i, s_j) \in \mathcal{Q}$ with $P_\theta^S(s_i), P_\theta^S(s_j) > 0$, a history $y_{<t}$ in the support, and a coupling $\gamma \in \Pi(\mu_{s_i}^\theta, \mu_{s_j}^\theta)$. Let $(X, X') \sim \gamma$ and $R_t \sim \mathbb{P}_{\eta,\rho}$ denote the minibatch sampling randomness at iteration t. For a draw r of R_t, let $\mathsf{I}_t(r) \subseteq [n]$ be the minibatch index (multi)set and $B_t(r) := |\mathsf{I}_t(r)|$ its size.

Step 1. By assumption, the minibatch size is deterministic, so we write $B_t := B_t(r) \ge 1$. For a dataset $x = (x_1,\dots,x_n)$, let $g_j(x)$ be the per-example gradient on x_j, and define the ℓ_2-clipped gradient

$$\tilde{g}_j(x) := g_j(x)\min\left\{1, \frac{C}{\|g_j(x)\|_2}\right\}.$$

Then $\|\tilde{g}_j(x)\|_2 \le C$ for all j and all x. The averaged clipped minibatch gradient at iteration t is

$$\bar{g}_t(\mathcal{B}(x;r)) := \frac{1}{B_t}\sum_{j\in\mathsf{I}_t(r)} \tilde{g}_j(x),$$

and similarly for x'. The discrepancy cap at iteration t (see (14)) is

$$K_t(x,x';r) := \sum_{j\in \mathsf{I}_t(r)} \mathbf{1}\{x_j \neq x'_j\}.$$

By definition, at most $K_t(x,x';r)$ indices in $\mathsf{I}_t(r)$ differ between x and x', so only those terms can change their contribution to the average of gradients.

Whenever $x_j \neq x'_j$, we have

$$\|\tilde{g}_j(x) - \tilde{g}_j(x')\|_2 \le \|\tilde{g}_j(x)\|_2 + \|\tilde{g}_j(x')\|_2 \le 2C.$$

Therefore,

$$\begin{aligned}\left\|\bar{g}_t(\mathcal{B}(x;r)) - \bar{g}_t(\mathcal{B}(x';r))\right\|_2 &\le \frac{1}{B_t} K_t(x,x';r)\,(2C)\\ &= \frac{2C}{B_t} K_t(x,x';r).\end{aligned}$$

For any slice $u_\ell \in \mathcal{U}$, this implies

$$\left|\langle \bar{g}_t(\mathcal{B}(x;r)) - \bar{g}_t(\mathcal{B}(x';r)), u_\ell\rangle\right| \le \frac{2C}{B_t} K_t(x,x';r). \tag{64}$$

Step 2. At iteration t, the pre-perturbation update is

$$f_t(x,y_{<t};r) = T_t\left(\bar{g}_t(\mathcal{B}(x;r)); y_{<t}\right),$$

and similarly $f_t(x',y_{<t};r) = T_t(\bar{g}_t(\mathcal{B}(x';r)); y_{<t})$. Let

$$z := \bar{g}_t(\mathcal{B}(x;r)), \qquad z' := \bar{g}_t(\mathcal{B}(x';r)).$$

By Assumption 5.2, for each $u_\ell \in \mathcal{U}$ there exists $L_{t,i} < \infty$ such that, for all $z, z' \in \mathbb{R}^d$ and all histories $y_{<t}$,

$$\left|u_\ell^\top \left(T_t(z;y_{<t}) - T_t(z';y_{<t})\right)\right| \le L_{t,i}\left|u_\ell^\top (z - z')\right|.$$

Applying this with the z, z' defined above and using (64), we obtain

$$\begin{aligned}\left|\langle f_t(x,y_{<t};r) - f_t(x',y_{<t};r), u_\ell\rangle\right| &\le L_{t,i}\,\frac{2C}{B_t} K_t(x,x';r)\\ &= \frac{2CL_{t,i}}{B_t} K_t(x,x';r).\end{aligned}$$

Squaring both sides yields

$$\left|\langle f_t(x,y_{<t};r) - f_t(x',y_{<t};r), u_\ell\rangle\right|^2 \le \left(\frac{2CL_{t,i}}{B_t} K_t(x,x';r)\right)^2.$$

Step 3. By definition of a feasible discrepancy cap at iteration t, K_t satisfies

$$K_t(x,x';r) \le K_t$$

for $\gamma \times \mathbb{P}_{\eta,\rho}$-almost every $((X,X'),R_t)$, uniformly over all $\theta \in \Theta$, $(s_i,s_j) \in Q$, histories $y_{<t}$ in support, and couplings γ. Thus, for $\gamma \times \mathbb{P}_{\eta,\rho}$-a.e. $((X,X'),R_t)$,

$$\left|\langle f_t(X,y_{<t};R_t) - f_t(X',y_{<t};R_t), u_\ell\rangle\right|^2 \le \left(\frac{2K_tL_{t,i}C}{B_t}\right)^2 =: h_{t,i}.$$

Since the bound holds for all $i = 1,\dots,m$ and is uniform over $\theta, (s_i,s_j), y_{<t}$, and γ, the vector $h_t = (h_{t,1},\dots,h_{t,m})$ with

$$h_{t,i} := \left(\frac{2K_tL_{t,i}C}{B_t}\right)^2$$

satisfies the HUC condition (13). Therefore h_t is a valid history-uniform cap at iteration t, as claimed. □

R Proof of Proposition 5.2

Fix an iteration t and a slice profile $\{\mathcal{U},\omega\}$ with $\mathcal{U} = \{u_i\}_{i=1}^m \subset \mathbb{S}^{d-1}$. Let $\theta \in \Theta$, $(s_i,s_j) \in Q$, a history $y_{<t}$ in the support, and a coupling $\gamma \in \Pi(\mu^\theta_{s_i}, \mu^\theta_{s_j})$ be arbitrary. Let $(X,X') \sim \gamma$. In addition, let $r_t \sim \mathbb{P}_{\eta,\rho}$ denote the subsampling randomness at iteration t. We write $\bar{g}_t(X;r_t)$ and $\bar{g}_t(X';r_t)$ for the clipped, averaged minibatch gradients, and $f_t(X,y_{<t};r_t)$, $f_t(X',y_{<t};r_t)$ for the corresponding pre-perturbation updates.

Step 1. By per-example ℓ_2-clipping at threshold C and the definition of the discrepancy $K_t(X,X';r_t)$, the argument in the proof of Proposition 5.1 gives, for every realization of r_t,

$$\left\|\bar{g}_t(X;r_t) - \bar{g}_t(X';r_t)\right\|_2 \le \frac{2C}{B_t} K_t(X,X';r_t).$$

Hence, for any slice $u_\ell \in \mathcal{U}$,

$$\left|\langle \bar{g}_t(X;r_t) - \bar{g}_t(X';r_t), u_\ell\rangle\right| \le \frac{2C}{B_t} K_t(X,X';r_t). \tag{65}$$

Define the pre-update difference

$$\Delta_t^{\mathrm{pre}}(X,X';r_t) := \bar{g}_t(X;r_t) - \bar{g}_t(X';r_t),$$

and the update difference

$$\Delta_t(X,X';r_t) := f_t(X,y_{<t};r_t) - f_t(X',y_{<t};r_t).$$

By Assumption 5.2, for each slice $u_\ell \in \mathcal{U}$ and every realization of r_t,

$$\left|\langle \Delta_t(X,X';r_t), u_\ell\rangle\right| \le L_{t,i}\left|\langle \Delta_t^{\mathrm{pre}}(X,X';r_t), u_\ell\rangle\right|.$$

Combining this with (65) yields the pathwise bound

$$\left|\langle \Delta_t(X,X';r_t), u_\ell\rangle\right| \le \frac{2L_{t,i}C}{B_t} K_t(X,X';r_t) \tag{66}$$

for all $i = 1,\dots,m$ and all η_t.

Step 2. Squaring both sides of (66) and taking expectations with respect to $\eta_t \sim \mathbb{P}_\eta$, we obtain, for each $i = 1,\dots,m$,

$$\mathbb{E}_{\eta_t}\left[\left|\langle \Delta_t(X,X';r_t), u_i\rangle\right|^2\right] \le \left(\frac{2L_{t,i}C}{B_t}\right)^2 \mathbb{E}_{\eta_t}\left[K_t(X,X';r_t)^2\right].$$

By the definition of the subsampling-aware discrepancy cap $\overline{K}_t^2$ in (17), we have, for γ-almost every (X,X'),

$$\mathbb{E}_{\eta_t}\left[K_t(X,X';r_t)^2\right] \le \overline{K}_t^2.$$

Therefore, for γ-almost every (X,X'),

$$\mathbb{E}_{\eta_t}\left[\left|\langle \Delta_t(X,X';r_t),u_\ell\rangle\right|^2\right] \le \left(\frac{2L_{t,i}C}{B_t}\right)^2 \overline{K}_t^2 = h_{t,i}^{\mathsf{sa}}, \quad i=1,\dots,m.$$

Step 3. Since $\theta\in\Theta$, $(s_i,s_j)\in\mathcal{Q}$, $y_{<t}$, and $\gamma\in\Pi(\mu_{s_i}^\theta,\mu_{s_j}^\theta)$ were arbitrary, the above bound holds for all admissible priors, secret pairs, histories, and couplings, and for γ-almost every (X,X'). By Definition 5.3 of a subsampling-aware history-uniform cap, this exactly means that

$$h_t^{\mathsf{sa}} = \left(h_{t,i}^{\mathsf{sa}}\right)_{i=1}^m \quad \text{and} \quad h_{t,i}^{\mathsf{sa}} = \left(\frac{2L_{t,i}C}{B_t}\right)^2 \overline{K}_t^2,$$

is a valid sa-HUC at iteration t. □

S Proof of Lemma 5.1

Fix t, a slice $u_i\in\mathbb{S}^{d-1}$, a prior $\theta\in\Theta$, and a secret pair $(s_i,s_j)\in\mathcal{Q}$ with $P_\theta^S(s_i),P_\theta^S(s_j)>0$. Let $\gamma\in\Pi(\mu_{s_i}^\theta,\mu_{s_j}^\theta)$ and draw $((X,X'),R_t)\sim\gamma\times\mathbb{P}_{\eta,\rho}$. Let $N_t\sim\mathcal{N}(0,\Sigma_t)$ be independent of $((X,X'),R_t)$ and set

$$\widetilde{Y}_t := \langle f_t(X,y_{<t};R_t),u_i\rangle + \langle N_t,u_i\rangle,$$
$$\widetilde{Y}_t' := \langle f_t(X',y_{<t};R_t),u_i\rangle + \langle N_t,u_i\rangle.$$

Then $\langle N_t,u_i\rangle\sim\mathcal{N}(0,v_{t,i})$ with $v_{t,i}=u_i^\top\Sigma_t u_i>0$.

Step 1. Let $S:=(X,X',R_t)$ and define the (random) mean shift

$$a(S) := \langle f_t(X,y_{<t};R_t)-f_t(X',y_{<t};R_t),u_i\rangle.$$

Conditioned on S, we have

$$\widetilde{Y}_t \mid S\sim\mathcal{N}(m(S)+a(S),v_{t,i}),\quad \widetilde{Y}_t'\mid S\sim\mathcal{N}(m(S),v_{t,i}),$$

for some $m(S)\in\mathbb{R}$ (which cancels in the divergence). Hence for every realization S,

$$\mathtt{D}_\alpha\left(\widetilde{Y}_t\mid S\,\middle\|\,\widetilde{Y}_t'\mid S\right) = \mathtt{D}_\alpha\left(\mathcal{N}(m+a,v_{t,i})\,\middle\|\,\mathcal{N}(m,v_{t,i})\right) = \frac{\alpha a(S)^2}{2v_{t,i}}.$$

By the HUC property (Definition 5.1), we have $a(S)^2\le h_{t,i}$ $\gamma\times\mathbb{P}_{\eta,\rho}$-a.s., therefore

$$\mathtt{D}_\alpha\left(\widetilde{Y}_t\mid S\,\middle\|\,\widetilde{Y}_t'\mid S\right)\le\frac{\alpha}{2v_{t,i}}h_{t,i}\quad\gamma\times\mathbb{P}_{\eta,\rho}\text{-a.s.}$$

Step 2. Let ν be the law of S under $\gamma\times\mathbb{P}_{\eta,\rho}$. For each $s\in\operatorname{supp}(\nu)$, write P_s as the probability distribution of $\widetilde{Y}_t$ given s and Q_s as the probability distribution of $\widetilde{Y}_t'$ given s. Then the unconditional one-dimensional output laws can be written as mixtures with the same weights:

$$P := \Pr(\langle Y_t,u_i\rangle\mid s_i,\theta) = \int P_s\,\nu(ds),$$
$$Q := \Pr(\langle Y_t,u_i\rangle\mid s_j,\theta) = \int Q_s\,\nu(ds).$$

By the standard Rényi-divergence bound for mixtures with common weights (equivalently, by applying data processing to the joint laws $\nu(ds)P_s(dy)$ and $\nu(ds)Q_s(dy)$ and then marginalizing out S), for any $\alpha>1$,

$$\mathtt{D}_\alpha(P\|Q)\le\operatorname*{ess\,sup}_s\mathtt{D}_\alpha(P_s\|Q_s).$$

Combining with the bound from Step 1 yields

$$\mathtt{D}_\alpha\left(\Pr(\langle Y_t,u_i\rangle\mid s_i,\theta)\,\middle\|\,\Pr(\langle Y_t,u_i\rangle\mid s_j,\theta)\right)\le\frac{\alpha}{2}\frac{h_{t,i}}{v_{t,i}}.$$

Since θ and (s_i,s_j) were arbitrary, the claim follows. □

T Proof of Theorem 5.6

Fix a slicing profile $\{\mathcal{U},\omega\}$ with $\mathcal{U}=\{u_\ell\}_{\ell=1}^m\subset\mathbb{S}^{d-1}$ and $\omega\in\Delta(\mathcal{U})$. Let $h=\{h_t\}_{t=1}^T$ be a sequence of valid HUC vectors, with $h_t=(h_{t,1},\dots,h_{t,m})$ for $t=1,\dots,T$. That is, for every prior $\theta\in\Theta$, every secret pair $(s_i,s_j)\in\mathcal{Q}$ with $P_\theta^S(s_i),P_\theta^S(s_j)>0$, every history $y_{<t}$ in the support, and every coupling $\gamma\in\Pi(\mu_{s_i}^\theta,\mu_{s_j}^\theta)$, we have $\gamma\times\mathbb{P}_{\eta,\rho}$-almost surely

$$\left|\langle f_t(X,y_{<t};R_t)-f_t(X',y_{<t};R_t),u_\ell\rangle\right|^2\le h_{t,\ell},\qquad(67)$$

for all $\ell=1,\dots,m$.

Let the SRPP-SGD mechanism use additive Gaussian noise

$$N_t\sim\mathcal{N}(0,\sigma^2I_d),\quad t=1,\dots,T,$$

independent across t and independent of (X,X',R_t), and write the noisy update at step t as

$$Y_t = f_t(X,y_{<t};R_t)+N_t.$$

Step 1. Fix $\alpha>1$, $\theta\in\Theta$, and $(s_i,s_j)\in\mathcal{Q}$ with positive prior mass. Fix an arbitrary history $y_{<t}$ in the support. Fix an arbitrary coupling $\gamma\in\Pi(\mu_{s_i}^\theta,\mu_{s_j}^\theta)$.

Since h_t is a valid HUC vector, for $\gamma\times\mathbb{P}_{\eta,\rho}$-a.s. in $((X,X'),R_t)$,

$$\left|\langle f_t(X,y_{<t};R_t)-f_t(X',y_{<t};R_t),u_\ell\rangle\right|^2\le h_{t,\ell},\quad\forall\ell\in[m].$$

By Fubini/Tonelli, there exists a measurable set $\mathcal{R}_t$ with $\Pr(R_t\in\mathcal{R}_t)=1$ such that for every $r\in\mathcal{R}_t$ the same bound holds γ-a.s. in (X,X'):

$$\left|\langle f_t(X,y_{<t};r)-f_t(X',y_{<t};r),u_\ell\rangle\right|^2\le h_{t,\ell}.$$

Condition on $R_t = r \in \mathcal{R}_t$. Then $\langle Y_t, u_\ell \rangle$ is Gaussian with variance $v_{t,\ell} = u_\ell^\top (\sigma^2 I_d) u_\ell = \sigma^2$ under both secrets, and the mean shift is $\langle f_t(X, y_{<t}; r) - f_t(X', y_{<t}; r), u_\ell \rangle$. Hence Lemma 5.1 yields, for all $r \in \mathcal{R}_t$,

$$\mathtt{D}_\alpha\Big(\Pr(\langle Y_t, u_\ell \rangle \mid s_i, \theta, y_{<t}, R_t = r) \Big\| \Pr(\langle Y_t, u_\ell \rangle \mid s_j, \theta, y_{<t}, R_t = r) \Big) \le \frac{\alpha}{2\sigma^2} h_{t,\ell}. \tag{68}$$

Now view $(R_t, \langle Y_t, u_\ell \rangle)$ as the released pair at round t. Because R_t is independent of the secret (and of $y_{<t}$), it has the same conditional law under s_i and s_j. Using (68) (which holds for R_t-a.e. r) and the definition of Rényi divergence, we obtain

$$\mathtt{D}_\alpha\Big(\Pr(R_t, \langle Y_t, u_\ell \rangle \mid s_i, \theta, y_{<t}) \Big\| \Pr(R_t, \langle Y_t, u_\ell \rangle \mid s_j, \theta, y_{<t}) \Big) \le \frac{\alpha}{2\sigma^2} h_{t,\ell}.$$

Finally, marginalizing out R_t is post-processing, so by data processing,

$$\mathtt{D}_\alpha\Big(\Pr(\langle Y_t, u_\ell \rangle \mid s_i, \theta, y_{<t}) \Big\| \Pr(\langle Y_t, u_\ell \rangle \mid s_j, \theta, y_{<t}) \Big) \le \frac{\alpha}{2\sigma^2} h_{t,\ell}. \tag{69}$$

This bound holds uniformly over all histories $y_{<t}$ in the support.

Step 2. Fix a slice index $\ell \in [m]$ and consider the 1-D projected mechanism

$$\mathcal{M}^{(\ell)} : X \longmapsto Z_{1:T}^{(\ell)} := \big(\langle Y_1, u_\ell \rangle, \ldots, \langle Y_T, u_\ell \rangle \big).$$

For each t, denote by $\mathrm{P}_{t,y_{<t}}^{(\ell)}$ and $\mathrm{Q}_{t,y_{<t}}^{(\ell)}$ the conditional probability measure of $\langle Y_t, u_\ell \rangle$ given $Z_{<t}^{(\ell)} = y_{<t}$ under secrets s_i and s_j, respectively. From the per–step bound (69), we have, for every history $y_{<t}$ in the support,

$$\mathtt{D}_\alpha\Big(\Pr\big(\langle Y_t, u_\ell \rangle \mid s_i, \theta, y_{<t}\big) \Big\| \Pr\big(\langle Y_t, u_\ell \rangle \mid s_j, \theta, y_{<t}\big) \Big) \le \frac{\alpha}{2\sigma^2} h_{t,\ell}.$$

Equivalently,

$$\mathtt{D}_\alpha\big(P_{t,y_{<t}}^{(\ell)} \,\big\|\, Q_{t,y_{<t}}^{(\ell)} \big) \le \varepsilon_{t,\ell}, \quad \varepsilon_{t,\ell} := \frac{\alpha}{2\sigma^2} h_{t,\ell}. \tag{70}$$

Let

$$\mathrm{P}^{(\ell)} := \Pr\big(Z_{1:T}^{(\ell)} \mid s_i, \theta\big), \quad \mathrm{Q}^{(\ell)} := \Pr\big(Z_{1:T}^{(\ell)} \mid s_j, \theta\big)$$

denote the trajectory probability measures of the projected mechanism along slice u_ℓ.

The family $\{\mathrm{P}_{t,y_{<t}}^{(\ell)}, \mathrm{Q}_{t,y_{<t}}^{(\ell)}\}_{t,y_{<t}}$ satisfies the assumptions of Lemma T.1 (stated later) with per–step bounds $\varepsilon_t = \varepsilon_{t,\ell}$. Applying Lemma T.1 yields

$$\mathtt{D}_\alpha\Big(\Pr\big(Z_{1:T}^{(\ell)} \mid s_i, \theta\big) \Big\| \Pr\big(Z_{1:T}^{(\ell)} \mid s_j, \theta\big) \Big) = \mathtt{D}_\alpha\big(\mathrm{P}^{(\ell)} \,\big\|\, \mathrm{Q}^{(\ell)} \big) \le \sum_{t=1}^{T} \varepsilon_{t,\ell} = \frac{\alpha}{2\sigma^2} \sum_{t=1}^{T} h_{t,\ell}. \tag{71}$$

Note that (71) is derived for the full projected trajectory $Z_{1:T}^{(\ell)}$. If Algorithm 1 outputs only the final iterate (or any measurable function of the trajectory), the corresponding Rényi divergence can only decrease by data processing, so the same bound (71) applies to the actual output distributions.

Step 3. By the definition of ω–Average Sliced Rényi Divergence (Definition 4.1), applied to the output probability measure of the mechanism (identified with the full projected trajectories), we have

$$\begin{aligned} &\mathtt{AveSD}_\alpha^\omega\Big(\Pr(M(X) \mid s_i, \theta) \Big\| \Pr(M(X) \mid s_j, \theta) \Big) \\ &= \sum_{\ell=1}^{m} \omega_\ell \mathtt{D}_\alpha\Big(\Pr\big(Z_{1:T}^{(\ell)} \mid s_i, \theta\big) \Big\| \Pr\big(Z_{1:T}^{(\ell)} \mid s_j, \theta\big) \Big). \end{aligned}$$

Using the per–slice bound (71), we get

$$\begin{aligned} &\mathtt{AveSD}_\alpha^\omega\Big(\Pr(M(X) \mid s_i, \theta) \Big\| \Pr(M(X) \mid s_j, \theta) \Big) \\ &\le \sum_{\ell=1}^{m} \omega_\ell \frac{\alpha}{2\sigma^2} \sum_{t=1}^{T} h_{t,\ell} \\ &= \frac{\alpha}{2\sigma^2} \sum_{t=1}^{T} \sum_{\ell=1}^{m} \omega_\ell h_{t,\ell}. \end{aligned}$$

Therefore, if

$$\sigma^2 \ge \frac{\alpha}{2\varepsilon} \sum_{t=1}^{T} \sum_{\ell=1}^{m} \omega_\ell h_{t,\ell},$$

then for every prior θ and every $(s_i, s_j) \in Q$ with positive prior mass, the ω–Ave-SRD between the corresponding output distributions is at most ε, i.e., Algorithm 1 is $(\alpha, \varepsilon, \omega)$–Ave-SRPP.

Step 4. For Joint-SRPP, we use the joint sliced Rényi divergence (Definition 4.3), which for the two output distributions can be written as

$$\begin{aligned} &\mathtt{JSD}_\alpha^\omega\Big(\Pr(M(X) \mid s_i, \theta) \Big\| \Pr(M(X) \mid s_j, \theta) \Big) \\ &= \frac{1}{\alpha - 1} \log \mathbb{E}_{U \sim \omega}\Big[\exp\big((\alpha - 1) \mathtt{D}_\alpha(\mathrm{P}_U \| \mathrm{Q}_U) \big) \Big], \end{aligned}$$

where, for each $u_\ell \in \mathcal{U}$, P_{u_ℓ} and Q_{u_ℓ} denote the distributions of $Z_{1:T}^{(\ell)}$ under s_i and s_j, respectively, and $\mathrm{P}_U, \mathrm{Q}_U$ are the corresponding random choices when $U \sim \omega$.

From (71), for each fixed slice u_ℓ we have

$$\mathtt{D}_\alpha(\mathrm{P}_{u_\ell} \| \mathrm{Q}_{u_\ell}) \le \frac{\alpha}{2\sigma^2} \sum_{t=1}^{T} h_{t,\ell}.$$

Thus, for the random $U \sim \omega$, it holds almost surely that

$$\mathtt{D}_\alpha(P_U \| Q_U) \le \frac{\alpha}{2\sigma^2} \sum_{t=1}^{T} h_{t,U} \le \frac{\alpha}{2\sigma^2} \sum_{t=1}^{T} \max_\ell h_{t,\ell}.$$

Plugging this into the definition of Joint-SRD yields

$$\begin{aligned}
&\mathtt{JSD}_\alpha^\omega\Big(\Pr(M(X)\mid s_i,\theta)\,\Big\|\,\Pr(M(X)\mid s_j,\theta)\Big)\\
&=\frac{1}{\alpha-1}\log\mathbb{E}_{U\sim\omega}\Big[\exp\big((\alpha-1)\mathtt{D}_\alpha(\mathrm{P}_U\|\mathrm{Q}_U)\big)\Big]\\
&\le\frac{1}{\alpha-1}\log\exp\Big((\alpha-1)\frac{\alpha}{2\sigma^2}\sum_{t=1}^{T}\max_\ell h_{t,\ell}\Big)\\
&=\frac{\alpha}{2\sigma^2}\sum_{t=1}^{T}\max_\ell h_{t,\ell}.
\end{aligned}$$

Therefore, if

$$\sigma^2\ge\frac{\alpha}{2\varepsilon}\sum_{t=1}^{T}\max_\ell h_{t,\ell},$$

then for every prior θ and every $(s_i,s_j)\in\mathcal{Q}$ the Joint-SRD between the corresponding output distributions is at most ε, i.e., Algorithm 1 is $(\alpha,\varepsilon,\omega)$–Joint-SRPP. □

Lemma T.1 (Adaptive composition of Rényi divergence)**.** *Let $\alpha>1$ and let P and Q be probability measures on the product space $\mathcal{Y}_1\times\cdots\times\mathcal{Y}_T$ with coordinate random variables $Y_1,\dots,Y_T$. For each $t=1,\dots,T$, let $\mathrm{P}_{t|y_{<t}}$ and $\mathrm{Q}_{t|y_{<t}}$ denote regular conditional distributions of Y_t given $Y_{<t}=y_{<t}$ under P and Q, respectively, where $y_{<t}=(y_1,\dots,y_{t-1})$. Suppose there exist constants $\varepsilon_t\ge0$ such that*

$$\mathtt{D}_\alpha\big(P_{t|y_{<t}}\,\big\|\,Q_{t|y_{<t}}\big)\le\varepsilon_t,$$

for all $t=1,\dots,T$ and all histories $y_{<t}$ in the support. Then the joint Rényi divergence satisfies

$$\mathtt{D}_\alpha(\mathrm{P}\,\|\,\mathrm{Q})\le\sum_{t=1}^{T}\varepsilon_t.$$

Proof of Lemma T.1. Let $L:=\frac{d\mathrm{P}}{d\mathrm{Q}}$ be the Radon–Nikodym derivative of P w.r.t. Q, assumed to exist (otherwise $\mathtt{D}_\alpha(\mathrm{P}\|\mathrm{Q})=+\infty$ and the inequality is straightforward). For each t, define the conditional likelihood ratio

$$L_t(y_{\le t}):=\frac{d\mathrm{P}_{t|y_{<t}}}{d\mathrm{Q}_{t|y_{<t}}}(y_t),\quad y_{\le t}=(y_1,\dots,y_t).$$

By the chain rule for densities,

$$L(Y_{1:T})=\prod_{t=1}^{T}L_t(Y_{\le t})$$

Q-almost surely.

By definition of Rényi divergence,

$$\begin{aligned}
\mathtt{D}_\alpha(\mathrm{P}\,\|\,\mathrm{Q})&=\frac{1}{\alpha-1}\log\mathbb{E}_{\mathrm{Q}}\big[L(Y_{1:T})^\alpha\big]\\
&=\frac{1}{\alpha-1}\log\mathbb{E}_{\mathrm{Q}}\Big[\prod_{t=1}^{T}L_t(Y_{\le t})^\alpha\Big].
\end{aligned}$$

For each t, define

$$M_t:=\sup_{y_{<t}}\mathbb{E}_{\mathrm{Q}}\big[L_t(Y_{\le t})^\alpha\,\big|\,Y_{<t}=y_{<t}\big].$$

Under our assumption,

$$\frac{1}{\alpha-1}\log\mathbb{E}_{\mathrm{Q}}\big[L_t(Y_{\le t})^\alpha\,\big|\,Y_{<t}=y_{<t}\big]=\mathtt{D}_\alpha\big(\mathrm{P}_{t|y_{<t}}\,\big\|\,\mathrm{Q}_{t|y_{<t}}\big)\le\varepsilon_t.$$

Thus,

$$\mathbb{E}_{\mathrm{Q}}\big[L_t(Y_{\le t})^\alpha\,\big|\,Y_{<t}=y_{<t}\big]\le e^{(\alpha-1)\varepsilon_t},$$

for all $y_{<t}$, and hence

$$M_t\le e^{(\alpha-1)\varepsilon_t}.$$

We now bound $\mathbb{E}_{\mathrm{Q}}[\prod_{t=1}^{T}L_t^\alpha]$ iteratively using the tower property:

$$\begin{aligned}
\mathbb{E}_{\mathrm{Q}}\Big[\prod_{t=1}^{T}L_t^\alpha\Big]&=\mathbb{E}_{\mathrm{Q}}\Big[\mathbb{E}_{\mathrm{Q}}\big[L_T^\alpha\,\big|\,Y_{<T}\big]\prod_{t=1}^{T-1}L_t^\alpha\Big]\\
&\le M_T\,\mathbb{E}_{\mathrm{Q}}\Big[\prod_{t=1}^{T-1}L_t^\alpha\Big].
\end{aligned}$$

Repeating this backward for $t=T-1,\dots,1$ yields

$$\mathbb{E}_{\mathrm{Q}}\Big[\prod_{t=1}^{T}L_t^\alpha\Big]\le\prod_{t=1}^{T}M_t\le\prod_{t=1}^{T}e^{(\alpha-1)\varepsilon_t}=e^{(\alpha-1)\sum_{t=1}^{T}\varepsilon_t}.$$

Taking $\frac{1}{\alpha-1}\log(\cdot)$ on both sides gives

$$\mathtt{D}_\alpha(\mathrm{P}\,\|\,\mathrm{Q})=\frac{1}{\alpha-1}\log\mathbb{E}_{\mathrm{Q}}\Big[\prod_{t=1}^{T}L_t^\alpha\Big]\le\sum_{t=1}^{T}\varepsilon_t,$$

as claimed.

□

U Proof of Theorem 5.7

Fix a slice profile $(\mathcal{U},\omega)$ with $\mathcal{U}=\{u_\ell\}_{\ell=1}^m$ and $\omega\in\Delta(\mathcal{U})$. Fix $\alpha>1$. Fix an arbitrary prior $\theta\in\Theta$ and an arbitrary secret pair $(s_i,s_j)\in\mathcal{Q}$.

Let $\eta=(\eta_1,\dots,\eta_T)$ denote the (possibly time-varying) subsampling parameters used by Algorithm 1 (e.g., sampling rates). Let R_t denote the *realized subsampling outcome* at round t (e.g., the sampled index multiset), and let $\mathbf{R}:=(R_1,\dots,R_T)$. For each realization $\mathbf{r}$ of $\mathbf{R}$, let $\mathcal{M}_s^{\theta,\mathbf{r}}$ denote the conditional law of the final output of Algorithm 1 under prior θ and secret s, conditional on $\mathbf{R}=\mathbf{r}$.

Throughout, the Gaussian noises are i.i.d. $N_t\sim\mathcal{N}(0,\sigma^2I_d)$, independent of all data and of $\mathbf{R}$. Hence for every slice u_ℓ,

$$v_{t,\ell}:=u_\ell^\top(\sigma^2I_d)u_\ell=\sigma^2.$$

Lemma U.1 (Expected Gaussian sliced Rényi cost)**.** *Fix iteration t and direction $u \in \mathbb{S}^{d-1}$. Let $N_t \sim \mathcal{N}(0, \sigma^2 I_d)$ be independent of all data and of R_t. Fix $\theta \in \Theta$, $(s_i, s_j) \in \mathcal{Q}$, a history $y_{<t}$ in the support, and a coupling $\gamma \in \Pi(\mu^{\theta}_{s_i}, \mu^{\theta}_{s_j})$. For γ-a.e. (X, X'), define the random directional shift*

$$\Delta_t^u(X, X'; R_t) := \big\langle f_t(X, y_{<t}; R_t) - f_t(X', y_{<t}; R_t), u \big\rangle.$$

Then, for any $\alpha > 1$,

$$\mathbb{E}_{R_t}\Big[\mathtt{D}_\alpha\Big(\Pr(\langle Y_t, u\rangle \mid s_i, \theta, R_t) \,\Big\|\, \Pr(\langle Y_t, u\rangle \mid s_j, \theta, R_t)\Big)\Big] = \frac{\alpha}{2\sigma^2}\,\mathbb{E}_{R_t}\Big[\big|\Delta_t^u(X, X'; R_t)\big|^2\Big],$$

where $Y_t = f_t(X, y_{<t}; R_t) + N_t$.

Proof. Fix $t, u, \theta, (s_i, s_j), y_{<t}, \gamma$ as stated and take γ-a.e. (X, X'). Condition on $R_t = r$. Then $\langle Y_t, u\rangle$ is Gaussian with variance σ^2 and mean $\langle f_t(X, y_{<t}; r), u\rangle$, while the corresponding mean under secret s_j is $\langle f_t(X', y_{<t}; r), u\rangle$. Hence the conditional mean shift equals $\Delta_t^u(X, X'; r)$ and the standard Gaussian identity gives

$$\mathtt{D}_\alpha\Big(\Pr(\langle Y_t, u\rangle \mid s_i, \theta, R_t = r) \,\Big\|\, \Pr(\langle Y_t, u\rangle \mid s_j, \theta, R_t = r)\Big) = \frac{\alpha}{2\sigma^2}\,\big|\Delta_t^u(X, X'; r)\big|^2.$$

Taking expectation over R_t yields the claim. □

Step 1. Fix an iteration $t \in \{1, \dots, T\}$ and a slice $u_\ell \in \mathcal{U}$. Fix any history $y_{<t}$ in the support and any coupling $\gamma \in \Pi(\mu^{\theta}_{s_i}, \mu^{\theta}_{s_j})$. Apply Lemma U.1 with $u = u_\ell$: for γ-a.e. (X, X'),

$$\mathbb{E}_{R_t}\Big[\mathtt{D}_\alpha\Big(\Pr(\langle Y_t, u_\ell\rangle \mid s_i, \theta, R_t) \,\Big\|\, \Pr(\langle Y_t, u_\ell\rangle \mid s_j, \theta, R_t)\Big)\Big] = \frac{\alpha}{2\sigma^2}\,\mathbb{E}_{R_t}\Big[\big|\langle f_t(X, y_{<t}; R_t) - f_t(X', y_{<t}; R_t), u_\ell\rangle\big|^2\Big].$$

Since h_t^{sa} is a valid sa-HUC vector (Definition 5.3), the expectation on the right-hand side is at most $h_{t,\ell}^{\mathsf{sa}}$. Therefore,

$$\mathbb{E}_{R_t}\Big[\mathtt{D}_\alpha\Big(\Pr(\langle Y_t, u_\ell\rangle \mid s_i, \theta, R_t) \,\Big\|\, \Pr(\langle Y_t, u_\ell\rangle \mid s_j, \theta, R_t)\Big)\Big] \le \frac{\alpha}{2\sigma^2}\,h_{t,\ell}^{\mathsf{sa}}. \tag{72}$$

Step 2 (sa-Ave-SRPP). Fix a realization $\mathbf{r} = (r_1, \dots, r_T)$ of the subsampling outcomes $\mathbf{R}$. Apply the same conditional SRPP accountant (composition argument) as used for Theorem 5.6 to the conditional mechanism $\mathcal{M}^{\theta,\mathbf{r}}$. This yields the pointwise (in $\mathbf{r}$) bound

$$\mathtt{AveSD}_\alpha^{\omega}\Big(\mathcal{M}_{s_i}^{\theta,\mathbf{r}} \,\Big\|\, \mathcal{M}_{s_j}^{\theta,\mathbf{r}}\Big) \le \sum_{t=1}^{T}\sum_{\ell=1}^{m}\omega_\ell\, \mathtt{D}_\alpha\Big(\Pr(\langle Y_t, u_\ell\rangle \mid s_i, \theta, R_t = r_t) \,\Big\|\, \Pr(\langle Y_t, u_\ell\rangle \mid s_j, \theta, R_t = r_t)\Big). \tag{73}$$

Take expectation over η and use linearity and the fact that the t-th summand depends only on η_t:

$$\mathbb{E}_\eta\Big[\mathtt{AveSD}_\alpha^{\omega}\Big(\mathcal{M}_{s_i}^{\theta,\eta} \,\Big\|\, \mathcal{M}_{s_j}^{\theta,\eta}\Big)\Big] \le \sum_{t=1}^{T}\sum_{\ell=1}^{m}\omega_\ell\,\mathbb{E}_{\eta_t}\Big[\mathtt{D}_\alpha\Big(\Pr(\langle Y_t, u_\ell\rangle \mid s_i, \theta, R_t) \,\Big\|\, \Pr(\langle Y_t, u_\ell\rangle \mid s_j, \theta, R_t)\Big)\Big].$$

Applying (72) yields

$$\mathbb{E}_\eta\Big[\mathtt{AveSD}_\alpha^{\omega}\Big(\mathcal{M}_{s_i}^{\theta,\eta} \,\Big\|\, \mathcal{M}_{s_j}^{\theta,\eta}\Big)\Big] \le \frac{\alpha}{2\sigma^2}\sum_{t=1}^{T}\sum_{\ell=1}^{m}\omega_\ell\, h_{t,\ell}^{\mathsf{sa}}.$$

Therefore, if

$$\sigma^2 \ge \frac{\alpha}{2\varepsilon}\sum_{t=1}^{T}\sum_{\ell=1}^{m}\omega_\ell\, h_{t,\ell}^{\mathsf{sa}},$$

then the right-hand side is at most ε. Since θ and (s_i, s_j) were arbitrary, Definition 5.4 implies that Algorithm 1 satisfies $(\alpha, \varepsilon, \omega)$-sa-Ave-SRPP, proving (i).

Step 3 (sa-Joint-SRPP). Fix a realization $\mathbf{r} = (r_1, \dots, r_T)$ of the subsampling outcomes $\mathbf{R}$. Apply the same conditional accountant as in Step 2 for the joint sliced divergence:

$$\mathtt{JSD}_\alpha^{\omega}\Big(\mathcal{M}_{s_i}^{\theta,\mathbf{r}} \,\Big\|\, \mathcal{M}_{s_j}^{\theta,\mathbf{r}}\Big) \le \sum_{t=1}^{T}\max_{\ell\in[m]}\mathtt{D}_\alpha\Big(\Pr(\langle Y_t, u_\ell\rangle \mid s_i, \theta, R_t = r_t) \,\Big\|\, \Pr(\langle Y_t, u_\ell\rangle \mid s_j, \theta, R_t = r_t)\Big), \tag{74}$$

Take expectation over r_t. and use linearity across t:

$$\mathbb{E}_\eta\Big[\mathtt{JSD}_\alpha^{\omega}\Big(\mathcal{M}_{s_i}^{\theta,\eta} \,\Big\|\, \mathcal{M}_{s_j}^{\theta,\eta}\Big)\Big] \le \sum_{t=1}^{T}\mathbb{E}_{\eta_t}\Big[\max_{\ell\in[m]}\mathtt{D}_\alpha\Big(\Pr(\langle Y_t, u_\ell\rangle \mid s_i, \theta, R_t) \,\Big\|\, \Pr(\langle Y_t, u_\ell\rangle \mid s_j, \theta, R_t)\Big)\Big].$$

Now invoke the explicit SRPP–SGD sa-HUC instantiation from Proposition 5.2. In that setting, for each t and ℓ there is a deterministic coefficient

$$a_{t,\ell} := \Big(\frac{2L_{t,\ell}C}{B_t}\Big)^2$$

and a (scalar) discrepancy random variable $K_t(R_t)$ (shared across slices at round t) such that for γ-a.e. (X, X') and all ℓ,

$$\big|\langle f_t(X, y_{<t}; R_t) - f_t(X', y_{<t}; R_t), u_\ell\rangle\big|^2 \le a_{t,\ell}\,K_t(R_t)^2.$$

Condition on $R_t = r$ and apply the same Gaussian identity as in Lemma U.1 to get

$$\mathtt{D}_\alpha\Big(\Pr(\langle Y_t, u_\ell\rangle \mid s_i, \theta, R_t = r) \,\Big\|\, \Pr(\langle Y_t, u_\ell\rangle \mid s_j, \theta, R_t = r)\Big) = \frac{\alpha}{2\sigma^2}\,\big|\langle f_t(X, y_{<t}; r) - f_t(X', y_{<t}; r), u_\ell\rangle\big|^2 \le \frac{\alpha}{2\sigma^2}\,a_{t,\ell}\,K_t(r)^2.$$

Taking the maximum over ℓ yields, for each r,

$$\max_{\ell\in[m]} \mathtt{D}_\alpha(\cdots \mid R_t = r) \le \frac{\alpha}{2\sigma^2}\left(\max_{\ell\in[m]} a_{t,\ell}\right) K_t(r)^2.$$

Now take expectation over R_t:

$$\mathbb{E}_{R_t}\left[\max_{\ell\in[m]} \mathtt{D}_\alpha(\cdots)\right] \le \frac{\alpha}{2\sigma^2}\left(\max_{\ell\in[m]} a_{t,\ell}\right)\mathbb{E}_{R_t}\left[K_t(R_t)^2\right].$$

By the subsampling-aware discrepancy cap (17), $\mathbb{E}_{R_t}[K_t(R_t)^2] \le \overline{K}_t^2$, hence

$$\mathbb{E}_{R_t}\left[\max_{\ell\in[m]} \mathtt{D}_\alpha(\cdots)\right] \le \frac{\alpha}{2\sigma^2}\left(\max_{\ell\in[m]} a_{t,\ell}\right)\overline{K}_t^2 = \frac{\alpha}{2\sigma^2}\max_{\ell\in[m]} h^{\mathsf{sa}}_{t,\ell},$$

where the last equality uses the closed form $h^{\mathsf{sa}}_{t,\ell} = a_{t,\ell}\overline{K}_t^2$ from Proposition 5.2. Substituting back,

$$\mathbb{E}_\eta\left[\mathtt{JSD}^\omega_\alpha\left(\mathcal{M}^{\theta,\eta}_{s_i} \,\middle\|\, \mathcal{M}^{\theta,\eta}_{s_j}\right)\right] \le \frac{\alpha}{2\sigma^2}\sum_{t=1}^{T}\max_{\ell\in[m]} h^{\mathsf{sa}}_{t,\ell}.$$

Therefore, if

$$\sigma^2 \ge \frac{\alpha}{2\varepsilon}\sum_{t=1}^{T}\max_{\ell\in[m]} h^{\mathsf{sa}}_{t,\ell},$$

then the right-hand side is at most ε. Since θ and (s_i,s_j) were arbitrary, Definition 5.5 implies that Algorithm 1 satisfies $(\alpha,\varepsilon,\omega)$-sa-Joint-SRPP, proving (ii). □

V Proof of Theorem 5.8

Fix a slice profile $\{\mathcal{U},\omega\}$ and a Pufferfish scenario $(\mathcal{S},\mathcal{Q},\Theta)$. Let $\vec{\mathcal{M}}$ be the composed mechanism in (20):

$$\vec{\mathcal{M}}(X) := \big(\mathcal{M}_1(X),\dots,\mathcal{M}_J(X)\big),$$

where X is the underlying common dataset. For every realized dataset x, we assume that the primitive randomness of $\mathcal{M}_1,\dots,\mathcal{M}_J$ is independent across j when the mechanisms are run on x.

Fix $\theta\in\Theta$ and $(s_i,s_j)\in\mathcal{Q}$ with $P^S_\theta(s_i),P^S_\theta(s_j)>0$. Let $\mu^\theta_{s_i}$ and $\mu^\theta_{s_j}$ be the corresponding dataset beliefs. For any fixed dataset realization x in the support of these beliefs, define the conditional output distribution

$$\mathrm{P}^x_j := \Pr\big(\mathcal{M}_j(x)\in\cdot\,\big|\,s_i,\theta\big),$$
$$\mathrm{Q}^x_j := \Pr\big(\mathcal{M}_j(x)\in\cdot\,\big|\,s_j,\theta\big),$$

and

$$\mathrm{P}^{\mathrm{comp},x} := \Pr\big(\vec{\mathcal{M}}(x)\in\cdot\,\big|\,s_i,\theta\big),$$
$$\mathrm{Q}^{\mathrm{comp},x} := \Pr\big(\vec{\mathcal{M}}(x)\in\cdot\,\big|\,s_j,\theta\big).$$

By independence of the primitive randomness across j for this fixed x, we have

$$\mathrm{P}^{\mathrm{comp},x} = \bigotimes_{j=1}^{J}\mathrm{P}^x_j, \quad\text{and}\quad \mathrm{Q}^{\mathrm{comp},x} = \bigotimes_{j=1}^{J}\mathrm{Q}^x_j.$$

For any Rényi order $\alpha>1$ and product measures $\bigotimes_{j=1}^J \mathrm{P}_j$ and $\bigotimes_{j=1}^J \mathrm{Q}_j$, the Rényi divergence becomes

$$\mathtt{D}_\alpha\Big(\bigotimes_{j=1}^J\mathrm{P}_j\,\Big\|\,\bigotimes_{j=1}^J\mathrm{Q}_j\Big) = \sum_{j=1}^{J}\mathtt{D}_\alpha(\mathrm{P}_j\|\mathrm{Q}_j). \tag{75}$$

Applying this with $\mathrm{P}_j=\mathrm{P}^x_j$ and $\mathrm{Q}_j=\mathrm{Q}^x_j$ yields, for every fixed dataset realization x,

$$\mathtt{D}_\alpha\big(\mathrm{P}^{\mathrm{comp},x}\big\|\mathrm{Q}^{\mathrm{comp},x}\big) = \sum_{j=1}^{J}\mathtt{D}_\alpha(\mathrm{P}^d_j\|\mathrm{Q}^x_j).$$

We now prove the composition bound for each SRPP flavor Ξ.

Case 1: $\Xi\in\{$Ave, sa-Ave$\}$. Fix $\theta\in\Theta$ and $(s_i,s_j)\in\mathcal{Q}$ with $P^S_\theta(s_i),P^S_\theta(s_j)>0$. Fix any coupling $\gamma\in\Pi(\mu^\theta_{s_i},\mu^\theta_{s_j})$ and take $(X,X')\sim\gamma$. For each j, write $W_j := \mathcal{M}_j(X)$ and $W'_j := \mathcal{M}_j(X')$.

Fix a slice $u_\ell\in\mathcal{U}$ and consider the 1-D release sequence

$$Z^{(\ell)}_{1:J} := \big(\langle W_1,u_\ell\rangle,\dots,\langle W_J,u_\ell\rangle\big),$$
$$Z'^{(\ell)}_{1:J} := \big(\langle W'_1,u_\ell\rangle,\dots,\langle W'_J,u_\ell\rangle\big).$$

Let $\mathrm{P}^{(\ell)} := \Pr(Z^{(\ell)}_{1:J}\in\cdot\mid s_i,\theta)$ and $\mathrm{Q}^{(\ell)} := \Pr(Z'^{(\ell)}_{1:J}\in\cdot\mid s_j,\theta)$.

Per-step conditional slice bounds. Because each $\mathcal{M}_j$ is an SRPP-SGD mechanism whose SRPP guarantee is obtained from an HUC/sa-HUC sequence (Theorem 5.6 or Theorem 5.7), its per-slice Rényi bound is uniform over all histories and (coupled) dataset pairs. In particular, for every $j\in[J]$, every $\ell\in[m]$, and every history $z^{(\ell)}_{<j}$ in the support, we have

$$\begin{aligned}\mathtt{D}_\alpha\Big(&\Pr(\langle W_j,u_\ell\rangle\in\cdot\mid s_i,\theta,Z^{(\ell)}_{<j}=z^{(\ell)}_{<j})\\ &\Big\|\ \Pr(\langle W'_j,u_\ell\rangle\in\cdot\mid s_j,\theta,Z'^{(\ell)}_{<j}=z^{(\ell)}_{<j})\Big)\le\varepsilon_{j,\ell},\end{aligned} \tag{76}$$

for some nonnegative numbers $\{\varepsilon_{j,\ell}\}_{\ell=1}^m$ satisfying

$$\sum_{\ell=1}^{m}\omega_\ell\,\varepsilon_{j,\ell}\le\varepsilon_j. \tag{77}$$

(Here $\varepsilon_{j,\ell}$ is the per-slice budget induced by the HUC/sa-HUC-based bound for $\mathcal{M}_j$.)

Applying Lemma T.1 to the sequence $Z^{(\ell)}_{1:J}$ (using the per-step bounds (76)) yields, for each fixed ℓ,

$$\mathtt{D}_\alpha(\mathrm{P}^{(\ell)}\|\mathrm{Q}^{(\ell)})\le\sum_{j=1}^{J}\varepsilon_{j,\ell}.$$

Now average over slices:

$$\begin{aligned}&\mathtt{AveSD}^\omega_\alpha\big(\Pr(\vec{\mathcal{M}}(X)\mid s_i,\theta)\,\big\|\,\Pr(\vec{\mathcal{M}}(X)\mid s_j,\theta)\big)\\ &=\sum_{\ell=1}^{m}\omega_\ell\,\mathtt{D}_\alpha(\mathrm{P}^{(\ell)}\|\mathrm{Q}^{(\ell)})\le\sum_{\ell=1}^{m}\omega_\ell\sum_{j=1}^{J}\varepsilon_{j,\ell}\\ &=\sum_{j=1}^{J}\sum_{\ell=1}^{m}\omega_\ell\varepsilon_{j,\ell}\le\sum_{j=1}^{J}\varepsilon_j,\end{aligned}$$

where the last inequality uses (77). Since θ and (s_i, s_j) were arbitrary, this proves $(\alpha, \sum_{j=1}^{J} \varepsilon_j, \omega)$-Ave-SRPP for $\vec{\mathcal{M}}$.

For sa-Ave-SRPP, the same argument applies verbatim after folding the subsampling randomness into each mechanism's output (i.e., applying the above proof to the expanded-output channels), so $\vec{\mathcal{M}}$ is also $(\alpha, \sum_{j=1}^{J} \varepsilon_j, \omega)$-sa-Ave-SRPP.

Case 2: $\Xi \in \{$Joint, sa-Joint$\}$. Fix x in the support of $\mu_{s_i}^{\theta}$ and $\mu_{s_j}^{\theta}$. Recall that Joint-SRD can be represented as the Rényi divergence of a randomized sliced channel. Define the (one-shot) slicing operator

$$\mathtt{Slice}(z) := (U, \langle z, U \rangle), \qquad U \sim \omega,$$

where U is independent of all other randomness. For a mechanism $\mathcal{M}$, write $\mathcal{N} := \mathtt{Slice} \circ \mathcal{M}$. Then, by the definition/representation of Joint-SRD (Definition 4.3 and App. B.3),

$$\begin{aligned} &\mathtt{JSD}_{\alpha}^{\omega}\big(\Pr(\mathcal{M}(X) \mid s_i, \theta) \,\big\|\, \Pr(\mathcal{M}(X) \mid s_j, \theta)\big) \\ &= \mathtt{D}_{\alpha}\big(\Pr(\mathcal{N}(X) \mid s_i, \theta) \,\big\|\, \Pr(\mathcal{N}(X) \mid s_j, \theta)\big). \end{aligned} \tag{78}$$

Now define, for each j, the sliced channel $\mathcal{N}_j := \mathtt{Slice} \circ \mathcal{M}_j$. Since $\mathcal{M}_j$ is $(\alpha, \varepsilon_j, \omega)$-Joint-SRPP, (78) implies

$$\mathtt{D}_{\alpha}\big(\Pr(\mathcal{N}_j(X) \mid s_i, \theta) \,\big\|\, \Pr(\mathcal{N}_j(X) \mid s_j, \theta)\big) \le \varepsilon_j.$$

Consider the *componentwise* slicing of the composed mechanism: sample $U_1, \ldots, U_J \overset{\text{i.i.d.}}{\sim} \omega$ independently of everything, and define

$$\mathtt{Slice}^J \circ \vec{\mathcal{M}}(X) := \big((U_1, \langle \mathcal{M}_1(X), U_1 \rangle), \ldots, (U_J, \langle \mathcal{M}_J(X), U_J \rangle)\big).$$

Equivalently, $\mathtt{Slice}^J \circ \vec{\mathcal{M}}(X) = (\mathcal{N}_1(X), \ldots, \mathcal{N}_J(X))$. By the assumed independence of the primitive randomness across j (and independence of U_j), the conditional laws under s_i and s_j factorize:

$$\Pr\big((\mathcal{N}_1(x), \ldots, \mathcal{N}_J(x)) \mid s_i, \theta\big) = \bigotimes_{j=1}^{J} \Pr\big(\mathcal{N}_j(x) \mid s_i, \theta\big),$$

and similarly under s_j. Therefore, by Rényi tensorization,

$$\begin{aligned} &\mathtt{D}_{\alpha}\Big(\Pr\big(\mathtt{Slice}^J \circ \vec{\mathcal{M}}(x) \mid s_i, \theta\big) \,\Big\|\, \Pr\big(\mathtt{Slice}^J \circ \vec{\mathcal{M}}(x) \mid s_j, \theta\big)\Big) \\ &= \sum_{j=1}^{J} \mathtt{D}_{\alpha}\Big(\Pr\big(\mathcal{N}_j(x) \mid s_i, \theta\big) \,\Big\|\, \Pr\big(\mathcal{N}_j(x) \mid s_j, \theta\big)\Big) \le \sum_{j=1}^{J} \varepsilon_j. \end{aligned}$$

Finally, by applying the representation (78) to the composed mechanism (with $\mathtt{Slice}^J$ as the slicing operator on $(\mathbb{R}^d)^J$), the left-hand side is exactly the Joint-SRD of $\vec{\mathcal{M}}$ under the same slice profile ω. Hence $\vec{\mathcal{M}}$ is $(\alpha, \sum_{j=1}^{J} \varepsilon_j, \omega)$-Joint-SRPP.

For sa-Joint-SRPP, the same argument applies after folding the subsampling randomness into each $\mathcal{M}_j$ (equivalently, applying the above proof to the corresponding expanded-output channels).

Combining Case 1 and Case 2, we conclude that for any $\Xi \in \{\text{Ave}, \text{Joint}, \text{sa-Ave}, \text{sa-Joint}\}$, if each $\mathcal{M}_j$ is $(\alpha, \varepsilon_j, \omega)$-$\Xi$-SRPP, then the composed mechanism $\vec{\mathcal{M}}$ is $(\alpha, \sum_{j=1}^{J} \varepsilon_j, \omega)$-$\Xi$-SRPP. □

W Experimental Details for Sec. 4.5

In this appendix, we provide the mathematical details for the numerical example in Sec. 4.5. The experiment compares full-dimensional RPP with finite-profile Ave-SRPP and Joint-SRPP under a fixed Pufferfish scenario.

Pufferfish scenario and query function. Consider the Pufferfish scenario $(\mathcal{S}, \mathcal{Q}, \Theta)$ with $\mathcal{S} = \{s_1, s_2, s_3\}$, $\mathcal{Q} = \{(s_i, s_j) : i \neq j\}$, and $\Theta = \{\theta^A, \theta^B\}$. The data domain contains tuples $x = (i, b, z)$, where $i \in \{1,2,3\}$, $b \in \{A, B\}$, and $z \in \mathcal{Z} := \{(0,0,0), (0.1,0.1,0.1)\}$. Define $f(x) := \mu_i^b + z \in \mathbb{R}^3$, where $\mu_1^A = e_1, \mu_2^A = e_2, \mu_3^A = e_3, \mu_1^B = e_1, \mu_2^B = \frac{1}{2}(e_1 + e_2)$, and $\mu_3^B = e_3$. Under prior θ^b, conditioning on secret s_i means that $X = (i, b, Z)$, where Z is uniform on $\mathcal{Z}$. Hence $P_{\theta^b}^{f, s_i}$ is the distribution of $\mu_i^b + Z$.

We apply the Gaussian mechanism $\mathcal{M}(X) = f(X) + N$, where $N \sim \mathcal{N}(0, \sigma^2 I_3)$, and set the Rényi order to $\alpha = 4$.

Full-dimensional Wasserstein sensitivity. Let $a := \mu_i^b - \mu_j^b$. We claim that

$$W_{\infty}(P_{\theta^b}^{f, s_i}, P_{\theta^b}^{f, s_j}) = \|a\|_2 = \|\mu_i^b - \mu_j^b\|_2.$$

For the upper bound, couple the two distributions by using the same draw of Z. That is, let $Y_i = \mu_i^b + Z$ and $Y_j = \mu_j^b + Z$. Then $Y_i - Y_j = a$ almost surely, so this coupling has essential supremum transport cost $\|a\|_2$. Hence

$$W_{\infty}(P_{\theta^b}^{f, s_i}, P_{\theta^b}^{f, s_j}) \le \|a\|_2.$$

For the reverse inequality, assume $a \neq 0$ and let $v := a / \|a\|_2$. Since all centers μ_i^b have coordinate sum one, the vector $a = \mu_i^b - \mu_j^b$ has coordinate sum zero. Hence v is orthogonal to the all-ones direction. Since Z varies only along the all-ones direction, we have $\langle v, Z \rangle = 0$ almost surely. Therefore, the projected distributions $v_{\#} P_{\theta^b}^{f, s_i}$ and $v_{\#} P_{\theta^b}^{f, s_j}$ are point masses:

$$v_{\#} P_{\theta^b}^{f, s_i} = \delta_{\langle v, \mu_i^b \rangle}, \quad v_{\#} P_{\theta^b}^{f, s_j} = \delta_{\langle v, \mu_j^b \rangle}.$$

Thus,

$$W_{\infty}(v_{\#} P_{\theta^b}^{f, s_i}, v_{\#} P_{\theta^b}^{f, s_j}) = |\langle v, \mu_i^b - \mu_j^b \rangle| = \|a\|_2.$$

Because the projection $y \mapsto \langle v, y\rangle$ is 1-Lipschitz, projecting cannot increase W_∞. Hence

$$W_\infty(P^{f,s_i}_{\theta^b}, P^{f,s_j}_{\theta^b}) \geq W_\infty(v_\# P^{f,s_i}_{\theta^b}, v_\# P^{f,s_j}_{\theta^b}) = \|a\|_2.$$

Combining the upper and lower bounds proves the claim. The case $a = 0$ is immediate.

Therefore, the full-dimensional RPP sensitivity is

$$\Delta^2_\infty = \max_{b\in\{A,B\},(s_i,s_j)\in\mathcal{Q}} \|\mu^b_i - \mu^b_j\|_2^2.$$

For θ^A, the three squared center distances are all 2. For θ^B, they are $\|\mu^B_1 - \mu^B_2\|_2^2 = \frac{1}{2}$, $\|\mu^B_1 - \mu^B_3\|_2^2 = 2$, and $\|\mu^B_2 - \mu^B_3\|_2^2 = \frac{3}{2}$. Thus, we have $\Delta^2_\infty = 2$.

For the Gaussian mechanism, the full-dimensional RPP Gaussian certificate is $\varepsilon_{\mathrm{RPP}}(\sigma^2) = \frac{\alpha\Delta^2_\infty}{2\sigma^2}$ [34]. Since $\alpha = 4$ and $\Delta^2_\infty = 2$, we have $\varepsilon_{\mathrm{RPP}}(\sigma^2) = \frac{4}{\sigma^2}$ and $\sigma^2_{\mathrm{RPP}}(\varepsilon) = \frac{4}{\varepsilon}$.

Per-slice Wasserstein sensitivities. For a direction $u \in \mathbb{S}^2$, let $\Psi^u(y) = \langle u, y\rangle$. The projected secret-conditioned clean query distribution is the distribution of $\langle u, \mu^b_i\rangle + \langle u, Z\rangle$. Since the same projected random component $\langle u, Z\rangle$ appears for every secret, the projected distributions are again translated copies. Hence

$$W_\infty\left(\Psi^u_\# P^{f,s_i}_{\theta^b}, \Psi^u_\# P^{f,s_j}_{\theta^b}\right) = |\langle u, \mu^b_i - \mu^b_j\rangle|.$$

Thus the per-slice Wasserstein sensitivity is

$$\Delta^u_\infty = \max_{b\in\{A,B\},(s_i,s_j)\in\mathcal{Q}} |\langle u, \mu^b_i - \mu^b_j\rangle|.$$

For a finite profile $\mathcal{U}_m = \{u_1, \ldots, u_m\} \subset \mathbb{S}^2$ with uniform weights, define

$$\overline{\Delta}^2(\mathcal{U}_m) = \frac{1}{m}\sum_{\ell=1}^m (\Delta^{u_\ell}_\infty)^2, \quad \Delta^2_\star(\mathcal{U}_m) = \max_{\ell\in[m]} (\Delta^{u_\ell}_\infty)^2.$$

These are the finite-profile sensitivity quantities used for Ave-SRPP and Joint-SRPP Gaussian SWM calibration, respectively.

Two forms of the Joint-SRPP certificate. For a fixed finite slice profile $\mathcal{U}_m = \{u_1, \ldots, u_m\}$ with uniform weights, the Joint-SRPP certificate can be evaluated in two related forms. The tight envelope form aggregates the per-slice Gaussian Rényi costs by the log-mean-exp expression

$$\varepsilon^{\mathrm{env}}_{\mathrm{Joint}}(\sigma^2; \mathcal{U}_m) = \frac{1}{\alpha - 1}\log\left[\frac{1}{m}\sum_{\ell=1}^m \exp\left((\alpha-1)\frac{\alpha(\Delta^{u_\ell}_\infty)^2}{2\sigma^2}\right)\right].$$

The closed-form calibration in Theorem 4.6 instead uses the worst-direction upper bound

$$\varepsilon^\star_{\mathrm{Joint}}(\sigma^2; \mathcal{U}_m) = \frac{\alpha\Delta^2_\star(\mathcal{U}_m)}{2\sigma^2}, \quad \Delta^2_\star(\mathcal{U}_m) = \max_{\ell\in[m]} (\Delta^{u_\ell}_\infty)^2.$$

These two quantities satisfy

$$\varepsilon^{\mathrm{env}}_{\mathrm{Joint}}(\sigma^2; \mathcal{U}_m) \leq \varepsilon^\star_{\mathrm{Joint}}(\sigma^2; \mathcal{U}_m),$$

because the log-mean-exp is upper bounded by the maximum per-slice cost. Both quantities certify Joint-SRPP for the same finite-profile mechanism, but the closed-form bound is more conservative and admits the direct inversion $\sigma^2 = \alpha\Delta^2_\star(\mathcal{U}_m)/(2\varepsilon)$. Accordingly, Experiment 1 uses the closed-form calibration to map a target ε to a noise variance, while Experiment 2 reports the tight envelope value assigned to a fixed Gaussian mechanism.

Bayes-optimal attack accuracy. In this construction, the shared random component Z is supported on the one-dimensional subspace spanned by $(1,1,1)$, whereas every center difference $\mu^b_i - \mu^b_j$ lies in the orthogonal contrast subspace. Indeed, all centers have coordinate sum one, so $\langle \mu^b_i - \mu^b_j, (1,1,1)\rangle = 0$ for all protected pairs. Thus, for distinguishing two secret-conditioned output distributions after adding isotropic Gaussian noise, the common Z component is a nuisance component shared by both hypotheses, and the likelihood ratio depends only on the contrast direction $\mu^b_i - \mu^b_j$. Consequently, for any protected pair with center distance $d = \|\mu^b_i - \mu^b_j\|_2$, the equal-prior Bayes-optimal binary attack accuracy is

$$\mathrm{Acc}^\star(d, \sigma^2) = \Phi\left(\frac{d}{2\sigma}\right),$$

where Φ is the standard normal CDF. The worst-protected pairs have $d = \Delta_\infty = \sqrt{2}$, so throughout the experiment we use $\mathrm{Acc}^\star(\sigma^2) = \Phi\left(\frac{\sqrt{2}}{2\sigma}\right)$.

Experiment 1: fixed certificate → protection strength. Experiment 1 fixes the same native numerical budget ε for RPP, Ave-SRPP, and Joint-SRPP, then compares the induced Bayes-optimal attack accuracy.

For RPP,

$$\sigma^2_{\mathrm{RPP}}(\varepsilon) = \frac{\alpha\Delta^2_\infty}{2\varepsilon}.$$

For Ave-SRPP under profile $\mathcal{U}_m$,

$$\sigma^2_{\mathrm{Ave}}(\varepsilon; \mathcal{U}_m) = \frac{\alpha\overline{\Delta}^2(\mathcal{U}_m)}{2\varepsilon}.$$

For Joint-SRPP under the Gaussian SWM calibration,

$$\sigma^2_{\mathrm{Joint}}(\varepsilon; \mathcal{U}_m) = \frac{\alpha\Delta^2_\star(\mathcal{U}_m)}{2\varepsilon}.$$

With $\alpha = 4$, these are

$$\sigma^2_{\mathrm{RPP}}(\varepsilon) = \frac{4}{\varepsilon}, \quad \sigma^2_{\mathrm{Ave}}(\varepsilon; \mathcal{U}_m) = \frac{2\overline{\Delta}^2(\mathcal{U}_m)}{\varepsilon},$$
$$\sigma^2_{\mathrm{Joint}}(\varepsilon; \mathcal{U}_m) = \frac{2\Delta^2_\star(\mathcal{U}_m)}{\varepsilon}.$$

The induced operational attack accuracy is then obtained by substituting each calibrated variance into $\mathrm{Acc}^{\star}(\sigma^2) = \Phi\left(\frac{\sqrt{2}}{2\sigma}\right)$.

Experiment 2: fixed protection strength → certificate gap. Experiment 2 fixes the Bayes-optimal attack accuracy a and compares the native certificate values assigned by RPP, Ave-SRPP, and Joint-SRPP to the same Gaussian mechanism. Let $t(a) := (\Phi^{-1}(a))^2$. Since $a = \Phi\left(\frac{\sqrt{2}}{2\sigma_a}\right)$, the common variance that realizes target attack accuracy a is $\sigma_a^2 = \frac{1}{2t(a)}$.

At this common variance, the RPP certificate is

$$\varepsilon_{\mathrm{RPP}}(a) = \frac{\alpha\Delta_\infty^2}{2\sigma_a^2} = 8t(a).$$

The Ave-SRPP certificate is

$$\varepsilon_{\mathrm{Ave}}(a; \mathcal{U}_m) = \frac{\alpha\overline{\Delta}^2(\mathcal{U}_m)}{2\sigma_a^2} = 4\overline{\Delta}^2(\mathcal{U}_m)t(a).$$

For Joint-SRPP, the finite-profile Gaussian SWM envelope certificate is

$$\varepsilon_{\mathrm{Joint}}(a; \mathcal{U}_m) = \frac{1}{\alpha-1}\log\left[\frac{1}{m}\sum_{\ell=1}^{m}\exp\left((\alpha-1)\frac{\alpha(\Delta_\infty^{u_\ell})^2}{2\sigma_a^2}\right)\right].$$

With $\alpha = 4$ and $\sigma_a^{-2} = 2t(a)$, this becomes

$$\varepsilon_{\mathrm{Joint}}(a; \mathcal{U}_m) = \frac{1}{3}\log\left[\frac{1}{m}\sum_{\ell=1}^{m}\exp\left(12(\Delta_\infty^{u_\ell})^2 t(a)\right)\right].$$

Ordering and interpretation. For each fixed finite profile $\mathcal{U}_m$ and fixed σ^2, there are three relevant certificate values:

$$\varepsilon_{\mathrm{Ave}}(\sigma^2; \mathcal{U}_m) = \frac{1}{m}\sum_{\ell=1}^{m} c_\ell,$$

$$\varepsilon_{\mathrm{Joint}}^{\mathrm{env}}(\sigma^2; \mathcal{U}_m) = \frac{1}{\alpha-1}\log\left(\frac{1}{m}\sum_{\ell=1}^{m} e^{(\alpha-1)c_\ell}\right),$$

and

$$\varepsilon_{\mathrm{Joint}}^{\star}(\sigma^2; \mathcal{U}_m) = \max_{\ell\in[m]} c_\ell, \quad c_\ell = \frac{\alpha(\Delta_\infty^{u_\ell})^2}{2\sigma^2}.$$

These quantities satisfy

$$\varepsilon_{\mathrm{Ave}}(\sigma^2; \mathcal{U}_m) \le \varepsilon_{\mathrm{Joint}}^{\mathrm{env}}(\sigma^2; \mathcal{U}_m) \le \varepsilon_{\mathrm{Joint}}^{\star}(\sigma^2; \mathcal{U}_m).$$

The first inequality follows from Jensen's inequality applied to the convex function $x \mapsto e^{(\alpha-1)x}$, and the second follows because the log-mean-exp is bounded above by the maximum per-slice cost. Thus, within a fixed finite profile, the tight Joint-SRPP envelope is at least the Ave-SRPP certificate, while the closed-form Joint-SRPP calibration is a conservative upper bound on the tight Joint envelope.

This ordering holds only within the same finite slice profile. Across different values of m, the profiles and uniform weights are different, so no monotone ordering is implied for Ave-SRPP, for the tight Joint-SRPP envelope, or for the closed-form Joint-SRPP bound. The profiles in the experiment are nested, but increasing m changes both the set of directions and the weights assigned to existing directions. Thus, the curves should be interpreted as certificates for different declared finite-profile SRPP notions under **Mode-I**, rather than as a monotone convergence sequence to RPP.

X Experiment Details

In this appendix, we provide additional details for the experiments in Section 6 and present further empirical results.

X.1 Experimental Setup: Gaussian Sliced Wasserstein Mechanisms for Static Privatization

This section provides additional details on our experimental methodology beyond what is presented in Section 6 of the main text.

Dataset configurations and prior distributions. As described in the main text, we evaluate our sliced SRPP mechanisms on three standard tabular benchmarks: Adult Census (30,162 records), Cleveland Heart Disease (303 records), and Student Performance (395 records). Each dataset is studied under two prior configurations to examine the impact of class imbalance on both privacy protection and utility.

For *Adult*, the imbalanced configuration uses the natural race distribution from the full dataset, where the majority class (White) accounts for approximately 86% of records, yielding a prior-only baseline accuracy of $\mathrm{Acc}_{\mathrm{maj}}^{0.86} = 0.86$ for attribute inference attacks. The balanced configuration constructs synthetic groups via deterministic hash-based assignment, partitioning records into five approximately equal groups with uniform prior $\pi(s) \approx 0.20$ for each synthetic secret value $s \in \{G_1, \ldots, G_5\}$, giving $\mathrm{Acc}_{\mathrm{maj}}^{0.20} = 0.20$.

For *Cleveland*, the imbalanced configuration uses the full dataset with its natural sex distribution (majority class probability ≈ 0.70), while the balanced configuration performs stratified subsampling to obtain equal numbers of male and female records, yielding a uniform prior with $\mathrm{Acc}_{\mathrm{maj}}^{0.50} = 0.50$.

For *Student Performance*, we similarly consider both the full dataset with its empirical sex distribution ($\mathrm{Acc}_{\mathrm{maj}}^{0.60} \approx 0.60$) and a balanced subsample with equal representation of each sex ($\mathrm{Acc}_{\mathrm{maj}}^{0.50} = 0.50$).

Detailed descriptions of each dataset configuration, including feature construction and sampling procedures, are provided in the following subsections.

Query construction. We consider two families of queries, both of which aggregate information at the group level rather than releasing individual records.

Summary statistics queries. For each secret value $s \in \mathcal{S}$, we compute a d-dimensional aggregate statistic over all records with $S = s$. For Adult ($d = 9$), the query includes: means and standard deviations of age, education level, and hours worked per week, together with the fraction of records with positive capital gain, positive capital loss, and income exceeding \$50K. For Cleveland ($d = 9$), we compute means and standard deviations of age, resting blood pressure, and cholesterol level, along with rates of high cholesterol (≥ 240 mg/dL), high blood pressure (≥ 140 mmHg), and presence of heart disease. For Student ($d = 9$), we use means and standard deviations of grades G1, G2, and G3, together with rates of high study time (≥ 3 hours/week), any past failures, and high absences (> 10 days).

Machine learning model queries. For the Adult dataset, we also treat trained model parameters as query outputs. For each race group $s \in \mathcal{S}$, we train a binary classifier to predict high income ($>$ \$50K) using a 9-dimensional feature vector constructed from demographic and employment attributes. The query output X consists of model parameters: for linear SVM and logistic regression, we extract the $d = 10$ coefficients (9 feature weights plus intercept); for random forest and gradient boosting, we extract the $d = 9$ feature importance scores. Each model is trained independently per secret group, so the query reflects group-specific predictive patterns.

Privacy mechanism and calibration. All experiments use isotropic Gaussian noise mechanisms of the form

$$Y = X + Z, \quad Z \sim \mathcal{N}(0, \sigma^2 I_d),$$

where the noise scale σ is calibrated via our sliced SRPP composition bounds.

For slicing, we generate $m = 200$ directions $\{u_\ell\}_{\ell=1}^m$ sampled uniformly from the unit sphere $\mathbb{S}^{d-1}$ with equal weights $\omega_\ell = 1/m$. Along each direction u_ℓ, we compute the one-dimensional Wasserstein-∞ sensitivity

$$\Delta_\infty^{u_\ell} = \max_{(s_i, s_j) \in Q} W_\infty\big(P^\theta_{u_\ell^\top X | S=s_i}, P^\theta_{u_\ell^\top X | S=s_j}\big),$$

where $P^\theta_{u_\ell^\top X | S=s}$ is the empirical distribution of the projected query values for secret s under world θ. These per-slice sensitivities are then aggregated according to the Ave-SRPP and Joint-SRPP definitions to determine the required noise level for target privacy parameters (α, ε).

We fix Rényi order $\alpha = 4$ throughout and sweep the privacy budget ε over the range $[0.05, 60]$ to study the privacy-utility tradeoff. Each configuration is evaluated over $N_{\text{trials}} = 3$ independent trials with different random seeds to account for Monte Carlo variability in the slicing directions and noise realization.

Utility and privacy evaluation. For utility, we report mean squared error (MSE), mean absolute error (MAE), average ℓ_2 distance per record, relative Frobenius error, and signal-to-noise ratio (SNR in dB). The primary utility metric is MSE:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n \|X_i - Y_i\|_2^2.$$

For empirical privacy auditing, we implement a prior-aware Gaussian maximum a posteriori (MAP) attack. Given a privatized output y_i, the attacker estimates the secret as

$$\hat{s}(y_i) = \arg\max_{s \in \mathcal{S}} \Big\{ \log \pi(s) + \log \phi\big(y_i; \mu_s, \sigma^2 I_d\big) \Big\},$$

where $\pi(s)$ is the prior distribution over secrets (either the empirical distribution from the dataset or the uniform prior for balanced configurations), $\mu_s = \mathbb{E}[X \mid S = s]$ is the mean query output for secret s computed from the unprivatized data, and $\phi(\cdot; \mu, \Sigma)$ denotes the Gaussian density with mean μ and covariance $\Sigma = \sigma^2 I_d$.

We report two metrics: (i) overall attack accuracy, and (ii) advantage over the prior-only baseline, defined as $\text{Adv} = \text{Acc} - \text{Acc}_{\text{maj}}$. For imbalanced Adult, the baseline is $\text{Acc}_{\text{maj}}^{0.86} = 0.86$; for balanced Adult, Cleveland, and Student (balanced), the baseline is $\text{Acc}_{\text{maj}}^{0.20} = 0.20$, $\text{Acc}_{\text{maj}}^{0.50} = 0.50$, and $\text{Acc}_{\text{maj}}^{0.50} = 0.50$, respectively; for imbalanced Cleveland and Student, the baselines are $\text{Acc}_{\text{maj}}^{0.70} = 0.70$ and $\text{Acc}_{\text{maj}}^{0.60} = 0.60$. An advantage close to zero indicates that the privacy mechanism successfully prevents the attacker from learning substantially more than what the prior distribution already reveals.

The following subsections provide detailed specifications of each dataset configuration.

X.1.1 Adult Census (race as secret)

We use the *Adult Census Income* dataset (30,162 records after preprocessing) in an attribute-privacy setting where the sensitive attribute is *race*. Let

$$S \in \mathcal{S} = \{\text{White}, \text{Black}, \text{Asian-Pac-Islander}, \text{Amer-Indian-Eskimo}, \text{Other}\},$$

with $k = |\mathcal{S}| = 5$, and let $X \in \mathbb{R}^d$ denote the per-record query vector constructed from demographic and socioeconomic features (see below for the specific query families). The Pufferfish scenario specifies the discriminatory set

$$Q = \{(s_i, s_j) : s_i, s_j \in \mathcal{S}, s_i \neq s_j\},$$

so that every ordered pair of distinct secret values is protected. We consider two configurations of the dataset that differ only in the prior over S.

Adult–0.86 prior (imbalanced). We use the full processed dataset with its empirical race distribution. Writing

$$\pi_{0.86}(s) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{S_i = s\}, \quad s \in \mathcal{S},$$

for the empirical prior over S, the corresponding *majority-class baseline* for any attribute-inference attack is

$$\mathrm{Acc}_{\mathrm{maj}}^{0.86} = \max_{s\in\mathcal{S}} \pi_{0.86}(s) \approx 0.86.$$

The associated Pufferfish world $\theta_D^{0.86}$ is given by the empirical joint distribution of (X,S) over all records in the full dataset.

Adult–0.20 prior (balanced). To study a uniform prior, we construct a balanced configuration by creating synthetic balanced groups via hash-based assignment. We partition the full dataset into $k = 5$ groups of approximately equal size using a deterministic hash function applied to each record's feature vector, yielding a subsample with an approximately uniform prior

$$\pi_{0.20}(s) = \frac{1}{n'}\sum_{i=1}^{n'} \mathbf{1}\{S_i = s\} \approx 0.20, \quad s \in \mathcal{S}.$$

In this configuration, the majority-class baseline is

$$\mathrm{Acc}_{\mathrm{maj}}^{0.20} = \max_{s\in\mathcal{S}} \pi_{0.20}(s) \approx 0.20.$$

The corresponding Pufferfish world $\theta_D^{0.20}$ is the empirical joint distribution of (X,S) restricted to this balanced subsample.

In both configurations, the Pufferfish family is instantiated by the empirical conditionals $\{P^{\theta}_{X|S=s}\}_{s\in\mathcal{S}}$ under the chosen world $\theta \in \{\theta_D^{0.86}, \theta_D^{0.20}\}$, together with the discriminatory set $\mathcal{Q}$ defined above. Our sliced SRPP mechanisms are calibrated to protect all pairs in $\mathcal{Q}$ for the specified Rényi order and privacy budget.

X.1.2 Cleveland Heart Disease (sex as secret)

We use the *processed Cleveland Heart Disease* dataset (303 records) in an attribute-privacy setting where the sensitive attribute is *sex*. Let

$$S \in \mathcal{S} = \{\text{female}, \text{male}\}, \qquad k = |\mathcal{S}| = 2,$$

and let $X \in \mathbb{R}^d$ denote the per-record query vector constructed from clinical features (see below for the specific query families). The Pufferfish scenario specifies the discriminatory set

$$\begin{aligned}\mathcal{Q} &= \{(s_i, s_j) : s_i, s_j \in \mathcal{S}, s_i \neq s_j\} \\ &= \{(\text{female}, \text{male}), (\text{male}, \text{female})\},\end{aligned}$$

so that every ordered pair of distinct secret values is protected. We consider two configurations of the dataset that differ only in the prior over S.

Cleveland–0.7 prior. We use the full processed dataset with its empirical sex distribution. Writing

$$\pi_{0.7}(s) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{S_i = s\}, \quad s \in \mathcal{S},$$

for the empirical prior over S, the corresponding *majority-class baseline* for any attribute-inference attack is

$$\mathrm{Acc}_{\mathrm{maj}}^{0.7} = \max_{s\in\mathcal{S}} \pi_{0.7}(s) \approx 0.7.$$

The associated Pufferfish world $\theta_D^{0.7}$ is given by the empirical joint distribution of (X,S) over all records in the full dataset.

Cleveland–0.5 prior. To study a uniform prior, we construct a balanced configuration by stratified subsampling on sex. Let n_{f} and n_{m} denote the numbers of female and male records in the full dataset. We draw $n' = \min(n_{\mathrm{f}}, n_{\mathrm{m}})$ records uniformly at random without replacement from each sex subset, yielding a subsample of size $2n'$ with an approximately uniform prior

$$\pi_{0.5}(s) = \frac{1}{2n'}\sum_{i=1}^{2n'} \mathbf{1}\{S_i = s\} \approx 0.5, \quad s \in \mathcal{S}.$$

In this configuration, the majority-class baseline is

$$\mathrm{Acc}_{\mathrm{maj}}^{0.5} = \max_{s\in\mathcal{S}} \pi_{0.5}(s) \approx 0.5.$$

The corresponding Pufferfish world $\theta_D^{0.5}$ is the empirical joint distribution of (X,S) restricted to this balanced subsample.

In both configurations, the Pufferfish family is instantiated by the empirical conditionals $\{P^{\theta}_{X|S=s}\}_{s\in\mathcal{S}}$ under the chosen world $\theta \in \{\theta_D^{0.7}, \theta_D^{0.5}\}$, together with the discriminatory set $\mathcal{Q}$ defined above. Our sliced SRPP mechanisms are calibrated to protect all pairs in $\mathcal{Q}$ for the specified Rényi order and privacy budget.

X.1.3 Student Performance (grade group as secret)

We use the *Student Performance* dataset [36] (395 records) in an attribute-privacy setting where the sensitive attribute is a discretized *grade group* derived from the final grade $G3 \in \{0, \ldots, 20\}$. We partition students into $k = 5$ grade groups using quantile-based binning:

$$S \in \mathcal{S} = \{\text{G1}, \text{G2}, \text{G3}, \text{G4}, \text{G5}\}, \qquad k = |\mathcal{S}| = 5,$$

where each group corresponds to a quintile of the final grade distribution (G1 = lowest quintile, G5 = highest quintile). Let $X \in \mathbb{R}^d$ denote the per-record query vector constructed from grade and study-related features. The Pufferfish scenario specifies the discriminatory set

$$\mathcal{Q} = \{(s_i, s_j) : s_i, s_j \in \mathcal{S}, s_i \neq s_j\},$$

so that every ordered pair of distinct grade groups is protected. We consider two configurations of the dataset that differ in the prior over S.

Student (full dataset). We use all 395 records with grade groups assigned via 5-quantile binning. Since quantile-based binning attempts to create approximately equal-sized groups by construction, the empirical prior is

$$\pi(s) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{S_i = s\}, \quad s \in \mathcal{S},$$

with majority-class baseline

$$\mathrm{Acc}_{\mathrm{maj}} = \max_{s\in\mathcal{S}} \pi(s).$$

Due to tied final grade values preventing perfect equal division, the empirical distribution may deviate slightly from uniform. The Pufferfish world θ_D^{full} is the empirical joint distribution of (X,S) over the full dataset.

Student (balanced subsample). To ensure an exactly uniform prior, we perform stratified subsampling: let n_s denote the number of records in grade group s in the full dataset, and set $n' = \min_{s\in\mathcal{S}} n_s$. We draw n' records uniformly at random without replacement from each grade group, yielding a subsample of size $5n'$ with exactly uniform prior

$$\pi^{\mathrm{bal}}(s) = \frac{1}{5n'}\sum_{i=1}^{5n'} \mathbf{1}\{S_i = s\} = 0.2, \quad \forall s \in \mathcal{S}.$$

The corresponding majority-class baseline is

$$\mathrm{Acc}_{\mathrm{maj}}^{0.2} = \max_{s\in\mathcal{S}} \pi^{\mathrm{bal}}(s) = 0.2.$$

The Pufferfish world θ_D^{bal} is the empirical joint distribution of (X,S) restricted to this balanced subsample.

In both configurations, the Pufferfish family is instantiated by the empirical conditionals $\{P^{\theta}_{X|S=s}\}_{s\in\mathcal{S}}$ under the chosen world $\theta \in \{\theta_D^{\mathrm{full}}, \theta_D^{\mathrm{bal}}\}$, together with the discriminatory set Q defined above. Our sliced SRPP mechanisms are calibrated to protect all pairs in Q for the specified Rényi order and privacy budget.

X.1.4 Additional Results.

Figure 6 presents analogous privacy-utility tradeoff curves for the Cleveland Heart Disease and Student Performance datasets, evaluating the same four query types (logistic regression, random forest, SVM, and summary statistics) under both balanced and imbalanced prior configurations.

The results exhibit qualitatively similar behavior to the Adult dataset experiments shown in Section 6. Across all configurations, Ave-SRPP and Joint-SRPP demonstrate the expected monotonic tradeoff: as the privacy budget ε increases, MSE decreases while attack accuracy rises from the prior baseline toward perfect inference. At small ε, both mechanisms successfully limit attack accuracy to near the prior baselines (Cleveland: 0.7 and 0.5; Student: 0.6 and 0.2), confirming effective privacy protection. At any fixed ε, Joint-SRPP remains more conservative than Ave-SRPP, inducing higher MSE but lower attack accuracy, with the gap most pronounced in the moderate privacy regime. These additional experiments validate that the comparative performance of Ave-SRPP and Joint-SRPP generalizes across datasets of varying size (Cleveland: 303 records, Student: 395 records versus Adult: 30,162 records), different numbers of secret classes ($k = 2$ for Cleveland and Student-sex versus $k = 5$ for Adult and Student-grade).

X.2 SRPP- sa-SRPP-SGD with Gradient Clipping

For our CIFAR-10 experiments (Figures 3i-3l) in Section 6, we use the ResNet22, adapted for compatibility with Opacus [45] by following the architecture of [42],

The network is a 22-layer residual network with three stages of channel sizes $(16,32,64)$ and $(3,4,3)$ residual blocks per stage. The first layer is a 3×3 convolution (16 channels, stride 1, padding 1) followed by group normalization and ELU. Stages 2–3 downsample via stride 2 in their first blocks; shortcuts use spatial subsampling with channel zero-padding.

Each residual block has the form

$$\begin{aligned} x &\mapsto \mathrm{GN}_1(\mathrm{ELU}(\mathrm{Conv}_1(x))) \\ &\mapsto \mathrm{GN}_2(\mathrm{Conv}_2(\cdot)) + \mathrm{shortcut}(x) \\ &\mapsto \mathrm{GN}_3(\mathrm{ELU}(\cdot)), \end{aligned}$$

where both convolutions are 3×3 with padding 1; Conv_1 uses stride 2 for downsampling blocks and stride 1 otherwise, while Conv_2 always uses stride 1. Group normalization layers $\mathrm{GN}_1, \mathrm{GN}_2, \mathrm{GN}_3$ use at most four groups without affine parameters.

After the final stage, we apply 3×3 adaptive average pooling and flatten to $z \in \mathbb{R}^{576}$. Before classification, we apply per-example standardization $\tilde{z} = (z-\mu(z))/(\sigma(z)+10^{-6})$ where $\mu(z)$ and $\sigma(z)$ are the sample mean and standard deviation, then compute logits via $\ell = W\tilde{z}+b$ with $W \in \mathbb{R}^{10\times 576}$ and $b \in \mathbb{R}^{10}$. Weights use Kaiming-normal initialization. The model contains approximately 270,000 parameters. This architecture is fully compatible with per-example gradient computation in Opacus; in particular, we replace non-picklable lambda-based shortcut layers with explicit `PaddingShortcut` modules.

Standard SGD With this ResNet22 model, we consider a *global Lipschitz constant* as a special case of Assumption 5.2: the update map $T_t(z; y_{<t}) = \xi_{t-1} - \eta_t(\mu v_{t-1} + z)$ for SGD with momentum satisfies

$$\left|u_i^{\top}\left(T_t(z;y_{<t}) - T_t(z';y_{<t})\right)\right| = \eta_t \left|u_i^{\top}(z-z')\right|$$

for all slices $u_i \in \mathcal{U}$ and all histories $y_{<t}$, yielding the *uniform slice-wise Lipschitz constant* $L_{t,i} = \eta_t$ for all $i \in [m]$. This simplifies the HUC from Proposition 5.1 to

$$h_{t,i} = \left(\frac{2K_t\eta_t C}{B_t}\right)^2, \quad \forall i \in [m],$$

which depends only on the learning rate η_t, clipping threshold C, batch size B_t, and discrepancy cap K_t, but is *independent of the slice direction*. As a result, Ave-SRPP and Joint-SRPP require the same noise calibration in this uniform case, though their privacy quantification schemes remain different.

X.3 Experiment Setup

Now, we provide detailed specifications for our SRPP-SGD and sa-SRPP-SGD experiments (Figures 3i-3l) in Section 6 of the main text.

Pufferfish two-world construction. We construct a Pufferfish privacy scenario on CIFAR-10 by creating two aligned worlds that differ minimally in the prevalence of a designated secret class. Specifically, we select `cat` as the secret class and define two target prevalences: $p_{\text{low}} = 0.10000$ and $p_{\text{high}} = 0.10004$.

The world construction proceeds in two steps to ensure minimal edit distance:

1. *Base labels → world s_0*: Starting from CIFAR-10's original training labels (50,000 examples), we minimally relabel examples to achieve exactly $n_0 = \lfloor p_{\text{low}} \cdot N \rfloor$ `cat` instances. If the original dataset contains fewer than n_0 cats, we randomly select non-cat examples and relabel them as cats; if it contains more, we randomly select cat examples and relabel them to other (non-cat) classes.
2. *World s_0 → world s_1*: From s_0, we perform minimal edits to reach exactly $n_1 = \lfloor p_{\text{high}} \cdot N \rfloor$ cat instances, producing world s_1. The edit operations are analogous to step 1.

This construction yields two label vectors y_0 and y_1 (both of length $N = 50{,}000$) that differ on exactly $\Delta = |n_1 - n_0| = 20$ examples, while the image data remains identical across both worlds. The mismatch rate between worlds is $p_{\text{realized}} = \Delta/N = 0.0004$.

For training, we fix a single *realized world*—either s_0 or s_1—and overwrite the CIFAR-10 training labels with the corresponding realized label vector. All subsequent training (private and non-private) operates exclusively on this realized world, ensuring that the privacy guarantee protects against an adversary who might observe the other world.

Privacy accounting: SRPP-SGD, sa-SRPP-SGD, and group-DP-SGD. We compare three privacy accounting methods.

(1) SRPP-SGD (deterministic worst-case cap): For each iteration t, we bound the number of differing examples in a batch by the deterministic cap $K_{\text{cap}} = \min(L, \Delta) = 20$. Under isotropic Gaussian noise with single-block clipping, the per-step HUC is

$$h_t^* = \left(\frac{2\eta_t C K_{\text{cap}}}{L}\right)^2 = \left(\frac{2\eta_t \cdot 4.0 \cdot 20}{512}\right)^2.$$

The total HUC over all T steps is $H_{\text{total}} = \sum_{t=1}^{T} h_t^*$, and the noise scale is calibrated via:

$$\sigma^2(\varepsilon) = \frac{\alpha}{2\varepsilon} \cdot H_{\text{total}},$$

where $\alpha = 16$ is the Rényi order. The Opacus noise multiplier is $\sigma(\varepsilon) \cdot L/C$.

(2) sa-SRPP-SGD (expected $\overline{K}^2$ via hypergeometric model): We model the number of differing examples K_t in each batch as a random variable following a hypergeometric distribution: $K_t \sim \text{Hypergeometric}(N, \Delta, L)$ under fixed-size sampling without replacement. For this distribution,

$$\mathbb{E}[K_t] = L \cdot \frac{\Delta}{N}, \quad \text{Var}[K_t] = L \cdot \frac{\Delta}{N} \cdot \left(1 - \frac{\Delta}{N}\right) \cdot \frac{N-L}{N-1},$$

and thus $\mathbb{E}[K_t^2] = \text{Var}[K_t] + \mathbb{E}[K_t]^2$. We define

$$\gamma^2 = a^2 \cdot \frac{\mathbb{E}[K_t^2]}{L^2},$$

where $a = 1$ for add-remove adjacency (our default) or $a = 2$ for replace adjacency. The total HUC is then

$$H_{\text{total}} = 4C^2\gamma^2 \sum_{t=1}^{T} \eta_t^2,$$

and the noise scale is calibrated identically to SRPP-SGD via $\sigma^2(\varepsilon) = (\alpha/2\varepsilon) \cdot H_{\text{total}}$.

(3) group-DP-SGD (baseline): As a baseline, we also implement group-DP-SGD using the standard DP-SGD accountant [1] with group size $G = \lceil \mathbb{E}_\eta[K_t^2]^{1/2} \rceil$ (derived from the hypergeometric model). The DP-SGD accountant computes a per-step privacy loss, which we then convert to group DP by scaling with the group size. App. G provides detailed discussions of the relationship between SRPP/sa-SRPP-SGD and group DP-SGD.

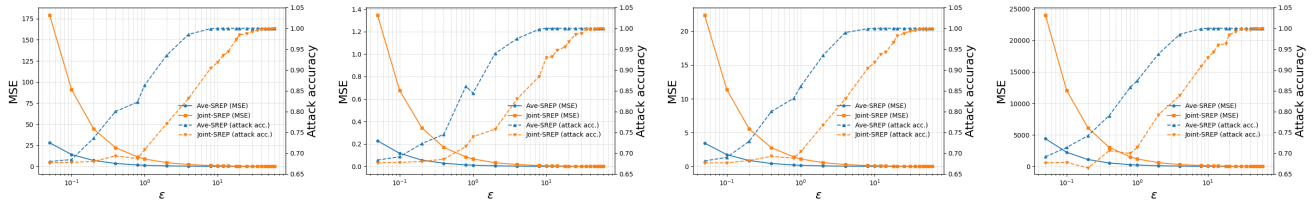
In the experiments for Figures 3i and 3k, we sweep the privacy budget ε over the range { 80, 70, 65, 60, 55, 50, 45, 40, 35, 30, 25, 20, 18, 14, 12, 10, 8, 4, 2, 1, 0.5 } and report test accuracy at each ε. We evaluate two sampling rates corresponding to batch sizes $L = 512$ (sampling rate $q \approx 0.01$) and $L = 1024$ (sampling rate $q \approx 0.02$) on the full 50,000-example training set. Training uses 40 epochs for $L = 512$ and 40 epochs for $L = 1024$, with clipping norm $C = 4.0$ (for $L = 512$) or $C = 5.0$ (for $L = 1024$), learning rate $\eta_0 = 0.2$ (for $L = 512$) or $\eta_0 \approx 0.14$ (for $L = 1024$) with cosine decay and 5% warmup, momentum 0.9, and weight decay 5×10^{-4}. All privacy accounting uses Rényi order $\alpha = 16$ and add-remove adjacency.

Evaluation metrics. For each privacy budget ε, we train the model to convergence (50 epochs) and report:

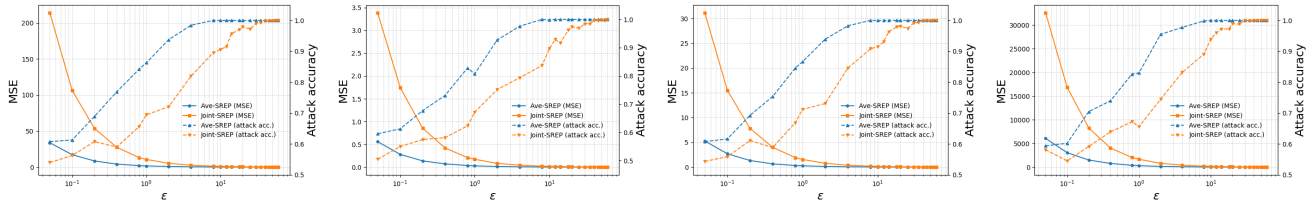
- *Test accuracy*: Classification accuracy on the CIFAR-10 test set (10,000 examples).
- *Training accuracy*: Classification accuracy on the training set (for diagnostic purposes).
- *Reference* ε: The privacy budget computed by Opacus's built-in accountant (at $\delta = 10^{-5}$) for comparison.

Overfitted regime and membership inference attacks. To perform empirical privacy auditing, we conduct additional experiments in an overfitted regime where membership attacks are most effective. We randomly subsample 1,000 examples from the CIFAR-10 training set and train models for 200 epochs with reduced regularization (weight decay set to zero) to induce overfitting. We use batch size $L = 256$, clipping norm $C = 5.0$, learning rate $\eta_0 = 0.1$ with cosine decay, and evaluate privacy budgets $\varepsilon \in \{2, 8, 20, 50, 70, 100, 200, 300, 500\}$ plus a non-private baseline.

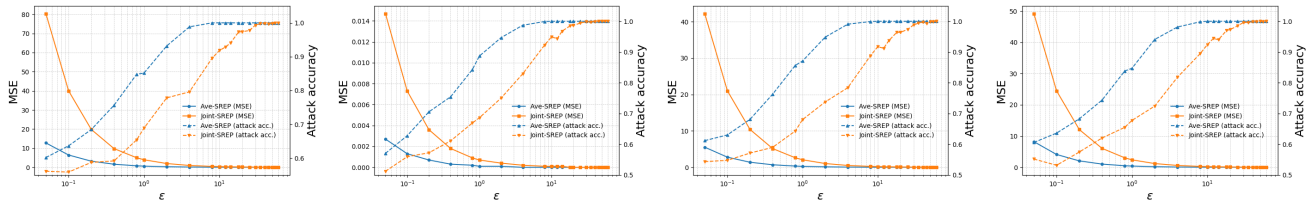
On these overfitted models, we apply the loss-threshold membership inference attack of [44]. For each example x with true label y, we compute the cross-entropy loss $\ell(x) = -\log p_\theta(y \mid x)$ and use the negative loss as a membership score: lower loss suggests the example is more likely a training member. We compute the ROC AUC for distinguishing the 1,000 training members from an equal-sized set of non-members sampled from the remaining CIFAR-10 training data. ROC AUC of 0.5 indicates random guessing (perfect privacy), while higher values indicate successful membership inference. We report ROC AUC as a function of ε to demonstrate how privacy protection degrades membership leakage.



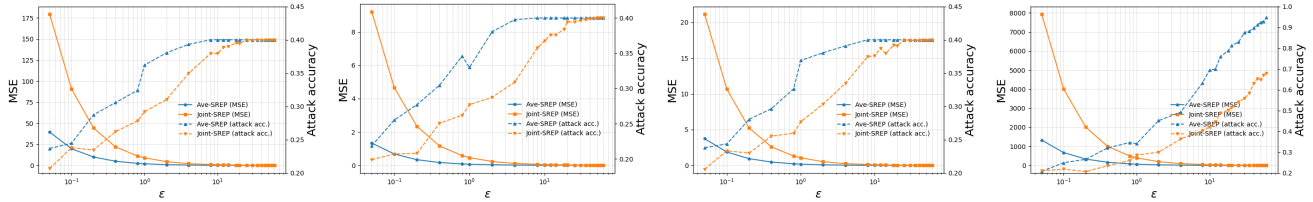
(a) Logistic Regression (b) Random Forest (c) SVM (d) Summary statistics



(e) Logistic Regression (f) Random Forest (g) SVM (h) Summary statistics



(i) Logistic Regression (j) Random Forest (k) SVM (l) Summary Statistics



(m) Logistic Regression (n) Random Forest (o) SVM (p) Summary Statistics

Figure 6: **Privacy-utility tradeoffs on Cleveland Heart Disease and Student Performance datasets.** (a)–(d): Cleveland dataset with imbalanced prior (majority-class baseline $\mathrm{Acc}_{\mathrm{maj}}^{0.7} = 0.7$). (e)–(h): Cleveland dataset with balanced prior ($\mathrm{Acc}_{\mathrm{maj}}^{0.5} = 0.5$). (i)–(l): Student Performance dataset with imbalanced prior (sex as secret, $\mathrm{Acc}_{\mathrm{maj}}^{0.6} = 0.6$). (m)–(p): Student Performance dataset with balanced prior (grade group as secret, $\mathrm{Acc}_{\mathrm{maj}}^{0.2} = 0.2$). Each configuration evaluates Ave-SRPP and Joint-SRPP mechanisms across privacy budgets $\varepsilon \in \{0.05, 0.1, 0.2, 0.4, 0.8, 1.0, 2.0, 4.0, 8.0, 10.0, 12.0, 14.0, 18.0, 20.0, 25.0, 30.0, 35.0, 40.0, 45.0, 50.0, 55.0, 60.0\}$ with Rényi order $\alpha = 4$. Metrics shown: mean squared error (MSE), attack accuracy, and attack advantage over the priors.