# A CERTIFIED ROBUSTNESS FRAMEWORK FOR BACKDOOR POISONING ATTACKS USING $f$-DIVERGENCES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Machine learning systems trained on untrusted data are vulnerable to data and backdoor poisoning attacks, in which an adversary corrupts training and test samples to manipulate predictions at inference time. Randomized smoothing yields provable defenses against such attacks, but existing applications to train-time robustness are limited to specific threat models and smoothing mechanisms. We propose a general framework that allows expressing complex adversary constraints on the number of corrupted training examples, feature and label perturbation magnitudes, and test-time corruptions. Our framework randomizes inputs by combining data subsampling with feature, label and test-sample noising. We propose a composable certification procedure based on $f$-divergences that handles arbitrary combinations of threat models and noising strategies, and recovers several prior certification regimes exactly as special cases. We demonstrate our framework experimentally by certifying, for the first time, general $\ell_p$-constraint poisoning adversaries under bagging with additive Laplace, Gaussian, and Uniform noises.

## 1 INTRODUCTION

Despite their continued adoption in wide-reaching domains, machine learning systems remain susceptible to adversarial attacks which subvert their reliability and can cause harmful outcomes. These attacks span a broad class of threat models, including both training- and inference-phase attacks. Of particular concern are *data* and *backdoor poisoning* attacks, in which an adversary corrupts a portion of the training data to degrade or control model behavior (Gu et al., 2019; Shafahi et al., 2018; Saha et al., 2020; Aghakhani et al., 2021; Geiping et al., 2021; Souri et al., 2022). To defend against such attacks, certified defenses have been proposed that provide provable guarantees on the consistency of a classifier's output around an input. The most popular of these techniques, randomized smoothing, works by lifting the classifier into a function over a noised version of the input and defining a new classifier by taking the expectation (Lecuyer et al., 2019; Cohen et al., 2019).

The intuitive reason randomized smoothing works is that averaging a function over a noised input induces a uniform continuity property whose stability depends primarily on the consistency of the function around the input, rather than the continuity properties of the base function. This principle applies whether the input and noise are simple (e.g., classification under Gaussian noise) or complex (e.g., an entire randomized dataset). In the case of data poisoning or backdoor attacks, several natural noising strategies exist, including data subsampling, feature noising, and label flipping. However, as the noising procedures become more complex, analyzing the certified guarantees becomes increasingly difficult. As a result, existing certified defenses for data poisoning attacks remain limited to smoothing strategies like simple bagging or discrete label flipping (Jia et al., 2021; Chen et al., 2022; Zhang et al., 2022).

In this work, we ask, *can we develop a highly general randomized smoothing framework for data poisoning and backdoor adversaries that easily handles complex threat models and smoothing strategies?* Our insight is to leverage the convex-optimization view of Randomized Smoothing first introduced by Dvijotham et al. (2020), in which the admissible perturbation set is replaced by a convex outer approximation defined by $f$-divergence constraints. The certification condition relies on the smoothing mechanism and threat model only through a set of divergence constraints corresponding to a certain dual problem with nice composition properties. We argue this is the "right" perspective for composite mechanisms such as those arising in the backdoor poisoning setting.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

**Contributions** We present a randomized smoothing framework for backdoor and data poisoning attacks incorporating rich attacker threat models, a multi-step randomization scheme, and efficient certification via convex relaxation using $f$-divergences. Our framework makes it straightforward to describe complex threat models spanning feature, label, and test instance corruptions, describe noising strategies that defend against them, and to derive sound robustness certificates under the resulting schemes. We explain how our framework recovers the tight certificates of several prior works as special cases. We use our framework to study certifiably robust classification under backdoor and data poisoning attacks in a wide range of settings not considered in prior works.

## 2 Background and related work

A growing body of work studies how to obtain certifiable robustness guarantees against data poisoning and backdoor attacks. Appendix C gives a more detailed treatment of selected works.

**Randomized smoothing** In response to the weaknesses of empirical defenses (Athalye et al., 2018), randomized smoothing emerged as a technique to obtain provable robustness to adversarial examples (Lecuyer et al., 2019; Cohen et al., 2019; Li et al., 2019; Salman et al., 2019). Early work established the standard certificate analysis methodology via the Neyman–Pearson Lemma (Neyman & Pearson, 1933) and the use of confidence-corrected estimates of the smoothed prediction. Other characterizations have also been explored, including functional optimization over $f$-divergence constraints (Dvijotham et al., 2020) and as a nonlinear Lipschitz condition (Salman et al., 2019).

**Learner-specific defenses** Some works derive robustness certificates for specific learning algorithms. Early work studied duality-based upper bounds on the total test error (Steinhardt et al., 2017). Rosenfeld et al. (2020) derive an efficient certified inference algorithm under label flipping attacks for classifiers trained using linear regression and label flipping noise. Drews et al. (2020) train certifiably robust decision trees with abstract interpretation. Jia et al. (2022) derive efficient certified guarantees for nearest neighbors classifiers under label flipping attacks. Sabanayagam et al. (2025) and Gosch et al. (2026) certify graph neural networks and SVMs against data poisoning, backdoor, and label flipping attacks using the neural tangent kernel and linear programming. Sosnin et al. (2025) extend abstract interpretation to gradient-based training to certify gradient-based learners.

**Differential privacy for robustness** Another line of work obtains robustness from differentially private (DP) learners. Ma et al. (2019) certify global accuracy using the stability of expectation under a differentially private algorithm, and Xie et al. (2023) study a similar property in federated learning. Liu et al. (2023) extend the results of Ma et al. (2019) to instance-wise poisoning by improving the group privacy analysis of the subsampled Gaussian mechanism. Saxena et al. (2026) connect the standard hypothesis testing framework for certification to the theory of DP privacy profiles, enabling efficient analysis of composed mechanisms like DP-SGD.

**Voting-based poisoning defenses** The most popular strategy for provable robustness to poisoning is to train an ensemble of learners over different partitions or randomized samples and to use robustness properties of the majority vote. One line of work considers deterministic partitioning strategies (Levine & Feizi, 2021; Wang et al., 2022; Rezaei et al., 2023; Hammoudeh & Lowd, 2024). Another line of work considers Randomized Smoothing by training ensembles on randomized samples of training data (Weber et al., 2023; Jia et al., 2021; Chen et al., 2022; Zhang et al., 2022).

## 3 Problem statement and preliminaries

### 3.1 Learning setup

We consider classification tasks with input space $\mathcal{X} \subseteq \mathbb{R}^d$ and label space $\mathcal{Y} = [C]$, and write $\mathcal{Z} \coloneqq \mathcal{X} \times \mathcal{Y}$ for the labeled input space and $\mathcal{S} \coloneqq \mathcal{Z}^*$ for the space of training sets. We absorb training and inference into a single procedure: define the joint input space $\mathcal{W} \coloneqq \mathcal{S} \times \mathcal{X}$ and

suppose $\mathcal{W} \subseteq \widetilde{W}$ for some ambient space $\widetilde{\mathcal{W}}$. [1] A *base predictor* $G\colon \widetilde{\mathcal{W}} \to \mathcal{P}(\mathcal{Y})$ maps a joint input $w = (S, x_{\mathrm{te}})$ to a distribution over labels. Concretely, $G$ composes a learning algorithm $\mathcal{A}$ with evaluation at a test point, $G(w) = G(S, x_{\mathrm{te}}) = \mathcal{A}(S)(x_{\mathrm{te}})$. The induced hard predictor is $g(w) \coloneqq \arg\max_{y \in \mathcal{Y}} G(w)_y$, with fixed tie-breaking rule. $\mathcal{A}$ can be randomized, in which case $G$ denotes the expectation over $\mathcal{A}$'s internal randomness (and thus $G$ is a deterministic function of $w$).

### 3.2 Threat model

We consider a replacement adversary who jointly corrupts the training set and test sample. For joint input $w = (S, x_{\mathrm{te}})$ with $S = ((x_i, y_i))_{i=1}^n$, denote the corrupted input by $w' = (S', x'_{\mathrm{te}})$, where $S' = ((x'_i, y'_i))_{i=1}^n$. In particular, the adversary may corrupt training features $x_i \mapsto x'_i$, training labels $y_i \mapsto y'_i$, and the test sample $x_{\mathrm{te}} \mapsto x'_{\mathrm{te}}$, but may not insert or delete training examples.

We model corruption constraints along four axes. The first axis constrains the number of corrupted examples in the training set. The remaining axes constrain per-example feature or label corruptions in terms of a feature-space distance function $d_{\mathcal{X}}$ or label space distance function $d_{\mathcal{Y}}$. The adversary's capability is thus parameterized by the pair $(d_{\mathcal{X}}, d_{\mathcal{Y}})$ together with a *threat vector* $\rho = (m_{\mathrm{p}}, \varepsilon_{\mathrm{f}}, \varepsilon_{\mathrm{l}}, \varepsilon_{\mathrm{te}})$. Define $I(w, w') \coloneqq \{i \in [n] : (x_i, y_i) \neq (x'_i, y'_i)\}$. The threat vector fixes the constraints (1) $|I| \leq m_{\mathrm{p}}$, (2) $\max_{i \in I} d_{\mathcal{X}}(x_i, x'_i) \leq \varepsilon_{\mathrm{f}}$, (3) $\max_{i \in I} d_{\mathcal{Y}}(y_i, y'_i) \leq \varepsilon_{\mathrm{l}}$, (4) $d_{\mathcal{X}}(x_{\mathrm{te}}, x'_{\mathrm{te}}) \leq \varepsilon_{\mathrm{te}}$. These constraints determine the *perturbation set* $\mathcal{B}_\rho(w) \coloneqq \{w' \in \mathcal{W} : \text{constraints (1)–(4) hold}\}$.

Decoupling the budget into four independent axes allows us to recover the threat models of several prior works as special cases (Rosenfeld et al., 2020; Jia et al., 2021; Zhang et al., 2022), as well as generalize to several previously unconsidered threat models (e.g., $m_{\mathrm{p}} < \infty$ with $d_{\mathcal{X}} = \|\cdot\|_p$ for $p \neq 0$). Usually we will take $d_{\mathcal{Y}}$ discrete and $d_{\mathcal{X}} \equiv \|\cdot\|_p$ for some $0 \leq p \leq \infty$.

### 3.3 Certified robustness

We seek to provably defend against attacks within our threat model under the following definition.

**Definition 3.1** ($\rho$-Robustness)**.** *Let $g\colon \mathcal{W} \to \mathcal{Y}$ be a deterministic hard predictor. We say that $g$ is* $\rho$-robust *at $w \in \mathcal{W}$ if $g(w) = g(w')$ for all $w' \in \mathcal{B}_\rho(w)$.*

Existing certified poisoning defenses typically report robustness as a scalar *radius* specifying the tolerable corruption magnitude along a single axis (e.g., $m_{\mathrm{p}}$ or $\varepsilon_{\mathrm{f}}$) (Rosenfeld et al., 2020; Levine & Feizi, 2021; Zhang et al., 2022). Instead, we are interested in characterizing the full certified region directly. For fixed $g$, define the *robustness region* $\mathcal{R}(w) \coloneqq \{\rho : g \text{ is } \rho\text{-robust at } w\}$. Defining the robustness region $\mathcal{R}$ encodes the idea that varying one threat parameter (e.g., $\varepsilon_{\mathrm{te}}$) can affect robustness against other parameters (e.g., $\varepsilon_{\mathrm{f}}$), and we study the entire object holistically.

## 4 Methodology

Our goal is to develop a general defense for training classifiers that handles the threat models from Section 3 and which admits an efficient certification procedure proving robustness.

### 4.1 Robust learning procedure

**Robust training via randomized smoothing** We design a certifiably robust training procedure following the methods of *Randomized Smoothing* (Lecuyer et al., 2019; Li et al., 2019; Cohen et al., 2019), initially introduced for certified robustness against evasion attacks. The main idea behind randomized smoothing is to transform a pre-existing *base learner* (in our case, the joint training-inference procedure $G$) into a provably robust *smoothed* classifier $\overline{G_{\mathcal{M}}}$. The smoothed classifier $\overline{G_{\mathcal{M}}}$ is obtained by applying a randomized *smoothing mechanism* $\mathcal{M}\colon \mathcal{W} \to \mathcal{P}(\widetilde{\mathcal{W}})$ to the inputs and lifting the base classifier $G$ to a classifier on the space of *distributions* over the joint input space according to the aggregation rule: $\overline{G_{\mathcal{M}}}(w) \coloneqq \mathbb{E}_{\widetilde{w} \sim \mathcal{M}(w)}[G(\widetilde{w})]$. $\overline{G_{\mathcal{M}}}$ is a deterministic soft classifier whose associated hard classifier is $\overline{g_{\mathcal{M}}}(w) = \arg\max_y \overline{G_{\mathcal{M}}}(w)_y$.

[1] We introduce $\widetilde{\mathcal{W}}$ because the randomization we introduce later means the base learner must process datasets from a slightly different space than $\mathcal{W}$.

**Randomized mechanism** The smoothing mechanism $\mathcal{M}$ should be chosen in a way to "wipe away" any influence the corruptions might have on the prediction. We design a multi-step randomization procedure that consists of a data subsampling phase followed by a feature and label noising phase. The resulting probability measure takes the form of a mixture distribution whose components correspond to the samples retained during the subsampling phase. Each component of the resulting mixture follows structurally identical randomization behavior, but instantiated with different parameters.

For a training set $S = ((x_i, y_i))_{i=1}^n$ and sample-index sequence $U = (u_1, \ldots, u_k) \in [n]^k$, define $S_U \coloneqq ((x_{u_i}, y_{u_i}))_{i=1}^k$. Let $\Pi \in \mathcal{P}([n]^*)$ be a subsampling distribution over retained sample indices $U$, $\phi_{\mathcal{X}}(\cdot)$ a *feature noise* mechanism (e.g., $\phi(x) = \mathcal{N}(x, \sigma^2 I)$), and $\phi_{\mathcal{Y}}(\cdot)$ a *label noise* mechanism (e.g., $\phi(y) = (1-\gamma)\delta_y + \gamma\,\mathrm{Unif}([C])$). Our smoothing mechanism $\mathcal{M} \coloneqq \mathcal{M}_{\Pi,\phi_{\mathcal{X}},\phi_{\mathcal{Y}}}$ is generated by the following procedure. **(1)** Draw retained indices $U \sim \Pi$. **(2)** Apply per-example noise to features and labels: For each $i \in [|U|]$, independently sample $\widetilde{x}_i \sim \phi_{\mathcal{X}}(\cdot \mid x_{u_i})$ and $\widetilde{y}_i \sim \phi_{\mathcal{Y}}(\cdot \mid y_{u_i})$. Construct $\widetilde{S} \coloneqq ((\widetilde{x}_i, \widetilde{y}_i))_{i=1}^{|U|}$. **(3)** Sample $\widetilde{x}_{\mathrm{te}} \sim \phi_{\mathcal{X}}(\cdot \mid x_{\mathrm{te}})$. [2] **(4)** Output $(\widetilde{S}, \widetilde{x}_{\mathrm{te}})$.

We make a few observations on the distributions resulting from the smoothing mechanism $\mathcal{M}$. By construction, sampling from $\mathcal{M}$ is efficient. $\mathcal{M}$ can be written as a mixture distribution whose components are product measures parameterized by a subsample of $S$. $\mathcal{M}$ has $|\mathrm{supp}\,\Pi|$ components, which is generally quite large (e.g., $n^k$ for bagging with replacement). This makes evaluating the density (or mass) at a point a priori intractable. For a large number of norm-based smoothing mechanisms (e.g., Gaussian), the product measure induced by $\phi_{\mathcal{X}}$ reduces to a well-understood measure over the full dataset space identified with $\mathbb{R}^{nd}$ (e.g., $\prod_{i=1}^n \mathcal{N}(0_d, \sigma^2 I_d) \cong \mathcal{N}(0_{nd}, \sigma^2 I_{nd})$).

**Concrete instantiations** We motivate our smoothing scheme with a few concrete instantiations. Natural choices of $\Pi$ are bagging (with or without replacement) or Poisson subsampling. Natural choices for $\phi_{\mathcal{X}}$ include Gaussian, Laplace, Uniform, or Discrete noise, but we can also accommodate more exotic noises such as the norm-based noises of Yang et al. (2020). Most often $\phi_{\mathcal{Y}}$ will be a flipping distribution $\phi(y) = \phi^{\mathrm{flip}}_{\gamma,K}(y) = (1-\gamma)\delta_y + \gamma\mathrm{Unif}([K])$ for some alphabet size $K$, but we can also accommodate structured noises $\phi_{\mathcal{Y}}(y) = M_{y\widetilde{y}}$ for Markovian transition matrix $M$.

The smoothing mechanisms of several prior works can be recovered for suitable choices of $\Pi, \phi_{\mathcal{X}}, \phi_{\mathcal{Y}}$. RAB (Weber et al., 2023) takes $\Pi$ as a point mass on the original dataset (no subsampling) and sets $\phi(x) = \mathcal{N}(x, \sigma^2 I_d)$. Jia et al. (2021) take $\Pi$ as bagging with replacement with no feature or label noise. BagFlip (Zhang et al., 2022) considers bagging with replacement, and with coupled feature-label noise $\phi_{\mathcal{X}}(x) = \bigotimes_{j=1}^d \phi^{\mathrm{flip}}_{\gamma,K}(x_j)$ and $\phi_{\mathcal{Y}}(y) = \phi^{\mathrm{flip}}_{\gamma,K}(y)$, where notably the feature and label alphabets are the same size and use the same flipping probability.

### 4.2 Certifying predictions: overview

**Core challenges** The main technical challenge we face is how to certify robustness for such a large class of general smoothing methods against the adversaries defined in Section 3.2. Prior works applying randomized smoothing to data poisoning or backdoor attacks (Weber et al., 2023; Jia et al., 2021; Zhang et al., 2022) apply techniques introduced by Cohen et al. (2019), who frame certification as a hypothesis testing problem and construct Neyman–Pearson tests to derive tight certificates. Unfortunately, this path is only accessible for simple smoothing schemes and threat models, and does not scale to our setting: the likelihood ratio structure of $\mathcal{M}$ is complicated due to its combinatorial mixture structure and is typically only directly tractable for simple distributions such as the discrete noises of Zhang et al. (2022). Furthermore, changing the mechanisms $\Pi, \phi_{\mathcal{X}}, \phi_{\mathcal{Y}}$ or the threat model $d_{\mathcal{X}}, d_{\mathcal{Y}}, \rho$ a priori requires a laborious new derivation of the certification condition by hand.

What we desire instead is a generic way to *automatically construct* the certification conditions given the threat model $d_{\mathcal{X}}, d_{\mathcal{Y}}, \rho$ and the smoothing mechanism $\Pi, \phi_{\mathcal{X}}, \phi_{\mathcal{Y}}$. All relevant information can be derived for each component independently, and different combinations can be composed automatically with basic computational rules. It turns out that such a procedure is possible if we turn to a certain dual formulation of Randomized Smoothing, first considered by Dvijotham et al. (2020) for the evasion setting in the language of *$f$-divergences*. We start by introducing some preliminaries.

[2] We couple the train and test noise mechanisms for simplicity; decoupling them is conceptually straightforward.

### 4.3 CERTIFICATION VIA CONVEX OPTIMIZATION

Define a *specification* as a bounded measurable $\psi\colon \Omega \to \mathbb{R}$, and call it *robustly certified* at $\nu \in \mathcal{P}(\Omega)$ w.r.t. $\mathcal{D} \subseteq \mathcal{P}(\Omega)$ if $\mathbb{E}_{\nu'}[\psi] \geq 0$ for every $\nu' \in \mathcal{D}$ (Definition A.1). For convex $f$ with $f(1) = 0$ (a *generator*), the induced $f$-divergence is $D_f(\nu'\|\nu) \coloneqq \mathbb{E}_\nu[f(d\nu'/d\nu)]$ (Definition B.1). Given generators $f_1, \ldots, f_M$ with bounds $\tau_i \geq 0$, define $\mathcal{D}_\mathcal{F} \coloneqq \bigcap_i \{\nu' : D_{f_i}(\nu'\|\nu) \leq \tau_i\}$ (Definition A.2).

We use the following result of Dvijotham et al. (2020), stated in full in Theorem A.2. Let $\underline{\theta_a}, \theta_b$ denote lower and upper bounds on the majority and runner-up predicted probabilities from $\overline{G_\mathcal{M}}(w)$, respectively, and write $\theta = (\theta_a, \theta_b, 1 - \theta_a - \theta_b)$. Then, every specification consistent with $\theta_a, \theta_b$ is robustly certified w.r.t. $\mathcal{D}_\mathcal{F}$ iff the solution to the following problem is non-negative:

$$\inf_{\zeta \in \mathbb{R}^3_+} \zeta_a - \zeta_b \quad \text{s.t.} \quad \zeta_a + \zeta_b + \zeta_c = 1, \quad D_{f_i}(\zeta\|\theta) \leq \tau_i, \;\; i = 1, \ldots, m. \tag{1}$$

Robustness certification in the sense of Definition A.1 is related to our own robustness notion (Definition 3.1). Writing $c \coloneqq \overline{g_\mathcal{M}}(w)$ and taking specifications $\psi_{c,c'}(\widetilde{w}) \coloneqq G(\widetilde{w})_c - G(\widetilde{w})_{c'}$, the smoothed prediction is constant on $\mathcal{B}_\rho(w)$ iff every $\psi_{c,c'}$ with $c \neq c'$ is robustly certified at $\mathcal{M}(w)$ w.r.t. $\mathcal{D}_{w,\rho} \coloneqq \{\mathcal{M}(w') \;:\; w' \in \mathcal{B}_\rho(w)\}$. We state and prove the formal statement in Theorem D.1.

In working with Equation (1) it suffices to choose any $\tau_i \geq D_{f_i}(\nu'\|\mathcal{M}(w))$ for every $\nu' \in \mathcal{D}_{w,\rho}$, with the tightest choice being

$$\tau^\star_f(w, \rho) \coloneqq \sup_{w' \in \mathcal{B}_\rho(w)} D_f(\mathcal{M}(w')\|\mathcal{M}(w)). \tag{2}$$

One can use any $f$-divergence for which an upper bound is available. The most interesting cases for us are the Hockey-stick divergences due to their characterization of tight certificates (Theorem A.3), and the Rényi divergences due to their convenient tensorization properties (Lemma B.7).

### 4.4 COMPUTING DIVERGENCE BOUNDS

Under Theorems A.2 and D.1, the problem of certifying robustness for $\overline{G_\mathcal{M}}$ at $w \in \mathcal{W}$ reduces to establishing a bound $\tau_f(w, \rho) \geq \tau^\star_f(w, \rho)$ for suitable $f$. The technical challenge is grappling with the complex structure of $\mathcal{M}$: subsampling through $\Pi$ induces a mixture with combinatorially many components, and noising through $\phi$ makes each component a product measure. Our analysis proceeds in two steps: first, reduce the combinatorial structure of the mixture distribution to lower-dimensional problem, and second, to understand divergences between the paired product components.

**Bounding divergences for mixtures** The first key observation for bounding divergences is that the components of any $\mathcal{M}(w')$ and $\mathcal{M}(w)$ share mixing weights when paired by subsample, which interact nicely with $D_f(\mathcal{M}(w')\|\mathcal{M}(w))$. Moreover, despite a combinatorial number of components, often the components can be grouped together, which gives us the following bound:

**Theorem 4.1** (Reorganized Joint Convexity Bound)**.** *Let $\mu = \sum_{i=1}^N \theta_i \mu_i$ and $\nu = \sum_{i=1}^N \theta_i \nu_i$ be mixtures with shared mixing weights $\theta_i$. Suppose there is a partition of indices $[N] = \bigsqcup_{i=1}^k S_i$ and sequence of nonnegative real numbers $(\delta_i)_{i=1}^k$ such that for all $i \in [k]$ and all $j \in S_i$, $D_f(\mu_j\|\nu_j) \leq \delta_i$. Then one has $D_f(\mu\|\nu) \leq \tau_f$ for $\tau_f \coloneqq \sum_{i=1}^k \beta_i \delta_i$, where $\beta_i \coloneqq \sum_{j \in S_i} \theta_j$.*

We prove Theorem 4.1 in Appendix D, along with an analogue for Rényi divergences (Corollary D.1). Theorem 4.1 does not depend on the specific geometry of the feature or label noise and only uses pairing of components and joint convexity of $D_f$. The grouping of components is important computationally as it reduces the sum to a tractable number of terms. For our noises, the component bound depends on $U$ only through the number $s(U)$ of corrupted indices. Grouping by $s$ leaves only few terms, e.g., $k+1$ for bagging and $\min(k, m_\mathrm{p})+1$ for bagging without replacement. This trick can be viewed as an extension of earlier techniques to partition the output space into regions of constant likelihood ratio (Lee et al., 2019; Bojchevski et al., 2020; Zhang et al., 2022) to $f$-divergences.

### 4.5 BOUNDING COMPONENT DIVERGENCES

It remains to find the bounds $\delta_j$. Consider a fixed corruption $w'$ of $w$, and apply Theorem 4.1. Writing $z_U \coloneqq (S_U, x_\mathrm{te})$ for an un-noised subsampled input, we obtain the component pairs $(\phi(z_U), \phi(z'_U))$

with $z'_U$ determined by $w'$. A small complication is that the $z_U$ are coupled through $w'$, so we relax the problem slightly and take $\delta_j \geq \sup_{z'} D_f(\phi(z')\|\phi(z))$ over all corruptions $z'$ of $z$ that are feasible given group $j$ (for example, a component might correspond to $U$ with $s$ corrupted examples). This is tight when some $w'$ attains every component supremum simultaneously.

**Adjacency parameters** The supremum $\sup_{z'} D_f(\phi(z')\|\phi(z))$ is natively over a high-dimensional space. However, for sufficiently symmetrical noises, the divergence between $(x, x')$ depends only on a low-dimensional *adjacency parameter* $\theta(x, x')$, in the sense that every $f$-divergence can be written in the form $D_f(\phi(x')\|\phi(x)) = \delta_f(\theta(x, x'))$ for some $\delta_f$. One way to see this is by invariance of $f$-divergences under bimeasurable bijections (Lemma B.3). As a concrete example, we can see that for Gaussian noise any $f$-divergence depends only on $r = \|x - x'\|_2/\sigma$.

Table 1: Divergence bounds derived in Appendix G. Here $\theta$ is the adjacency parameter of Definition G.1 and $\Delta \coloneqq x - x'$. Notation $p_+$, $p_0$, $p_-$, $\kappa_\alpha(t)$, $d_\beta$, and $P_d$ are defined in Appendix G.

| **Noise** | **Adjacency** $\theta$ | $D_{\mathrm{R},\alpha}(\mu\|\nu)$ | $D_{\mathrm{HS},\beta}(\mu\|\nu)$ |
|---|---|---|---|
| $\mathrm{Unif}(\gamma)$ | $1 - \prod_i (1 - \frac{\lvert\Delta_i\rvert}{2\gamma})_+$ | $\frac{1}{\alpha-1}\log(1-\theta)$ | $\min(1, \beta)\,\theta$ |
| $\mathcal{N}(\cdot, \sigma^2 I)$ | $r = \lVert\Delta\rVert_2/\sigma$ | $\alpha r^2/2$ | $\Phi\big(\frac{r}{2} - \frac{\log\beta}{r}\big) - \beta\Phi\big(-\frac{r}{2} - \frac{\log\beta}{r}\big) - (1-\beta)_+$ |
| $\mathrm{Lap}(\cdot, b)$ | $\mathbf{t} = \lvert\Delta\rvert/b$ | $\sum_j \kappa_\alpha(\mathbf{t}_j)$ | $\big[(1 - \sqrt{\beta}e^{-t/2})_+ - (1-\beta)_+\big]_+$ |
| $\phi^{\text{flip}}_{\gamma,C}$ | $s = \lVert\Delta\rVert_0$ | $\frac{s}{\alpha-1}\log\big(p_+^\alpha p_-^{1-\alpha} + p_+^{1-\alpha}p_-^\alpha + p_0\big)$ | $\sum_{d \geq d_\beta}\big(1 - \beta(p_+/p_-)^{-d}\big)P_d$ |

**Worst-case pairs** A priori Equation (2) admits a different maximizer for each $f$. In most cases, this is fortunately not the case. Write $(\mu', \nu') \preceq (\mu, \nu)$ when $D_f(\mu'\|\nu') \leq D_f(\mu\|\nu)$ for every generator $f$. We show that $\preceq$ is entirely determined by the Hockey-stick envelope (Lemma G.1): $(\mu', \nu') \preceq (\mu, \nu)$ iff $D_{\mathrm{HS},\beta}(\mu'\|\nu') \leq D_{\mathrm{HS},\beta}(\mu\|\nu)$ for all $\beta > 0$. In such cases a single $(\mu^\star, \nu^\star)$ generates every $\tau_f^\star$ simultaneously. This property also composes nicely under product measures (Lemma G.2), which is useful for combining feature and label flipping. We show that such a pair exists for Gaussian noise under all $\ell_p$ norms, for flipping noise under $\ell_0$, and for Laplace and Uniform noise under $\ell_1$ and $\ell_\infty$ norms. For Laplace and Uniform noise under $\ell_2$, the behavior is more complicated, and we leverage sound but loose relaxations (Appendices G.3 and G.4).

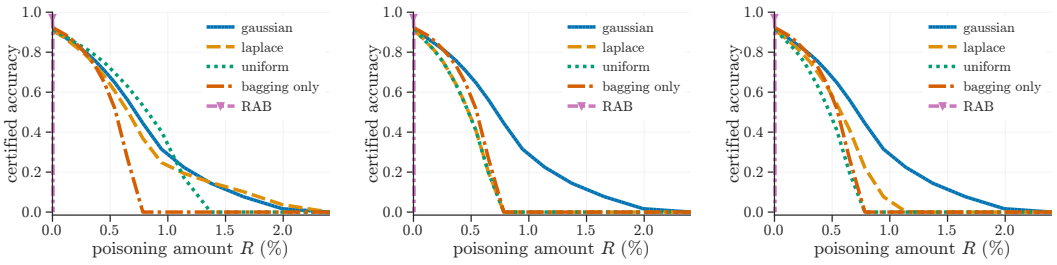
**Composition under products** Our framework involves applying independent noise (between features and labels within an example, and independence between examples), which result in the need to understand $f$-divergences between product measures. Rényi divergences are attractive here due to their nice tensorization properties (Lemma B.7) and can be computed immediately. We show in Lemma G.3 how to compute Hockey-stick divergences $D_{\mathrm{HS},\beta}(\mu_1 \otimes \mu_2\|\nu_1 \otimes \nu_2)$ in closed form via the formula $\widetilde{D}_{\mathrm{HS},\beta}(\mu\|\nu) = \mathbb{E}_{\mu_1}[\widetilde{D}_{\mathrm{HS},\beta/\Lambda_1}(\mu_2\|\nu_2)]$. This is especially useful for modeling dirty-label adversaries, where corresponding noise $\phi$ can be written as a product of a label flipping measure $\mu_1 = \phi^{\text{flip}}_{\gamma,C}$ and a continuous noise $\mu_2$ (e.g., $\mathcal{N}(x, \sigma^2 I)$). Lemma G.3 also translates into a computational method for numerically computing divergences when every component but one is discrete, and the hockey stick profile of the remaining one is understood.

**Worked example** Consider $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = [C]$, and $S = ((x_i, y_i))_{i=1}^n$. Let $\Pi$ be bagging with replacement of $k$ samples, and let $\phi_{\mathcal{X}}(x) = \mathcal{N}(x, \sigma^2 I_d)$, so that $\mathcal{M}(S) = \sum_{U \in [n]^k} n^{-k}\mathcal{N}(S_U, \sigma^2 I_{dk})$ has $n^k$ components. Under an $\ell_2$ adversary, a component with $s$ corrupted samples has adjacency parameter $r = \|S_U - S'_U\|_2/\sigma \leq \varepsilon_{\mathrm{f}}\sqrt{s}/\sigma$, which is attained at maximal corruption magnitude. Since $s \sim \mathrm{Binom}(k, m_{\mathrm{p}}/n)$ under $\Pi$, Theorem 4.1 gives $\tau_f = \sum_{s=0}^k \mathrm{Binom}(s; k, m_{\mathrm{p}}/n)\delta_f(\varepsilon_{\mathrm{f}}\sqrt{s}/\sigma)$, where $\delta_f$ depends on $f$ (e.g., $\delta_{\mathrm{TV}}(r) = 2\Phi(r/2) - 1$; see Table 1). If we also model a backdoor trigger, we can replace $\varepsilon_{\mathrm{f}}^2\, s$ by $\varepsilon_{\mathrm{f}}^2\, s + \varepsilon_{\mathrm{te}}^2$ to give $\delta_f(\sqrt{\varepsilon_{\mathrm{f}}^2\, s + \varepsilon_{\mathrm{te}}^2}/\sigma)$ in the sum. If we model label flipping adversaries and noise labels, we can compute $\delta_\beta$ using Lemma G.3 to be a weighted sum of Gaussian hockey stick values at thresholds $\beta(p_+/p_-)^{-d}$ (Theorem G.5) computable numerically.

# 5 EVALUATION

We present several applications of our framework to data poisoning and backdoor poisoning attacks. Our experiments focus on image classification of the MNIST (Lecun et al., 1998) and CIFAR10 (Krizhevsky, 2009) datasets, in addition to class-subsampled variants MNIST-17 and CIFAR10-02 (Zhang et al., 2022; Weber et al., 2023). We train ensembles of size $N = 10,000$ and $N = 1,000$ for MNIST and CIFAR10 datasets, respectively. We use confidence level $\alpha = 0.001$.

## 5.1 $\ell_p$ FEATURE ADVERSARIES




(a) $\ell_1$ adversary, $\varepsilon_{\mathrm{f}} = 1$. (b) $\ell_2$ adversary, $\varepsilon_{\mathrm{f}} = 1$. (c) $\ell_\infty$ adversary, $\varepsilon_{\mathrm{f}} = 1/28$.

Figure 1: Certified accuracies on MNIST for clean-label triggerless attacks under different $\ell_p$ adversaries. The budget at each is calibrated to correspond to the same $\ell_2$ budget, which guarantees the Gaussian noise certifies the same radius across all curves.

**Setup** We consider a setting in which the attacker corrupts at most $m_{\mathrm{p}}$ training examples, each by at most $\|x - x'\|_p \leq \varepsilon_{\mathrm{f}}$. The test sample and labels are not corrupted. We instantiate $\Pi$ as bagging and compare Gaussian, Laplace, and Uniform feature noise. We also evaluate pure bagging (Jia et al., 2021) and Gaussian noise without subsampling (Weber et al., 2023) baselines.

No prior method makes effective use of both $m_{\mathrm{p}}$ and $\varepsilon_{\mathrm{f}}$ constraints. For example, Jia et al. (2021) only subsamples and pessimistically assumes a single retained poison compromises the learner, while Weber et al. (2023) does not subsample data. Zhang et al. (2022) uses discrete noise that only works under $\ell_0$ constraint, which cannot certify more points than plain bagging: while $\ell_p$ constraints can be converted into (looser) $\ell_q$ constraints for $p, q > 0$, they cannot convert into nontrivial $\ell_0$ constraints.

For $p > 0$, we compare with the certificate obtainable from the RAB defense (Weber et al., 2023), which adds Gaussian noise to the entire training set without bagging. RAB's certificate can be used in our threat model by computing the maximum admissible $\ell_2$ perturbation under our threat model. In particular, an $\ell_p$ bound can be converted into an $\ell_2$ bound by taking $\varepsilon_{\mathrm{f}}^{(2)} \leq \varepsilon_{\mathrm{f}}^{(1)}$ or $\varepsilon_{\mathrm{f}}^{(2)} \leq \sqrt{d}\varepsilon_{\mathrm{f}}^{(\infty)}$, respectively, and with $m_{\mathrm{p}}$ poisoned examples we obtain the full dataset $\ell_2$ bound $\sqrt{m_{\mathrm{p}}}\varepsilon_{\mathrm{f}}^{(2)}$.

**Findings** We make some interesting observations about the clean-label poisoning robustness reported in Figure 1. First, pure bagging always attains a higher non-robust certified accuracy, which makes sense since adding noise to data should only reduce classifier accuracy. Second, uniform noise is competitive only for $\ell_1$ and for moderate poisoning amounts ($< 1.2\%$). This finding corroborates with Yang et al. (2020), who find in the evasion setting that Gaussian noise yields the largest certified accuracies for $\ell_2$ and $\ell_\infty$ whereas Uniform noise works best for $\ell_1$. Laplace noise is only competitive for $\ell_1$ adversaries at high poisoning rates, where it marginally beats Gaussian noise for $R \geq 1.5$.

We show the dependence on $\varepsilon_{\mathrm{f}}$ for MNIST, MNIST-17, CIFAR10, and CIFAR10-02 in Figure 2 when training under Gaussian noise. In this case the advantage of constraining $\varepsilon_{\mathrm{f}}$ deteriorates quickly: on MNIST, the certified radius is substantially larger for $\varepsilon_{\mathrm{f}} = 1$, but by $\varepsilon_{\mathrm{f}} = 2$ the certificate nearly matches that of the one at $\varepsilon_{\mathrm{f}} = \infty$ (corresponding to the bagging certificate of Jia et al. (2021)).

## 5.2 LABEL FLIPPING AND BACKDOOR ADVERSARIES

**Setup** We study the impact of different threat models on certifiable robustness. We consider two main attacker settings: joint feature-label attacks and joint train-test (backdoor) attacks. In both

settings the adversary can corrupt at most $m_{\mathrm{p}}$ training examples and, for each corrupted sample, modify features up to $\varepsilon_{\mathrm{f}}$ distance in $\ell_2$. In feature-label attacks, the adversary can additionally corrupt the label. In backdoor attacks, the adversary also corrupts the test sample with test budget $\varepsilon_{\mathrm{te}} = \varepsilon_{\mathrm{f}}$.

**Findings** We report the robustness tradeoffs for different datasets. The label-flipping adversary is reported in Table 2, and the backdoor advesary is reported in Table 3.

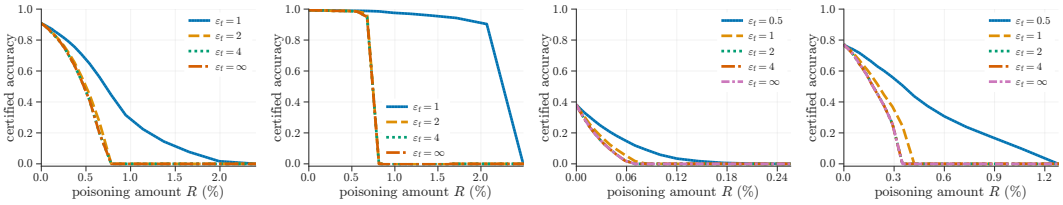
Our method significantly outperforms RAB under backdoor adversaries. On MNIST-17, while RAB reaches trivial certified accuracy after $R = 0.5\%$, our method retains over 45% certified at $R = 2\%$. The improvement is due to bagging: with high probability we retain only a small number $\mathcal{O}(k \cdot m_{\mathrm{p}}/n)$ of poisoned samples, whereas RAB retains all poisoned samples. For RAB to certify to comparable poisoned radii, it would need to train with $\sigma = \mathcal{O}(\varepsilon_{\mathrm{f}}\sqrt{m_{\mathrm{p}}})$, which would destroy predictive performance, whereas for our method $\sigma = \mathcal{O}(\varepsilon_{\mathrm{f}}\sqrt{m_{\mathrm{p}}\, k/n})$ can still yield large certificates.

Table 2: Robustness tradeoff for joint feature & label adversaries at poisoned fractions $R$, with $\varepsilon_{\mathrm{l}} = 1$ and $\ell_2$ budget $\varepsilon_{\mathrm{f}} = 1$ (MNIST, MNIST-17) or 0.5 (CIFAR10, CIFAR10-02). Bag sizes $k$, and noise scales $\sigma_a < \sigma_b$, are fixed per dataset (Table 7). Label noise retains the true label with $p = 0.9$.

| | **MNIST** | | | | **MNIST-17** | | | | **CIFAR10** | | | | **CIFAR10-02** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $R$ (%) | 0 | 0.5 | 1 | 1.5 | 0 | 1 | 2.5 | 3.5 | 0 | 0.05 | 0.1 | 0.15 | 0 | 0.8 | 1.6 | 2.4 |
| Bagging | **82.3** | **56.8** | **31.9** | **12.0** | **99.1** | **98.0** | 0.0 | 0.0 | **32.5** | **14.7** | **5.3** | **1.1** | **78.3** | **64.4** | **40.2** | 2.4 |
| Ours, $\sigma_a$ | 81.2 | 50.0 | 23.5 | 5.1 | 98.7 | 96.6 | **48.6** | 0.0 | 22.3 | 9.8 | 3.3 | 0.6 | 66.9 | 48.5 | 26.6 | 0.6 |
| Ours, $\sigma_b$ | 78.7 | 44.8 | 17.9 | 3.7 | 98.8 | 93.0 | 35.3 | **5.3** | 16.7 | 6.4 | 1.7 | 0.1 | 64.2 | 45.4 | 27.5 | **3.4** |

Table 3: Robustness tradeoff for clean label backdoor adversaries on MNIST, MNIST-17, CIFAR10, CIFAR10-02 at different poisoned fractions $R$. The $\ell_2$ constraint is $\varepsilon_{\mathrm{f}} = \varepsilon_{\mathrm{te}} = 1$ on MNIST and MNIST-17 and $\varepsilon_{\mathrm{f}} = \varepsilon_{\mathrm{te}} = 0.5$ on CIFAR10 and CIFAR10-02. $\sigma_a, \sigma_b$ are as in Table 7. RAB trains on the full training set.

| | **MNIST** | | | | **MNIST-17** | | | | **CIFAR10** | | | | **CIFAR10-02** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $R$ (%) | 0 | 0.1 | 0.5 | 1 | 0 | 0.5 | 1 | 2 | 0 | 0.02 | 0.06 | 0.1 | 0 | 0.2 | 0.4 | 0.6 |
| RAB, $\sigma_a$ | **74.1** | 0.0 | 0.0 | 0.0 | **95.9** | 0.0 | 0.0 | 0.0 | 18.4 | 0.0 | 0.0 | 0.0 | 63.5 | 0.0 | 0.0 | 0.0 |
| RAB, $\sigma_b$ | 66.5 | 0.0 | 0.0 | 0.0 | 94.7 | 0.0 | 0.0 | 0.0 | **19.3** | 0.0 | 0.0 | 0.0 | **67.5** | 0.0 | 0.0 | 0.0 |
| Ours, $\sigma_a$ | 48.6 | **42.2** | **20.5** | 5.0 | 92.9 | **88.0** | **76.6** | 41.8 | 6.8 | 5.6 | 3.8 | 2.5 | 40.9 | 34.4 | 27.5 | 20.2 |
| Ours, $\sigma_b$ | 41.9 | 36.6 | 17.8 | **6.2** | 91.3 | 86.1 | 76.1 | **45.4** | 8.7 | **7.9** | **6.5** | **5.3** | 54.0 | **51.4** | **48.7** | **45.8** |




(a) MNIST (b) MNIST-17 (c) CIFAR10 (d) CIFAR10-02

Figure 2: Certified accuracy versus poisoning fraction $R$ at each per-example budget $\varepsilon_{\mathrm{f}}$, under an $\ell_2$ adversary with Gaussian feature noise. $\sigma = 0.5$ for MNIST and $\sigma = 0.35$ for CIFAR10.

**When is more noise useful?** Aggressive noise erases corruptions, which corresponds to larger certified radii via smaller divergence bounds, but also degrades the base learner, lowering the vote confidence. In general, two training data noising strategies exist: either we use aggressive plain bagging (hoping to throw away bad examples), or we try to recover performance from retained bad examples by adding additional noise to them. It is therefore important to understand how different noises affect the resulting certificates.

As a simple example, consider $\ell_2$ corruptions with Gaussian noise. To keep the component divergences small, the noise level $\sigma$ must grow with the adversary budget $\varepsilon_{\mathrm{f}}$. If $\sigma$ becomes large, the input becomes impossible to classify, and the base learner accuracy will drop dramatically. In this regime, it is preferable to use no feature noising at all and rely on pure bagging to provide certificates.

Label noising is especially delicate (Table 2): we only observe improvements over the plain bagging method for binary classification datasets MNIST-17 and CIFAR10-02. This is intuitive: whereas features can be noised individually while still preserving high information content taken over the full input, the label serves as the only supervision signal for the learning task, so noising it is especially destructive. Still, label flipping can be effective for sufficiently easy tasks: on MNIST-17, we observe certified radius out to $3.5\%$ of the training dataset achieved by aggressive feature and label noising compared with the plain bagging baseline.

### 5.3 Comparison with prior works

Our method recovers the exact certificates of Zhang et al. (2022) under all threat models and smoothing strategies they consider (Appendix E.3, Figure 3), and those of Jia et al. (2021) for replacement adversaries (Appendix E.2).

Another useful finding is that, by writing down the hockey stick divergences in closed-form, we are able to implement certification for a fixed $\varepsilon_{\mathrm{f}}$ bound in a much more efficient $\mathcal{O}(k^2\varepsilon_{\mathrm{f}}^2)$ compared with BagFlip's $\mathcal{O}(d^3 + k^2 d^2)$ preprocessing cost when $\varepsilon_{\mathrm{f}} \ll d$. Notably, we only depend on $d$ insofar as through $\varepsilon_{\mathrm{f}}$. Our formulation is also amenable to parallelized precomputation across all desired $\varepsilon_{\mathrm{f}}$, whereas the linear recurrence of Zhang et al. (2022) is not.

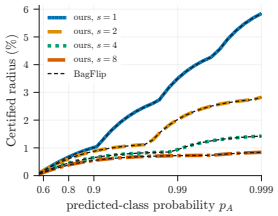



Figure 3: Tightness of our certificate with BagFlip's.

Because BagFlip couples the feature space to the label space, they can only certify dirty-label attacks when $|\mathcal{X}| = |\mathcal{Y}|$ and when labels are flipped with the same distribution as features. This in particular means they cannot certify dirty-label attacks in the multi-class setting if the features are binary. Our framework in contrast handles both of these cases directly. We are able to compute certificates in a setting where the label noise is different from the feature flipping noise, or where the number of distinct feature values is different from the number of classes.

## 6 Conclusion

We provide a randomized smoothing framework for certified robustness against backdoor and data poisoning attacks. Our framework decomposes threat model axes into independent dimensions bounding corruption magnitudes, and applies composite noise comprised of data subsampling, feature noising, and label flipping. We reduce certification in this regime to a convex program over $f$-divergence constraints and show how to compute the corresponding constraint sets comprehensively. Our constraint computation leverages rearranged joint convexity to efficiently grapple with the mixture distribution structure of subsampling and uses general, composable principles to accommodate general smoothing measures. We demonstrate this generality by certifying, for the first time, different $\ell_1$, $\ell_2$, and $\ell_\infty$ adversaries on models trained with Gaussian, Laplace, Uniform, and discrete noise.

**Limitations and open questions** Our method trains thousands of base learners for sound prediction estimation. We mitigate this cost with an efficient parallelized training implementation, but future methods might seek efficient closed-form solutions such as those considered in prior work (Rosenfeld et al., 2020; Weber et al., 2023). Our method does not exploit properties specific to the dataset or base learner. Future work should consider whether using such information can yield stronger certificates. An interesting remaining challenge is to precisely characterize the structural properties of optimal noising schemes for complex backdoor poisoning attack threat models. For example, we do not understand very well the tradeoff between subsampling and data noising. Finally, learner-based robustness properties (such as data-smoothness) might be leveraged to obtain even tighter certificates.

### AI USE STATEMENT

In this work, we used generative AI tools to implement methods, search for information, discover relevant literature, formulate mathematical claims, provide critical ingredients for proving mathematical claims, and assist in the writing of proofs. We have not used generative AI tools to format datasets, generate synthetic data sets, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, support qualitative and thematic data analysis, or interpret results. Developing theoretical models or conceptual frameworks is not applicable to this work. We have reviewed all AI-assisted work: the formulation of theoretical results was performed by the authors and all proof assistance was reviewed and reproduced manually. All LLM-generated code was verified manually. Activities involving discovery of prior work were conducted with multiple methods including LLM-guided deep research, academic search engines, and manual citation. All bibliographic metadata was collected manually. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

### REPRODUCIBILITY STATEMENT

Section 4 defines the smoothing mechanism and certification procedure. Appendix F gives additional experimental details including training parameters. Key theoretical claims are proved in Appendices D, E and G. We make the code available at `https://anonymous.4open.science/r/certified-backdoor-9E74/`, which contains all scripts needed to reproduce our experiments.

## REFERENCES

Hojjat Aghakhani, Dongyu Meng, Yu-Xiang Wang, Christopher Kruegel, and Giovanni Vigna. Bullseye polytope: A scalable clean-label poisoning attack with improved transferability. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 159–178, 2021. doi: 10.1109/EuroSP51992.2021.00021.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 274–283. PMLR, July 2018. URL `https://proceedings.mlr.press/v80/athalye18a.html`.

Borja Balle and Yu-Xiang Wang. Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 394–403. PMLR, July 2018. URL `https://proceedings.mlr.press/v80/balle18a.html`.

Aleksandar Bojchevski, Johannes Gasteiger, and Stephan Günnemann. Efficient robustness certificates for discrete data: Sparsity-aware randomized smoothing for graphs, images and more. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1003–1013. PMLR, July 2020. URL `https://proceedings.mlr.press/v119/bojchevski20a.html`.

Ruoxin Chen, Zenan Li, Jie Li, Junchi Yan, and Chentao Wu. On collective robustness of bagging against data poisoning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 3299–3319. PMLR, July 2022. URL `https://proceedings.mlr.press/v162/chen22k.html`.

C. J. Clopper and Egon S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, December 1934. ISSN 0006-3444. doi: 10.1093/biomet/26.4.404. URL `https://doi.org/10.1093/biomet/26.4.404`.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning*

*Research*, pp. 1310–1320. PMLR, June 2019. URL https://proceedings.mlr.press/v97/cohen19c.html.

Imre Csiszár and Paul C. Shields. Information theory and statistics: A tutorial. *Foundations and Trends in Communications and Information Theory*, 1(4):417–528, November 2004. ISSN 1567-2190. doi: 10.1561/0100000004. URL https://doi.org/10.1561/0100000004.

Samuel Drews, Aws Albarghouthi, and Loris D'Antoni. Proving data-poisoning robustness in decision trees. In *Proceedings of the 41st ACM SIGPLAN Conference on Programming Language Design and Implementation*, PLDI 2020, pp. 1083–1097, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450376136. doi: 10.1145/3385412.3385975. URL https://doi.org/10.1145/3385412.3385975.

Krishnamurthy (Dj) Dvijotham, Jamie Hayes, Borja Balle, Zico Kolter, Chongli Qin, Andras Gyorgy, Kai Xiao, Sven Gowal, and Pushmeet Kohli. A framework for robustness certification of smoothed classifiers using f-divergences. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJlKrkSFPH.

Jonas Geiping, Liam H Fowl, W. Ronny Huang, Wojciech Czaja, Gavin Taylor, Michael Moeller, and Tom Goldstein. Witches' brew: Industrial scale data poisoning via gradient matching. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=01olnfLIbD.

Lukas Gosch, Xichuan Chen, Yan Scholten, and Stephan Günnemann. Certifying graph neural networks against label and structure poisoning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=kQuvO5gaNr.

Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019. doi: 10.1109/ACCESS.2019.2909068.

Zayd Hammoudeh and Daniel Lowd. Provable robustness against a union of $\ell_0$ adversarial attacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19):21134–21142, March 2024. doi: 10.1609/aaai.v38i19.30106. URL https://ojs.aaai.org/index.php/AAAI/article/view/30106.

Jinyuan Jia, Xiaoyu Cao, and Neil Zhenqiang Gong. Intrinsic certified robustness of bagging against data poisoning attacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9):7961–7969, May 2021. doi: 10.1609/aaai.v35i9.16971. URL https://ojs.aaai.org/index.php/AAAI/article/view/16971.

Jinyuan Jia, Yupei Liu, Xiaoyu Cao, and Neil Zhenqiang Gong. Certified robustness of nearest neighbors against data poisoning and backdoor attacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(9):9575–9583, June 2022. ISSN 2159-5399. doi: 10.1609/aaai.v36i9.21191. URL http://dx.doi.org/10.1609/aaai.v36i9.21191.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 656–672, 2019. doi: 10.1109/SP.2019.00044.

Guang-He Lee, Yang Yuan, Shiyu Chang, and Tommi Jaakkola. Tight certificates of adversarial robustness for randomly smoothed classifiers. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/fa2e8c4385712f9a1d24c363a2cbe5b8-Paper.pdf.

Alexander Levine and Soheil Feizi. Deep partition aggregation: Provable defenses against general poisoning attacks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YUGG2tFuPM.

Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/335cd1b90bfa4ee70b39d08a4ae0cf2d-Paper.pdf.

Friedrich Liese and Igor Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006. doi: 10.1109/TIT.2006.881731.

Shijie Liu, Andrew C. Cullen, Paul Montague, Sarah M. Erfani, and Benjamin I. P. Rubinstein. Enhancing the antidote: Improved pointwise certifications against poisoning attacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7):8861–8869, June 2023. doi: 10.1609/aaai.v37i7.26065. URL https://ojs.aaai.org/index.php/AAAI/article/view/26065.

Yuzhe Ma, Xiaojin Zhu, and Justin Hsu. Data poisoning against differentially-private learners: Attacks and defenses. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 4732–4738. International Joint Conferences on Artificial Intelligence Organization, July 2019. doi: 10.24963/ijcai.2019/657. URL https://doi.org/10.24963/ijcai.2019/657.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pp. 263–275, 2017. doi: 10.1109/CSF.2017.11.

Jerzy Neyman and Egon Sharpe Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A: Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, February 1933. ISSN 0264-3952. doi: 10.1098/rsta.1933.0009. URL https://doi.org/10.1098/rsta.1933.0009.

Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2025.

Keivan Rezaei, Kiarash Banihashem, Atoosa Chegini, and Soheil Feizi. Run-off election: Improved provable defense against data poisoning attacks. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29030–29050. PMLR, July 2023. URL https://proceedings.mlr.press/v202/rezaei23a.html.

Elan Rosenfeld, Ezra Winston, Pradeep Ravikumar, and Zico Kolter. Certified robustness to label-flipping attacks via randomized smoothing. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8230–8241. PMLR, July 2020. URL https://proceedings.mlr.press/v119/rosenfeld20b.html.

Mahalakshmi Sabanayagam, Lukas Gosch, Stephan Günnemann, and Debarghya Ghoshdastidar. Exact certification of (graph) neural networks against label poisoning. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 25663–25691, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/401aa72e0e3be680348a5b0ffdb1a5aa-Paper-Conference.pdf.

Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. Hidden trigger backdoor attacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11957–11965, April 2020. ISSN 2159-5399. doi: 10.1609/aaai.v34i07.6871. URL http://dx.doi.org/10.1609/aaai.v34i07.6871.

Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/3a24b25a7b092a252166a1641ae953e7-Paper.pdf.

Aman Saxena, Jan Schuchardt, Yan Scholten, and Stephan Günnemann. Provable robustness against backdoor attacks via the primal-dual perspective on differential privacy, May 2026. URL https://arxiv.org/abs/2605.21780.

Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/22722a343513ed45f14905eb07621686-Paper.pdf.

Philip Sosnin, Mark Niklas Mueller, Maximilian Baader, Calvin Tsay, and Matthew Robert Wicker. Certified robustness to data poisoning in gradient-based training. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=9WHifn9ZVX.

Hossein Souri, Liam Fowl, Rama Chellappa, Micah Goldblum, and Tom Goldstein. Sleeper agent: Scalable hidden trigger backdoors for neural networks trained from scratch. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 19165–19178. Curran Associates, Inc., 2022. doi: 10.52202/068431-1393. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/79eec295a3cd5785e18c61383e7c996b-Paper-Conference.pdf.

Jacob Steinhardt, Pang Wei Koh, and Percy S Liang. Certified defenses for data poisoning attacks. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/9d7311ba459f9e45ed746755a32dcd11-Paper.pdf.

Wenxiao Wang, Alexander J Levine, and Soheil Feizi. Improved certified defenses against data poisoning with (deterministic) finite aggregation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 22769–22783. PMLR, July 2022. URL https://proceedings.mlr.press/v162/wang22m.html.

Stanley L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965. doi: 10.1080/01621459.1965.10480775. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1965.10480775. PMID: 12261830.

Maurice Weber, Xiaojun Xu, Bojan Karlaš, Ce Zhang, and Bo Li. RAB: Provable robustness against backdoor attacks. In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 1311–1328, 2023. doi: 10.1109/SP46215.2023.10179451.

Chulin Xie, Yunhui Long, Pin-Yu Chen, Qinbin Li, Sanmi Koyejo, and Bo Li. Unraveling the connections between privacy and certified robustness in federated learning against poisoning attacks. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, CCS '23, pp. 1511–1525, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400700507. doi: 10.1145/3576915.3623193. URL https://doi.org/10.1145/3576915.3623193.

Greg Yang, Tony Duan, J. Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning*

*Research*, pp. 10693–10705. PMLR, July 2020. URL https://proceedings.mlr.press/v119/yang20c.html.

Yuhao Zhang, Aws Albarghouthi, and Loris D' Antoni. BagFlip: A certified defense against data poisoning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 31474–31483. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/cc19e4ffde5540ac3fcda240e6d975cb-Paper-Conference.pdf.

Yuqing Zhu, Jinshuo Dong, and Yu-Xiang Wang. Optimal accounting of differential privacy via characteristic function. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 4782–4817. PMLR, March 2022. URL https://proceedings.mlr.press/v151/zhu22c.html.

# A Extended Background on Randomized Smoothing

## A.1 Tightness of Randomized Smoothing

Historically, several analytical approaches have been applied to derive sound robustness guarantees under randomized smoothing procedures. Originally, the differential privacy guarantees of mechanisms like adding Gaussian noise were used to derive sound but slack bounds on the certified radius Lecuyer et al. (2019). Other, tighter, but still slack bounds were derived by analyzing connection between Renyi divergence and certified robustness Li et al. (2019). Cohen et al. Cohen et al. (2019) were the first to derive provably optimal robustness certificates by analyzing the connection between robustness certification and hypothesis testing. Salman et al. Salman et al. (2019) derived the same tight certificate as a consequence of a nonlinear Lipschitz continuity property of Weierstrass transforms.

Dvijotham et al. Dvijotham et al. (2020) greatly extend the results of Li et al. Li et al. (2019) to formulate general robustness certification as a convex optimization problem. This optimization problem involves bounding various statistical divergences between noised versions of the clean and corrupted algorithm inputs. The general formulation not tight under most choices of divergence. However, Dvijotham et al. demonstrate a connection with hockey-stick divergences which can recover optimal certificates in the general case. Moreover, hockey-stick divergences are connected to differential privacy (specifically, $(\varepsilon, \delta)$-approximate differential privacy), thereby connecting privacy, convex optimization and hypothesis testing together as tools for randomized smoothing certification.

## A.2 $f$-Divergences

In this section we present the certified robustness framework of Dvijotham et al. (2020), based on convex programming. Dvijotham et al. (2020) present three main results important for our work. The first result, Theorem A.1, characterizes robustness of a specification under $f$-divergence constraints in terms of a convex program. The second result, Theorem A.2, characterizes the robustness over a broader space of specifications, aligning with the more typical certified robustness setting. The third results, Theorem A.3, demonstrates that under the setting of Theorem A.2, the Hockey-stick divergences of all orders are sufficient to recover tight certificates in the original primal certification problem.

**Definition A.1** (Specification; robust certification)**.** *A* specification *is a bounded measurable function* $\psi\colon \Omega \to \mathbb{R}$*. We say* $\psi$ *is* robustly certified at $\nu$ with respect to $\mathcal{D}$ *if* $\mathbb{E}_{w\sim\nu'}[\psi(w)] \geq 0$ *for all* $\nu' \in \mathcal{D}$*.*

**Definition A.2** ($f$-divergence constraint set)**.** *For convex* $f\colon \mathbb{R}_+ \to \mathbb{R}$ *with* $f(1) = 0$*, the* $f$*-divergence from* $\nu$ *to* $\nu'$ *is* $D_f(\nu'\|\nu) \coloneqq \mathbb{E}_{\omega\sim\nu}[f(d\nu'(\omega)/d\nu(\omega))]$*. Given a collection* $\mathcal{F}$ *of functions* $f$ *each defining an* $f$*-divergence, and to each* $f$ *an associated bound* $\tau_f \geq 0$*, define the* $\mathcal{F}$*-constraint set as*

$$\mathcal{D}_{\mathcal{F}} \coloneqq \textstyle\bigcap_{f\in\mathcal{F}} \mathcal{D}_f = \{\nu : \forall f \in \mathcal{F} \;\; D_f(\nu'\|\nu) \leq \tau_f\}. \tag{3}$$

**Theorem A.1** (Dvijotham et al. (2020))**.** *Let* $\mathcal{D}_{\mathcal{F},\tau}$ *be as in Definition A.2. For* $\lambda = (\lambda_1, \ldots, \lambda_M) \in \mathbb{R}_+^M$*, define* $f_\lambda(u) \coloneqq \sum_{i=1}^M \lambda_i f_i(u)$ *and denote its convex conjugate by* $f_\lambda^*$*. A specification* $\psi$ *is*

*robustly certified at $\nu$ with respect to $\mathcal{D}_{\mathcal{F},\tau}$ (in the sense of Definition A.1) if and only if the optimal value of the convex program*

$$\sup_{\lambda\in\mathbb{R}_+^M,\kappa\in\mathbb{R}} \kappa - \sum_{i=1}^{M}\lambda_i\tau_i - \mathop{\mathbb{E}}_{w\sim\nu}\left[f_\lambda^*(\kappa-\psi(w))\right] \tag{4}$$

*is non-negative.*

**Theorem A.2** (Dvijotham et al. (2020))**.** *Let $\mathcal{D}_{\mathcal{F},\tau}$ be as in Definition A.2, and let $\theta_a,\theta_b\geq 0$ satisfy $\theta_a+\theta_b\leq 1$. Define the specification class*

$$\mathcal{G}(\theta_a,\theta_b) \coloneqq \left\{ \psi\colon \mathcal{W}\to\{-1,0,+1\} \;\middle|\; \begin{array}{l} \Pr\limits_{w\sim\nu}[\psi(w)=+1]\geq\theta_a \\ \Pr\limits_{w\sim\nu}[\psi(w)=-1]\leq\theta_b \end{array} \right\}. \tag{5}$$

*Every $\psi\in\mathcal{G}(\theta_a,\theta_b)$ is robustly certified at $\nu$ with respect to $\mathcal{D}_{\mathcal{F},\tau}$* if and only if *the optimal value of the convex program*

$$\begin{aligned} \inf_{\zeta_a,\zeta_b,\zeta_c\geq 0} \quad & \zeta_a-\zeta_b \\ \textit{s.t.} \quad & \zeta_a+\zeta_b+\zeta_c=1, \\ & D_{f_i}(\zeta\|\theta)\leq\tau_i, \quad i=1,\dots,M \end{aligned} \tag{6}$$

*is non-negative, where $\theta=(\theta_a,\theta_b,1-\theta_a-\theta_b)$ and $\zeta=(\zeta_a,\zeta_b,\zeta_c)$ are interpreted as probability distributions.*

**Theorem A.3** (Dvijotham et al. (2020))**.** *Let $\theta_a,\theta_b,\mu,\mathcal{D},S$ be as in Definition A.1. Define $\tau_\beta\coloneqq\sup_{\nu\in\mathcal{D}}D_{\mathrm{HS},\beta}(\nu\|\mu)$ for all $\beta\geq 0$ and let $\beta_a^\star,\beta_b^\star$ be defined as*

$$\beta_a^\star,\beta_b^\star = \operatorname*{arg\,max}_{\beta_a\geq\beta_b\geq 0} \quad 1-\beta_a(1-\theta_a)-\beta_b\theta_b-(\tau_{\beta_a}+[1-\beta_a]_+)-(\tau_{\beta_b}+[1-\beta_b]_+). \tag{7}$$

*Define the constraint set*

$$\mathcal{D}_{\mathrm{HS}} \coloneqq \left\{ \nu\in\mathcal{P}(\mathcal{X}) \;\middle|\; \begin{array}{l} D_{\mathrm{HS},\beta_a^\star}(\nu\|\mu)\leq\tau_{\beta_a^\star} \\ D_{\mathrm{HS},\beta_b^\star}(\nu\|\mu)\leq\tau_{\beta_b^\star} \end{array} \right\}. \tag{8}$$

*Then, $S$ is information-limited robustly certified at $\mu$ with respect to $\mathcal{D}$* if and only if *$S$ is information-limited robustly certified at $\mu$ with respect to $\mathcal{D}_{\mathrm{HS}}$. Thus, the optimal information-limited certificate in this case can be obtained by applying the information-limited certificate of Theorem A.2 to $\mathcal{D}_{\mathrm{HS}}$.*

$$\tau_\beta^\star \coloneqq \sup_{w'\in\mathcal{B}_\rho(w)} D_{\mathrm{HS},\beta}(\mathcal{M}(w')\,\|\,\mathcal{M}(w)), \qquad \beta\geq 0 \tag{9}$$

## B Background on $f$-divergences

In this section we formally define $f$-divergences and state some useful facts for computing with them. We state here only the most essential results for understanding our derivations; a more complete treatment can be found in, e.g., Csiszár & Shields (2004); Liese & Vajda (2006) or Polyanskiy & Wu (2025, ch. 7). Additionally, we state and prove a few additional useful lemmata.

**Definition B.1** ($f$-divergence)**.** *Let $f\colon\mathbb{R}^+\to\mathbb{R}$ be convex with $f(1)=0$, and take by convention*

$$f(0)\coloneqq\lim_{x\downarrow 0}f(x), \qquad f'(\infty)\coloneqq\lim_{x\downarrow 0}xf(1/x), \qquad 0\cdot f'(\infty)\coloneqq 0.$$

*Let $\mu,\nu$ be two probability measures on a measurable space $(\Omega,\mathcal{A})$. If $\mu\ll\nu$ then the* $f$-divergence from $\nu$ to $\mu$ *is defined as*

$$D_f(\mu\|\nu)\coloneqq\mathbb{E}_\nu\left[f\left(\tfrac{d\mu}{d\nu}\right)\right] \tag{10}$$

*where $\frac{d\mu}{d\nu}$ is a Radon–Nikodym derivative. More generally, let $\lambda$ be any common dominating measure. Then, writing $p\coloneqq\frac{d\mu}{d\lambda}$ and $q\coloneqq\frac{d\nu}{d\lambda}$, we may define*

$$D_f(\mu\|\nu)\coloneqq\int_{q>0}q(x)f\left(\frac{p(x)}{q(x)}\right)d\lambda+f'(\infty)\mu(q=0). \tag{11}$$

*Equation* (11) *does not depend on the choice of common dominating measure $\lambda$, and moreover when $\mu\ll\nu$ gives the same result as Equation* (10)*. We call the function $f$ the* generator *of the $f$-divergence $D_f$.*

Table 4: $f$-divergences and associated generators and integral form, with $p \coloneqq \frac{d\mu}{d\lambda}$ and $q \coloneqq \frac{d\nu}{d\lambda}$.

| $f$-Divergence | Regime | Generator $f(u)$ | Integral form |
|---|---|---|---|
| Total variation $D_{\mathrm{TV}}(\mu\|\nu)$ | | $\frac{1}{2}\lvert u-1\rvert$ | $\frac{1}{2}\int \lvert p-q\rvert\, d\lambda$ |
| Hockey-stick $D_{\mathrm{HS},\beta}(\mu\|\nu)$ | $\beta \geq 0$ | $(u-\beta)_+ - (1-\beta)_+$ | $\int (p-\beta q)_+\, d\lambda - (1-\beta)_+$ |
| Kullback–Leibler $D_{\mathrm{KL}}(\mu\|\nu)$ | | $u \log u$ | $\int p \log \frac{p}{q}\, d\lambda$ |
| Squared Hellinger $D_{\mathrm{H}^2}(\mu\|\nu)$ | | $\frac{1}{2}(\sqrt{u}-1)^2$ | $1 - \int \sqrt{pq}\, d\lambda$ |
| Hellinger $D_{\mathrm{H},\alpha}(\mu\|\nu)$ | $\alpha \in (0,1)\cup(1,\infty)$ | $\frac{u^\alpha - 1}{\alpha - 1}$ | $\frac{1}{\alpha-1}(\int p^\alpha q^{1-\alpha}\, d\lambda - 1)$ |
| Pearson $D_{\chi^2_{\mathrm{P}}}(\mu\|\nu)$ | | $(u-1)^2$ | $\int \frac{(p-q)^2}{q}\, d\lambda$ |
| Neyman $D_{\chi^2_{\mathrm{N}}}(\mu\|\nu)$ | | $\frac{(u-1)^2}{u}$ | $\int \frac{(p-q)^2}{p}\, d\lambda$ |

**Unnormalized hockey-stick divergences** We write

$$\widetilde{D}_{\mathrm{HS},\beta}(\mu\|\nu) \coloneqq \int (p-\beta q)_+\, d\lambda = D_{\mathrm{HS},\beta}(\mu\|\nu) + (1-\beta)_+ = \sup_A \mu(A) - \beta\,\nu(A). \tag{12}$$

**$f$-divergences from likelihood ratios** If we define the likelihood and log-likelihood ratios

$$\Lambda \coloneqq p/q \in [0,\infty], \qquad L \coloneqq \log \Lambda \in [-\infty, +\infty], \tag{13}$$

we can rewrite Equation (11) in the form

$$D_f(\mu\|\nu) = \mathbb{E}_\nu[f(\Lambda)] + f'(\infty)\mu(\Lambda = 0). \tag{14}$$

**Rényi divergences and the Hellinger integral** The generator $f_\alpha(x) \coloneqq \frac{x^\alpha - 1}{\alpha - 1}$ is convex for every $\alpha \in (0,1)\cup(1,\infty)$, and defines the *Hellinger divergence of order* $\alpha$, $D_{\mathrm{H},\alpha}(\mu\|\nu) \coloneqq D_{f_\alpha}(\mu\|\nu)$. The *Hellinger integral of order* $\alpha$ is

$$Z_\alpha(\mu\|\nu) \coloneqq \int p^\alpha q^{1-\alpha}\, d\lambda = 1 + (\alpha-1)\, D_{\mathrm{H},\alpha}(\mu\|\nu), \tag{15}$$

with the convention that $Z_\alpha = \infty$ if $\alpha \geq 1$ and $\mu \not\ll \nu$. The *Rényi divergence of order* $\alpha$ is related via the monotone transformation:

$$D_{\mathrm{R},\alpha}(\mu\|\nu) \coloneqq \frac{1}{\alpha-1}\log Z_\alpha(\mu\|\nu) = \frac{1}{\alpha-1}\log\left(1 + (\alpha-1)D_{\mathrm{H},\alpha}(\mu\|\nu)\right). \tag{16}$$

Note $Z_\alpha$ is *not*, strictly speaking, an $f$-divergence, since $x \mapsto x^\alpha$ does not vanish at 1. Writing $L \coloneqq \log \frac{d\mu}{d\nu}$ for the log-likelihood ratio, $Z_\alpha(\mu\|\nu) = \mathbb{E}_\mu\left[e^{(\alpha-1)L}\right]$, which is the moment generating function of $Z$ under $\mu$ evaluated at $\alpha - 1$. By taking limits, one may also define the Rényi divergences at orders $\alpha \in \{0, 1, \infty\}$:

$$\begin{aligned} D_{\mathrm{R},0}(\mu\|\nu) &\coloneqq -\log \nu(\{p > 0\}), \quad p = d\mu/d\lambda, \\ D_{\mathrm{R},1}(\mu\|\nu) &\coloneqq D_{\mathrm{KL}}(\mu\|\nu), \\ D_{\mathrm{R},\infty}(\mu\|\nu) &\coloneqq \log \operatorname{ess\,sup}_\nu \tfrac{d\mu}{d\nu}. \end{aligned}$$

In general, it is more convenient to state the associated Rényi divergence $D_{\mathrm{R},\alpha}$ rather than the raw Hellinger divergence $D_{\mathrm{H},\alpha}$, due to the nice tensorization and continuity properties of $D_{\mathrm{R},\alpha}$ (see Lemma B.7).

**Special cases** The most useful $f$-divergences are the hockey stick divergences (due to their association with tight certificates via Theorem A.3) and the Rényi divergences (due to their attractive computational properties; see, e.g., Lemma B.7). Moreover, several of the divergences in Table 4 can be expressed in terms of a hockey stick or Rényi divergence of an appropriate order. We have the following relationships: $D_{\mathrm{TV}} = D_{\mathrm{HS},1}$, $D_{\mathrm{H},2} = D_{\chi^2_{\mathrm{P}}}$, $D_{\mathrm{H}^2} = \frac{1}{2}D_{\mathrm{H},1/2} = 1 - Z_{1/2}$.

**Practical considerations** Practically speaking, the relationships between divergences in Table 4 mean we should utilize the Hockey-stick divergences directly whenever possible, and otherwise simply compute the Rényi divergences of all orders (which is useful when tight Hockey-stick divergences are inaccessible). Other specific divergences, like $D_{\mathrm{H}^2}$ and $D_{\chi^2_{\mathrm{P}}}$, restate information already encoded in the Rényi divergences of the appropriate orders.

**Definition B.2** (Perspective function)**.** *Let $f$ be a generator as in Definition B.1 with the usual conventions for $f(0)$ and $f'(\infty)$. The* perspective *of $f$ is the function $g_f\colon \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R} \cup \{+\infty\}$ given by*

$$g_f(a,b) \coloneqq \begin{cases} bf(a/b) & b > 0, \\ af'(\infty) & b = 0, a > 0, \\ 0 & a = b = 0. \end{cases} \tag{17}$$

**Lemma B.1** (Properties of perspective functions)**.** *$g_f$ is jointly convex and positively homogeneous, i.e., $g_f(ta, tb) = tg_f(a,b)$ for every $t > 0$. Consequently $g_f$ is subadditive:*

$$g_f(a_1 + a_2, b_1 + b_2) \le g_f(a_1, b_1) + g_f(a_2, b_2). \tag{18}$$

*Moreover $g_f$ admits the representation*

$$g_f(a,b) = \sup_{s \in \operatorname{dom} f^*} [sa - f^*(s)b] \tag{19}$$

*where $f^*$ is the convex conjugate of $f$, and for $b > 0$ the supremum is attained exactly at those $s \in \partial f(a/b)$.*

**Definition B.3** (Convex Conjugate)**.** *For a function $f\colon \mathbb{R} \to \mathbb{R}$, its* convex conjugate *is defined by $f^*(u) \coloneqq \sup_v (uv - f(v))$.*

**Lemma B.2** (Basic properties of $f$-divergences)**.** *Let $\mu, \nu, \rho$ be probability measures and $f$ a generator as in Definition B.1.*

*(i)* (Nonnegativity.) *$D_f(\mu\|\nu) \ge 0$. If $f$ is strictly convex at 1, then $D_f(\mu\|\nu) = 0$ if and only if $\mu = \nu$.*

*(ii)* (Reversal.) *With the conjugate generator $\overline{f}(x) \coloneqq xf(1/x)$, one has $D_{\overline{f}}(\mu\|\nu) = D_f(\nu\|\mu)$.*

*(iii)* (Cancellation of common parts.) *$D_f(\mu \otimes \rho\|\nu \otimes \rho) = D_f(\mu\|\nu)$.*

*(iv)* (Joint convexity.) *Let $0 < \lambda < 1$ and suppose $\mu = \lambda\mu_0 + (1-\lambda)\mu_1$ and $\nu = \lambda\nu_0 + (1-\lambda)\nu_1$ for probability measures $\mu_i, \nu_i$. Then $D_f(\mu\|\nu) \le \lambda D_f(\mu_0\|\nu_0) + (1-\lambda) D_f(\mu_1\|\nu_1)$.*

*(v)* (Perspective representation.) *$D_f(\mu\|\nu) = \int g_f(p,q)\, d\lambda$ for common dominating $\lambda$.*

*(vi)* (Variational representation.) *With $f^*$ the convex conjugate of $f$,*

$$D_f(\mu\|\nu) = \sup_{g\colon \Omega \to \operatorname{dom}(f^*)} \left\{ \mathop{\mathbb{E}}_{\mu}[g] - \mathop{\mathbb{E}}_{\nu}[f^* \circ g] \right\}.$$

**Lemma B.3** (Data-processing inequality)**.** *Let $K$ be a Markov kernel from $(\Omega, \mathcal{A})$ to $(\Omega', \mathcal{A}')$, and write $K\mu$ for the pushforward $A \mapsto \int K(\omega, A)\, \mu(d\omega)$. Then for every generator $f$,*

$$D_f(K\mu\|K\nu) \le D_f(\mu\|\nu). \tag{20}$$

*Equality holds whenever $K$ is induced by a bimeasurable bijection $T\colon \Omega \to \Omega'$, so that $D_f(T_\#\mu\|T_\#\nu) = D_f(\mu\|\nu)$, and more generally whenever $K$ is a sufficient statistic for $\{\mu, \nu\}$.*

**Lemma B.4** (Reversal of Hockey-Stick divergences)**.** *Fix $\mu, \nu$ probability measures on $(\Omega, \mathcal{A})$, and $\beta > 0$. Define the dual parameter $\overline{\beta} \coloneqq 1/\beta$, extended to $\beta \in \{0, \infty\}$ in the usual way. Then one has the following relationship:*

$$\widetilde{D}_{\mathrm{HS},\beta}(\mu\|\nu) = \beta\, \widetilde{D}_{\mathrm{HS},\bar{\beta}}(\nu\|\mu) + 1 - \beta, \qquad D_{\mathrm{HS},\beta}(\mu\|\nu) = \beta\, D_{\mathrm{HS},\bar{\beta}}(\nu\|\mu). \tag{21}$$

*In particular, the TV distance $D_{\mathrm{TV}}$ is symmetric since it corresponds to the self-dual $\beta = 1$. Moreover, the Hockey-Stick profile for $(\mu, \nu)$ over $0 \le \beta \le 1$ is entirely determined by the profile of the reversed pair $(\nu, \mu)$ for $\beta \ge 1$.*

**Lemma B.5** (Integral representation by Hockey-stick divergences)**.** *Let $f$ be a generator as in Definition B.1, finite on $(0,\infty)$, and let $\rho$ denote the Lebesgue-Stieltjes measure of its right derivative $f'_+$, that is, the unique Borel measure on $(0,\infty)$ with $\rho((a,b]) = f'_+(b) - f'_+(a)$ for $0 < a < b < \infty$. Then for all $\mu \ll \nu$,*

$$D_f(\mu\|\nu) = \int_{(0,\infty)} D_{\mathrm{HS},\beta}(\mu\|\nu)\,\rho(d\beta). \tag{22}$$

**Lemma B.6** (Coupling characterization of TV divergence)**.** *Let $\mu$ and $\nu$ be probability measures on $(\Omega,\mathcal{A})$. Then*

1. *For any coupling $(X,Y)$ of $\mu$ and $\nu$ (that is, $X \sim \mu$, $Y \sim \nu$, but $X$ and $Y$ may not be independent), one has that $D_{\mathrm{TV}}(\mu\|\nu) \le \mathbb{P}[X \neq Y]$.*
2. *There exists a maximal coupling $(X^\star, Y^\star)$ of $\mu$ and $\nu$ such that $D_{\mathrm{TV}}(\mu\|\nu) = \mathbb{P}[X^\star \neq Y^\star]$.*

**Lemma B.7** (Tensorization Properties for $f$-Divergences)**.** *For product measures $\mu = \bigotimes_{i=1}^n \mu_i$ and $\nu = \bigotimes_{i=1}^n \nu_i$ on the same product space, $Z_\alpha(\mu\|\nu) = \prod_{i=1}^n Z_\alpha(\mu_i\|\nu_i)$ for every $\alpha \in (0,1)\cup(1,\infty)$. Furthermore, $D_{\mathrm{R},\alpha}(\mu\|\nu) = \sum_{i=1}^n D_{\mathrm{R},\alpha}(\mu_i\|\nu_i)$.*

Table 5: Tensorization of $f$-divergences under product measures. Throughout, $\mu = \bigotimes_1^n \mu_i$ and $\nu = \bigotimes_{i=1}^n \nu_i$ are product measures on a common product space.

| Divergence | | Tensorization | Notes |
|---|---|---|---|
| Hellinger integral | $Z_\alpha(\mu\|\nu)$ | $= \prod_{i=1}^n Z_\alpha(\mu_i\|\nu_i)$ | $\alpha \in (0,1)\cup(1,\infty)$ |
| Rényi | $D_{\mathrm{R},\alpha}(\mu\|\nu)$ | $= \sum_{i=1}^n D_{\mathrm{R},\alpha}(\mu_i\|\nu_i)$ | $\alpha \in [0,\infty]$ |
| Kullback–Leibler | $D_{\mathrm{KL}}(\mu\|\nu)$ | $= \sum_{i=1}^n D_{\mathrm{KL}}(\mu_i\|\nu_i)$ | case $\alpha = 1$ above |
| Squared Hellinger | $D_{\mathrm{H}^2}(\mu\|\nu)$ | $= 1 - \prod_{i=1}^n \big(1 - D_{\mathrm{H}^2}(\mu_i\|\nu_i)\big)$ | case $\alpha = 1/2$ above |
| Pearson $\chi^2$ | $D_{\chi^2_{\mathrm{P}}}(\mu\|\nu)$ | $= \prod_{i=1}^n \big(1 + D_{\chi^2_{\mathrm{P}}}(\mu_i\|\nu_i)\big) - 1$ | case $\alpha = 2$ above |
| Neyman $\chi^2$ | $D_{\chi^2_{\mathrm{N}}}(\mu\|\nu)$ | $= \prod_{i=1}^n \big(1 + D_{\chi^2_{\mathrm{N}}}(\mu_i\|\nu_i)\big) - 1$ | Lemma B.2(ii) w/ $D_{\chi^2_{\mathrm{P}}}$ |
| Hockey-stick | $D_{\mathrm{HS},\beta}(\mu\|\nu)$ | $\le 1 - \prod_{i=1}^n \big(1 - D_{\mathrm{HS},\beta_i}(\mu_i\|\nu_i)\big)$ | $\beta = \prod_i \beta_i$, $\beta_i \ge 1$ |
| Total variation | $D_{\mathrm{TV}}(\mu\|\nu)$ | $\le 1 - \prod_{i=1}^n \big(1 - D_{\mathrm{TV}}(\mu_i\|\nu_i)\big)$ | |

As a result of Lemma B.7, the divergences of Table 4 obtained by manipulating a Rényi-type divergence themselves possess nice tensorization properties. We document these and more tensorization properties in Table 5. In addition, one has upper bounds on product measures for Hockey-stick and TV divergences, although these bounds are not tight in general.

### B.1 Additional lemmata

In this section we state and prove some additional useful lemmata regarding various $f$-divergences.

**Lemma B.8** ($f$-divergence between measures with shared components)**.** *Let $\mu,\nu$ be probability measures on $(\Omega,\mathcal{A})$, let $\lambda$ be any dominating measure, and write $p := d\mu/d\lambda$ and $q := d\nu/d\lambda$. Define the overlap $S := \{p > 0\} \cap \{q > 0\}$ (up to a $\lambda$-null set). Suppose $\mu(E) = \nu(E)$ for every measurable $E \subseteq S$. Then, denoting $m := \mu(S^c) = \nu(S^c)$, one has $m = D_{\mathrm{TV}}(\mu\|\nu)$ and moreover for every generator $f$ as in Definition B.1,*

$$D_f(\mu\|\nu) = (f(0) + f'(\infty))m = (f(0) + f'(\infty))D_{\mathrm{TV}}(\mu\|\nu).$$

*Proof.* Taking $E = S$ gives $\mu(S) = \nu(S)$ and hence $\mu(S^c) = \nu(S^c)$. Set

$$S_\mu := \{p > 0, q = 0\}, \qquad S_\nu := \{p = 0, q > 0\}, \qquad N := \{p = q = 0\}$$

so that $\Omega = S \sqcup S_\mu \sqcup S_\nu \sqcup N$ with $\mu(N) = \nu(N) = 0$. Evaluating the divergence via Lemma B.2(v) gives

$$\begin{aligned} D_f(\mu\|\nu) &= \int_S g_f(p,q)\,d\lambda + \int_{S_\mu} g_f(p,0)\,d\lambda + \int_{S_\nu} g_f(0,q)\,d\lambda + \int_N g_f(0,0)\,d\lambda \\ &= \int_S qf(1)\,d\lambda + \int_{S_\mu} pf'(\infty)\,d\lambda + \int_{S_\nu} qf(0)\,d\lambda + 0 \\ &= f'(\infty)\mu(S_\mu) + f(0)\nu(S_\nu). \end{aligned}$$

To conclude, note $\mu(S_\mu) = \nu(S_\nu) = m$, and take $f(u) = \frac{1}{2}|u-1|$ for which $f(0) = f'(\infty) = \frac{1}{2}$. □

**Corollary B.1** (Hockey-stick envelope for measures with common component)**.** *In the setting of Lemma B.8, for every $\beta \geq 0$,*

$$D_{\mathrm{HS},\beta}(\mu\|\nu) = \min(1,\beta)m,$$

*where $m = \mu(S^c) = \nu(S^c)$. In particular, $D_{\mathrm{HS},\beta}(\mu\|\nu) = D_{\mathrm{TV}}(\mu\|\nu) = m$ whenever $\beta \geq 1$, independently of $\beta$.*

*Proof.* The generator $f_\beta(u) = (u-\beta)_+ - (1-\beta)_+$ has $f_\beta(0) = -[1-\beta]_+$ and $f'_\beta(\infty) = \lim_{x\downarrow 0} x f_\beta(1/x) = 1$, so $f_\beta(0) + f'_\beta(\infty) = 1 - [1-\beta]_+ = \min(1,\beta)$. □

The significance of Lemma B.8 is that for distributions exhibiting "simple" likelihood ratio behavior, the TV distance encodes the entire hockey stick envelope (indeed, *any* $f$-divergence), and in particular this means via Theorem A.3 that the information-limited certificate is strictly a function of the TV distance between those distributions.

# C Detailed Comparison with Prior Work

## C.1 Threat model and mechanism comparison

Table 6 expresses the threat model and smoothing mechanism of several prior certified poisoning defenses in our notation and describes whether our framework recovers their certificate.

**RAB** RAB bounds a single aggregate $\ell_2$ perturbation applied to the entire training dataset. Strictly speaking, our threat model of Section 3.2 does not capture theirs, since in RAB an adversary may spread their corruption over all training examples. Given our threat model with $m_{\mathrm{p}}$ corrupted examples and $\ell_2$ budget $\varepsilon_{\mathrm{f}}$, an outer RAB approximation may be obtained by taking the aggregate bound $\varepsilon_{\mathrm{f}}\sqrt{m_{\mathrm{p}}}$. In the case of backdoor attacks we can arrive at a similar bound. This is how we use their bound in Section 5. Under their own threat model, our certification procedure still recovers their certificate if we compute the divergences directly: the noise induces a single Gaussian pair (Definition G.3) at distance $\|\Delta\|_2/\sigma$ with closed form hockey-stick divergence (Theorem G.2). This certificate reduces to the closed form of Cohen et al. (2019).

**Bagflip** BagFlip's adversary is mostly recoverable under our threat modeling. The only nuance is in their treatment of label-flipping adversaries. First, they require that the finite label space and feature space coincide, $\mathcal{X} = \mathcal{Y} = [K]$. Then, they treat the label as an extra feature and reduce certification to the feature-only case. In particular, a single budget is shared between the features and label. In this case, we can technically recover their threat model by taking $\rho = (m_{\mathrm{p}}, s, 0, \varepsilon_{\mathrm{te}})$ on the "merged" feature-label space.

If we start with our label-flipping threat model, $\rho = (m_{\mathrm{p}}, \varepsilon_{\mathrm{f}}, 1, \varepsilon_{\mathrm{te}})$, we might ask how the resulting certificate compares to BagFlip's (which would take $s = \varepsilon_{\mathrm{f}} + 1$). First, observe that $\rho$ is more restrictive in the sense that our adversary *must* place one of its corruptions on the label, or not at all, whereas BagFlip's adversary could reallocate a label corruption into an extra feature corruption. Despite this, it turns out that the resulting certificates are identical. Since features and labels share the alphabet $[K]$ and the flipping kernel, the divergences depend only on the number of changed coordinates (Theorem G.4), and both worst cases corrupt $\varepsilon_{\mathrm{f}} + 1$ coordinates.

This simplification fails if $\mathcal{X} \neq \mathcal{Y}$ or if the feature and label noises differ (for example, to enable stronger noise on features than labels), in which case our framework can still compute tight certificates and is notably more general than Zhang et al. (2022).

Table 6: Threat models and smoothing mechanisms of prior randomized smoothing works expressed in the notation of Sections 3 and 4

| Work | Threat vector $\rho$ | Mechanism $(\Pi, \phi_{\mathcal{X}}, \phi_{\mathcal{Y}})$ | Certificate |
|---|---|---|---|
| Jia et al. (2021) | $(m_{\mathrm{p}}, \infty, 1, 0)$ | bagging, none, none | identical (Theorem E.2) |
| Rosenfeld et al. (2020) | $(m_{\mathrm{p}}, 0, 1, 0)$ | full set, none, flipping | tight[a] |
| Zhang et al. (2022) | $(m_{\mathrm{p}}, s, \varepsilon_{\mathrm{l}}, \varepsilon_{\mathrm{te}})$ | bagging, flipping, flipping | identical (Theorem E.3) |
| Weber et al. (2023) | aggregate $\ell_2$ on $S$ | full set, Gaussian, none | identical[b] |

[a] We can obtain provably tight certificates, but must estimate the vote probabilities with an ensemble. Rosenfeld et al. (2020) give an efficient, loose closed form for learners trained with linear regression.
[b] Under their threat model, the certificate reduces to the standard Gaussian certificate of Cohen et al. (2019), which is already understood in Dvijotham et al. (2020).

### C.2 Comparison with Saxena et al. (2026)

Saxena et al. (2026) develop a randomized smoothing-based backdoor defense from the DP perspective. They view the joint train-inference procedure as a randomized mechanism composed of a stack of two private algorithms.

We note several differences between our works. We do not make any assumptions about properties of the learner, like privacy, instead placing assumptions on, and then noising, the input data. It is these data noising schemes we analyze in this work, using techniques to handle subsampling, noise products, and then the individual noise factors. In contrast, the analysis of Saxena et al. (2026) is placed on understanding the composition of DP mechanisms. Our threat modeling is more expressive in its per-example budgets for replacement adversaries, handling arbitrary $\ell_p$ constraints and label corruptions, whereas Saxena et al. (2026) focuses primarily on symmetric set distance adversaries (corresponding to the standard adjacency relation notion in DP), with one notable exception being a special case for $\ell_2$ feature adversaries and Gaussian noise.

One interesting connection between our work and the methods of Saxena et al. (2026) is in the use of the hockey-stick profile, also called the privacy profile in the DP literature, to characterize the robustness condition. Whereas Saxena et al. (2026) use the language and tools of DP, we formulate our results here in the language of $f$-divergences and convex optimization.

To demonstrate the connection, we derive one of the results of Saxena et al. (2026) in our language. Saxena et al. (2026) consider a "randomized DPA" mechanism (motivated by (Levine & Feizi, 2021)) which works as follows. DPA assigns each training record to one of $N$ buckets by some fixed rule (e.g., index or hashing) and trains one model per bucket. In the randomized view, the mechanism samples one bucket uniformly at random and trains a model on it. In our notation, this means taking $\Pi$ to be uniform over $N$ disjoint buckets with no feature or label noise, and the test input to be noised by kernel $\phi_{\mathcal{X}}$.

**Theorem C.1** (DPA with test-time noise; Theorem D.4 of Saxena et al. (2026))**.** *Let $\Pi$ be uniform over pairwise disjoint index buckets $U_1, \ldots, U_N$. Let noise $\phi_{\mathcal{X}}$ be applied only to the test sample. Suppose the corruption changes records in at most $r$ buckets, and let $\tau_\beta^{\mathrm{te}} \geq D_{\mathrm{HS},\beta}(\phi_{\mathcal{X}}(x'_{\mathrm{te}}) \| x_{\mathrm{te}})$ for all admissible test corruptions. Then, for every $\beta > 0$,*

$$D_{\mathrm{HS},\beta}(\mathcal{M}(w') \| \mathcal{M}(w)) \leq \tfrac{r}{N} \min(1, \beta) + (1 - \tfrac{r}{N}) \tau_\beta^{\mathrm{te}} \tag{23}$$

*Equality holds when exactly $r$ buckets are changed, the $N$ buckets of $S$ are pairwise distinct sub-datasets, and the test change attains $\tau_\beta^{\mathrm{te}}$.*

*Proof.* Apply Theorem 4.1 with two groups corresponding to corrupted and uncorrupted buckets. For an uncorrupted bucket, the component divergence reduces to divergence between test components $\tau_\beta^{\mathrm{te}}$ (Lemma B.2(iii)). For a corrupted bucket, the component pair is singular and its hockey stick is simply $\min(1, \beta)$. Combining these terms with the mixing weights $\frac{r}{N}$ and $1 - \frac{r}{N}$ gives the bound.

If the equality conditions hold, the component supports are disjoint, and the bound is attained by Corollary E.1. □

The advantage of the randomized DPA mechanism is that, because the partition size $N$ is small, and because no feature or label noise is added to training data, the training can be done once for all buckets and re-used for arbitrarily many certifications (inference on a sampled model must still be performed). Another interesting observation is that DPA can be used to obtain bounds for symmetric set difference adversaries who can add or remove training samples. For example, if bucket assignment is performed by hashing the training example (as done in Levine & Feizi (2021)), one has $r \leq \min(2m_{\mathrm{p}}, N)$.

# D Deferred proofs

## D.1 Equivalence of robustness conditions

**Theorem D.1** (Robustness characterization)**.** *Fix base learner $G$, smoothing mechanism $\mathcal{M}$, input $w \in \mathcal{W}$, and threat vector $\rho$. Write $c \coloneqq \overline{g_{\mathcal{M}}(w)}$. For each $c' \in \mathcal{Y}$, define the specification $\psi_{c,c'} : \widetilde{W} \to \mathbb{R}$ by $\psi_{c,c'}(\widetilde{w}) \coloneqq G(\widetilde{w})_c - G(\widetilde{w})_{c'}$. Then the following are equivalent:*

1. *$c \in \arg\max_{y \in \mathcal{Y}} \overline{G}(w')_y$ for every $w' \in \mathcal{B}_\rho(w)$.*
2. *For every $c' \neq c$, the specification $\psi_{c,c'}$ is robustly certified at $\mathcal{M}(w)$ with respect to $\mathcal{D}_{w,\rho} \coloneqq \{\mathcal{M}(w') \,:\, w' \in \mathcal{B}_\rho(w)\}$.*

*In particular, either condition implies $\rho$-robustness of $\overline{g_{\mathcal{M}}}$ at $w$ whenever the certified margins are strict.*

*Proof.* Fix $w' \in \mathcal{B}_\rho(w)$ and $c, c' \in \mathcal{Y}$. By linearity of expectation together with the definition of the smoothed classifier,

$$\begin{aligned}\mathop{\mathbb{E}}_{\widetilde{w}\sim\mathcal{M}(w')}[\psi_{c,c'}(\widetilde{w})] &= \mathop{\mathbb{E}}_{\widetilde{w}\sim\mathcal{M}(w')}[G(w)_c - G(w)_{c'}] \\ &= \mathop{\mathbb{E}}_{\widetilde{w}\sim\mathcal{M}(w')}[G(w)_c] - \mathop{\mathbb{E}}_{\widetilde{w}\sim\mathcal{M}(w')}[G(w)_{c'}] \\ &= \overline{G_{\mathcal{M}}}(w')_c - \overline{G_{\mathcal{M}}}(w')_{c'}. \end{aligned} \tag{24}$$

Suppose (1) holds; that is, for every $w' \in \mathcal{B}_\rho(w)$, we have that $c \in \arg\max_{y\in\mathcal{Y}}$. Then for every $c' \neq c$ we have $\overline{G_{\mathcal{M}}}(w')_c \geq \overline{G_{\mathcal{M}}}(w')_{c'}$, so by Equation (24) $\mathbb{E}_{\widetilde{w}\sim\mathcal{M}(w')}[\psi_{c,c'}] \geq 0$. Since every element of $\mathcal{D}_{w,\rho}$ is of the form $\mathcal{M}(w')$ for some $w' \in \mathcal{B}_\rho(w)$, it follows that each $\psi_{c,c'}$ is robustly certified at $\mathcal{M}(w)$ with respect to $\mathcal{D}_{w,\rho}$.

On the other hand, suppose (1) does not hold. Then there exists some $w' \in \mathcal{B}_\rho(w)$ and $c' \neq c$ with $\overline{G_{\mathcal{M}}}(w')'_c > \overline{G_{\mathcal{M}}}(w')_c$. Then condition Equation (24) gives $\mathbb{E}_{\widetilde{w}\sum\mathcal{M}(w')}[\psi_{c,c'}] < 0$, so that $\psi_{c,c'}$ is not robustly certified in the sense of Definition A.1 and condition (2) fails.

Finally, if the certified margins are strict, i.e., $\overline{G_{\mathcal{M}}}(w')_c > \overline{G_{\mathcal{M}}}(w')_{c'}$ for all $c' \neq c$ and all $w' \in \mathcal{B}_\rho(w)$, then $c$ is the unique maximizer of $y \mapsto \overline{G_{\mathcal{M}}}(w')_y$, and we recover robustness in the desired sense of Definition 3.1. □

**Corollary D.1** (Rényi mixture bound)**.** *With the setting in Theorem 4.1, the Rényi divergence $D_{\mathrm{R},\alpha}(\mu\|\nu)$ satisfies the following LogSumExp bound:*

$$D_{\mathrm{R},\alpha}(\mu\|\nu) \leq \left(\tfrac{1}{\alpha-1}\right) \log\left(\sum_{i=1}^{N} \theta_i \exp((\alpha-1)D_{\mathrm{R},\alpha}(\mu_i\|\nu_i))\right) \tag{25}$$

*for all $\alpha \in (0,1) \cup (1,\infty)$. Moreover, given $\beta_i$ and $\delta_i$ in the setting of Theorem 4.1, we have the reorganized bound*

$$D_{\mathrm{R},\alpha}(\mu\|\nu) \leq \left(\tfrac{1}{\alpha-1}\right) \log\left(\sum_{i=1}^{k} \beta_i \exp((\alpha-1)\delta_i)\right). \tag{26}$$

### D.2 Proof of Theorem 4.1

*Proof.* The result follows by applying joint convexity of $D_f$ (Lemma B.2(iv)) and rearranging terms according to the partition:

$$\begin{aligned} D_f(\mu\|\nu) &\leq \sum_{i=1}^{N} \theta_i D_f(\mu_i\|\nu_i) = \sum_{i=1}^{k} \sum_{j\in S_i} \theta_j D_f(\mu_j\|\nu_j) \\ &\leq \sum_{i=1}^{k} \sum_{j\in S_i} \theta_j \delta_i = \sum_{i=1}^{k} \delta_i \sum_{j\in S_i} \theta_j = \sum_{i=1}^{k} \beta_i \delta_i. \end{aligned}$$

□

### D.3 Proof of Corollary D.1

*Proof.* We first note that Equation (26) follows straightforwardly from Equation (25) due to monotonicity of $\log$ and $\exp$.

We prove Equation (25) for the cases $0 < \alpha < 1$ and $\alpha > 1$ separately.

**Case 1:** $0 < \alpha < 1$ The Rényi divergence is related to the $f$-divergence generated by $f(u) = 1 - u^\alpha$ by $D_{\mathrm{R},\alpha}(\mu\|\nu) = (\frac{1}{\alpha-1}) \log(1 - D_f(\mu\|\nu))$, where in particular joint convexity of $D_f$ and monotonicity of $\log((1-\cdot)^{-1})$ give the bound

$$\begin{aligned} D_{\mathrm{R},\alpha}(\mu\|\nu) &\leq \left(\tfrac{1}{1-\alpha}\right) \log\left(\left(1 - \sum_{i=1}^{N} \theta_i D_f(\mu_i\|\nu_i)\right)^{-1}\right) \\ &= \left(\tfrac{1}{\alpha-1}\right) \log\left(1 - \sum_{i=1}^{N} \theta_i D_f(\mu_i\|\nu_i)\right). \end{aligned}$$

Using the fact that $D_f(\mu_i\|\nu_i) = 1 - \exp((\alpha-1)D_{\mathrm{R},\alpha}(\mu_i\|\nu_i))$, we get

$$\begin{aligned} D_{\mathrm{R},\alpha}(\mu\|\nu) &\leq \left(\tfrac{1}{\alpha-1}\right) \log\left(1 - \sum_{i=1}^{N} \theta_i (1 - \exp((\alpha-1)D_{\mathrm{R},\alpha}(\mu_i\|\nu_i)))\right) \\ &= \left(\tfrac{1}{\alpha-1}\right) \log\left(\sum_{i=1}^{N} \theta_i \exp((\alpha-1)D_{\mathrm{R},\alpha}(\mu_i\|\nu_i))\right) \end{aligned}$$

where the final step follows since $\sum_i \theta = 1$.

**Case 2:** $\alpha > 1$ The Rényi divergence is related to the $f$-divergence generated by $f(u) = u^\alpha - 1$ by $D_{\mathrm{R},\alpha}(\mu\|\nu) = (\frac{1}{\alpha-1}) \log(1 + D_f(\mu\|\nu))$, where in particular joint convexity of $D_f$ and monotonicity of $\log(1+\cdot)$ give the bound

$$D_{\mathrm{R},\alpha}(\mu\|\nu) \leq \left(\tfrac{1}{\alpha-1}\right) \log\left(1 + \sum_{i=1}^{N} \theta_i D_f(\mu_i\|\nu_i)\right).$$

Using the fact that $D_f(\mu_i\|\nu_i) = \exp((\alpha-1)D_{\mathrm{R},\alpha}(\mu_i\|\nu_i)) - 1$, we get

$$\begin{aligned} D_{\mathrm{R},\alpha}(\mu\|\nu) &\leq \left(\tfrac{1}{\alpha-1}\right) \log\left(1 + \sum_{i=1}^{N} \theta_i (\exp((\alpha-1)D_{\mathrm{R},\alpha}(\mu_i\|\nu_i)) - 1)\right) \\ &= \left(\tfrac{1}{\alpha-1}\right) \log\left(\sum_{i=1}^{N} \theta_i \exp((\alpha-1)D_{\mathrm{R},\alpha}(\mu_i\|\nu_i))\right) \end{aligned}$$

where the final step follows since $\sum_i \theta_i = 1$. □

# E TIGHTNESS OF CERTIFICATES

Since Cohen et al. (2019) proved optimality of their certificate for Randomized Smoothing with Gaussian noise, it has been customary to pair derivation of certified guarantees with an analysis of the tightness of the resulting certificate. In the ideal case, one can prove that a given Randomized Smoothing certificate is *tight* in the following precise sense. Given an input $w \in \mathcal{W}$, the certification procedure determines a robustness region $\mathcal{R}$ at which the input $w$ is robust ($\overline{g}(w) = \overline{g}(w')$ for all $w' \in \mathcal{B}_\rho(w)$ where $\rho \in \mathcal{R}$). We can say that the region $\mathcal{R}$ is *tight* if, for any $\rho \notin \mathcal{R}$, there exists a "worst-case" input $w' \in \mathcal{B}_\rho(w)$ and base algorithm $G$ (subject to some constraints determined by evaluating $G$ at $w$) with the property that $\overline{g(w')} \neq \overline{g(w)}$. In the information-limited setting of Theorem A.2, $G$ is constrained by the aggregated vote probabilities $0 \leq \theta_b \leq \theta_a \leq 1$. In the full-information setting of Theorem A.1, more information about the vote distribution can be used.

Turning to our framework, our analysis introduces slack at two key steps. First, we replace the induced constraint set of $\mathcal{D}_{w,\rho}$ of Theorem D.1 by a convex outer approximation $\mathcal{D}_{\mathcal{F}}$. Second, we replace the exact divergence bounds $\tau_f^\star = \sup_{w' \in \mathcal{B}_\rho(w)} D_f(\mathcal{M}(w') \| \mathcal{M}(w))$ by analytical upper bounds of Theorem 4.1. [3]

The first source of slack corresponds to the convex relaxation itself, and is already reasonably well-understood. By Theorem A.3, if the exact Hockey-stick profile $\tau_\beta^\star$ (see Equation (9)) is available at all orders $\beta \geq 0$, then the resulting certificate is tight in the information-limited setting. This means no procedure that observes only the vote statistics $\theta_a, \theta_b$ can certify a larger region without further assumptions on the base learner $G$.

Thus, to understand the tightness of our certificates, we should understand the error introduced by Theorem 4.1 and how it affects certification. For a pair of mixture distributions $\mu = \sum_i \theta_i \mu_i$ and $\nu = \sum_i \theta_i \nu_i$ on a shared space. Define the quantity

$$E_f(\mu, \nu; \theta) \coloneqq \sum_{i=1}^{N} \theta_i D_f(\mu_i \| \nu_i) - D_f(\mu \| \nu) \;\geq\; 0. \tag{27}$$

In this appendix, we seek to understand the second source of slack arising from our use of the joint convexity bound Theorem 4.1, and how it affects the resulting certificates generated by our framework. This analysis can be broken into several steps. Appendix E.1 characterizes $E_f$ exactly as the information gained by releasing the mixture component labels along with the noised observation. In particular, the joint convexity bound is exact in precisely the cases when the mixture component label can be recovered from the noised observation. Appendices E.2 and E.3 show that the smoothing schemes of two prior works, Jia et al. (2021) and Zhang et al. (2022), satisfy this hypothesis, and thus we recover their certificates exactly.

## E.1 JOINT-CONVEXITY SLACK FROM LABEL RECOVERY

Throughout this subsection fix mixtures $\mu = \sum_{i=1}^N \theta_i \mu_i$ and $\nu = \sum_{i=1}^N$ on $(\Omega, \mathcal{A})$ with common weights $\theta_i > 0$, let $\lambda$ be a common dominating measure for $\{\mu_i, \nu_i\}_{i=1}^N$, and write $p_i \coloneqq d\mu_i/d\lambda$ and $q_i \coloneqq d\nu_i/d\lambda$.

Sampling from $\mu$ is a two-stage process: first, draw index $i \sim \theta$, then, draw $\omega \sim \mu_i$. The mixture $\mu$ reflects the law of $\omega$ marginalized over $i$. For purposes of understanding the slack $E_f$, it is useful to also define the labeled pair $(i, \omega)$:

**Definition E.1** (Labeled mixture)**.** *The* labeled mixtures *of $\mu$ and $\nu$ are the probability measures on $[N] \times \Omega$*

$$\widehat{\mu} \coloneqq \sum_{i=1}^{N} \theta_i (\delta_i \otimes \mu_i) \qquad \widehat{\nu} \coloneqq \sum_{i=1}^{N} \theta_i (\delta_i \otimes \nu_i)$$

*We write $\pi \colon [N] \times \Omega \to \Omega$, $\pi(i, \omega) = \omega$ for the projection onto the observation $\omega$.*

Note that $\pi$ is a deterministic Markov kernel with $\pi_\# \widehat{\mu} = \mu$ and $\pi_\# \widehat{\nu} = \nu$.

[3] To be precise, slackness could also enter through the reorganization process of Theorem 4.1. However, in the cases we consider, this reorganization is exact.

**Lemma E.1** (Additivity of labeled mixtures)**.** *For every generator $f$ as in Definition B.1,*

$$D_f(\widehat{\mu}\|\widehat{\nu}) = \sum_{i=1}^{N} \theta_i D_f(\mu_i\|\nu_i). \tag{28}$$

*Proof.* The measure $\lambda' := \mathrm{count} \otimes \lambda$ dominates $\widehat{\mu}$ and $\widehat{\nu}$, with densities $(i,\omega) \mapsto \theta_i p_i(\omega)$ and $(i,\omega) \mapsto \theta_i q_i(\omega)$. By the perspective representation (Lemma B.2(v)) and positive homogeneity of $g_f$ (Lemma B.1),

$$D_f(\widehat{\mu}\|\widehat{\nu}) = \sum_{i=1}^{N} \int g_f(\theta_i p_i, \theta_i q_i)\, d\lambda = \sum_{i=1}^{N} \theta_i \int g_f(p_i, q_i)\, d\lambda = \sum_{i=1}^{N} \theta_i D_f(\mu_i\|\nu_i). \qquad \square$$

Combining Lemma E.1 with the data-processing inequality (Lemma B.3) applied to $\pi$ recovers Theorem 4.1:

$$D_f(\mu\|\nu) = D_f(\pi_\#\widehat{\mu}\|\pi_\#\widehat{\nu}) \le D_f(\widehat{\mu}\|\widehat{\nu}) = \sum_{i=1}^{N} \theta_i D_f(\mu_i\|\nu_i).$$

In particular,

$$E_f = D_f(\widehat{\mu}\|\widehat{\nu}) - D_f(\mu\|\nu). \tag{29}$$

Equation (29) states that the slack $E_f$ is precisely the difference in $f$-divergence between labeled and unlabeled mixtures. In our setting the index is the retained index sequence $U \sim \Pi$.

**Theorem E.1** (Exact divergence slack)**.** *Let $f$ be a generator as in Definition B.1 with perspective $g_f$, and define the pointwise error*

$$\Delta_f(\omega) := \sum_{i=1}^{N} g_f\left(\theta_i p_i(\omega), \theta_i q_i(\omega)\right) - g_f\left(\textstyle\sum_{i=1}^{N} \theta_i p_i(\omega), \sum_{i=1}^{N} \theta_i q_i(\omega)\right). \tag{30}$$

*Then $\Delta_f(\omega) \ge 0$ pointwise and $E_f = \int \Delta_f(\omega)\, d\lambda$. In particular, $E_f = 0$ if and only if $\Delta_f = 0$ $\lambda$-a.e. Moreover, writing $I(\omega) := \{i : (p_i(\omega), q_i(\omega)) \ne (0,0)\}$:*

1. (Sufficient.) *$\Delta_f(\omega) = 0$ whenever there exists a single $s \in \mathrm{dom}\, f^*$ attaining the supremum in Equation* (19) *simultaneously for every pair $(p_i(\omega), q_i(\omega))$, $i \in I(\omega)$.*
2. (Necessary and sufficient in the regular case.) *If $p_i(\omega), q_i(\omega) > 0$ for every $i \in I(\omega)$, then*

$$\Delta_f(\omega) = 0 \iff \bigcap_{i \in I(\omega)} \partial f\left(\frac{p_i(\omega)}{q_i(\omega)}\right) \ne \emptyset. \tag{31}$$

*Proof.* Nonnegativity of $\Delta_f$ follows from subadditivity of $g_f$ (Lemma B.1). The identity $E_f = \int \Delta_f(\omega)\, d\lambda$ follows from Lemma B.2(v) for the second term and Lemma E.1 (equivalently, positive homogeneity) for the first.

For the remaining claims, Equation (19) exhibits $g_f$ on $\mathbb{R}^2_+$ as the support function of $C := \mathrm{cl}\,\mathrm{conv}\{(s, -f^*(s)) : s \in \mathrm{dom}\, f^*\}$. A support function is additive on a finite family of vectors if and only if some point of $C$ is simultaneously maximizing for all of them, which gives Item 1. For Item 2, if $q_i(\omega) > 0$, then the maximizing set for $(\theta_i p_i(\omega), \theta_i q_i(\omega))$ is $\{(s, -f^*(s)) : s \in \partial f(p_i(\omega)/q_i(\omega))\}$ by Lemma B.1, and since $f$ is finite and convex on $(0,\infty)$ this set is nonempty and compact whenever $p_i(\omega)/q_i(\omega) \in (0,\infty)$. Maximizers therefore exist for each vector and the common-point criterion is exact, giving Equation (31). Components with $i \notin I(\omega)$ can be ignored as $g_f(0,0) = 0$. $\square$

**Corollary E.1** (Exactness for identifiable mixtures)**.** *If the sets $T_i := \{p_i > 0\} \cup \{q_i > 0\}$ are pairwise disjoint up to $\lambda$-null sets, then $|I(\omega)| \le 1$ almost everywhere, so $\Delta_f \equiv 0$ and*

$$D_f(\mu\|\nu) = \sum_{i=1}^{N} \theta_i D_f(\mu_i\|\nu_i) \tag{32}$$

*for* every *generator $f$.*

Corollary E.1 can be interpreted in the following way: if the components have disjoint supports, the projection $\pi$ is invertible $\lambda$-a.e., and so preserves $f$-divergences via the bijective form of Lemma B.3.

### E.2 TIGHTNESS OF PURE SUBSAMPLING

We use these results to recover the bagging certificate of Jia et al. (2021) as a special case of our framework. Take $\phi_{\mathcal{X}}(\cdot \mid x) = \delta_x$ and $\phi_{\mathcal{Y}}(\cdot \mid y) = \delta_y$ as the feature and label smoothing mechanisms (that is, no noise is added to features or labels). Call an input $w = (S, x_{\mathrm{te}})$ and corruption $w' \in \mathcal{B}_\rho(w)$ *record-distinct* if the labeled records of $S$ are pairwise distinct and every corrupted record $(x'_i, y'_i), i \in I(w, w')$, differs from every record of $S$.

**Theorem E.2** (Exactness for pure subsampling)**.** *Let $\mathcal{M}$ subsample without noise and let $w' \in \mathcal{B}_\rho(w)$ be record-distinct. Write $m(w, w') \coloneqq \mathbb{P}_{U\sim\Pi}[\exists j.\ u_j \in I(w, w')]$ for the probability that the subsample retains a corrupted record. Then, for every generator $f$,*

$$D_f(\mathcal{M}(w')\|\mathcal{M}(w)) = \sum_U \Pi(U) D_f\left(\delta_{S'_U}\middle\|\delta_{S_U}\right) = (f(0) + f'(\infty))\ m(w, w').$$

*So Theorem 4.1 is exact, and the Hockey-stick profile is*

$$D_{\mathrm{HS},\beta}(\mathcal{M}(w')\|\mathcal{M}(w)) = \min(1, \beta) m(w, w') \qquad \beta \geq 0. \tag{33}$$

*Proof.* By record-distinctness, the point masses $\delta_{S_U}$ and $\delta_{S'_U}$ are pairwise distinct across $U$ whenever $U$ contains at least one corrupted index. Otherwise, $S_U = S'_U$. The sets $T_U$ of Corollary E.1 are therefore pairwise disjoint and exactness follows. For Equation (33), observe that the mixtures agree on their intersection $\{U : U \cap I = \emptyset\}$, and so we may apply Lemma B.8 with overlap mass $1 - m(w, w')$. □

**Corollary E.2** (Recovery of bagging certificate)**.** *In the setting of Theorem E.2, write $m^\star_\rho \coloneqq \sup_{w'\in\mathcal{B}_\rho(w)} m(w, w')$. Then $\mathcal{G}(\theta_a, \theta_b)$ is robustly certified at $\mathcal{M}(w)$ with respect to $\mathcal{D}_{w,\rho}$ if and only if*

$$\theta_a - \theta_b \geq 2m^\star_\rho. \tag{34}$$

*For $\rho$ constraining only $m_{\mathrm{p}}$, $m^\star_\rho = 1 - (1 - m_{\mathrm{p}}/n)^k$ under size-$k$ sampling with replacement and $m^\star_\rho = 1 - \binom{n-m_{\mathrm{p}}}{k}\binom{n}{k}^{-1}$ without replacement.*

*Proof.* Substituting $\tau_\beta = \min(1, \beta) m^\star_\rho$ into the objective of Theorem A.3, we get $\beta^\star_a = \beta^\star_b = 1$ whenever $\theta_a > m^\star_\rho$ and $\theta_b + m^\star_\rho < 1$, giving optimal value of $\theta_a - \theta_b - 2m^\star_\rho$ in Equation (6). □

A noteworthy condition for our tightness proof was record-distinctness. In the usual randomized smoothing setting (Cohen et al., 2019; Jia et al., 2021; Zhang et al., 2022) this is a mild assumption, as the certificates are derived without using additional information about the input. If record-distinctness does not hold (either due to repetition in $S$ or impossibility of choosing an appropriate $w'$), we cannot directly apply Lemma B.8 to obtain the worst-case $\tau_\beta$ (although it remains a valid upper bound).

### E.3 RECOVERING BAGFLIP'S CERTIFICATE

In Appendix E.2, we exploited Corollary E.1 to obtain a tight certificate based on the TV distance in the case of pure bagging. Once feature or label noise is introduced, the mixture components are no longer identifiable. Thus, in principle, the slack Equation (27) is nontrivial.

However, it turns out that Zhang et al. (2022) compute the certificate only for the *label-released* mixture (see Zhang et al. (2022, Appendix A)), and thus Corollary E.1 does apply to the setting actually considered by BagFlip. This means two things. First, the certificates obtained within our framework match the certificates of Zhang et al. (2022) exactly in the analytical sense. Second, the certificate of Zhang et al. (2022) may not be tight in the sense that base learners that do not use the mixture component label (which is true of BagFlip's learners) are actually more robust than the certificate guarantees.

**Theorem E.3** (Recovery of BagFlip certificate)**.** *Consider the BagFlip instantiation taking $\mathcal{X} = [K]^d$, $\Pi$ uniform on $[n]^k$, and $\phi_{\mathcal{X}}, \phi_{\mathcal{Y}}$ the flipping measures of Definition G.5 with rate $\gamma$. Let*

*$\widehat{\mathcal{M}(w)}, \widehat{\mathcal{M}(w')}$ denote the labeled mixture of $\mathcal{M}$ as in Definition E.1. Then the bound of Theorem 4.1, applied with partition of $[n]^k$ by the number $c$ of retained corrupted examples, evaluates to*

$$\tau_\beta \;=\; D_{\mathrm{HS},\beta}\left(\widehat{\mathcal{M}(w')} \,\middle\|\, \widehat{\mathcal{M}(w)}\right) \;=\; \sum_{c=0}^{k} \mathrm{Binom}(c; k, \tfrac{m_{\mathrm{p}}}{n}) \delta_\beta\big(c(\varepsilon_{\mathrm{f}} + \varepsilon_{\mathrm{l}}) + \varepsilon_{\mathrm{te}}\big), \tag{35}$$

*where $\delta_\beta(s) \coloneqq D_{\mathrm{HS},\beta}(\phi_\gamma(y) \| \phi_\gamma(y'))$ for any $y, y'$ with $\|y - y'\|_0 = s$, given in closed form by Equation* (68)*. Consequently, the certificate obtained by applying Theorem A.3 to the profile given by Equation* (35) *coincides with the radius of Zhang et al. (2022, Thm. 5.1).*

*Proof.* Conditioned on the sampled indices $U$, the pair $(\mu_U, \nu_U)$ consists of two flipping measures on $[K]^{kd} \times [C]^k \times [K]^d$ whose parameters disagree in exactly $c(\varepsilon_{\mathrm{f}} + \varepsilon_{\mathrm{l}}) + \varepsilon_{\mathrm{te}}$ coordinates, where $c = |\{j : u_j \in I(w, w')\}|$. By Theorem G.4 the component divergence is determined as a function of $c$, so the regrouping step preserves the exact joint convexity bound. Thus, we recover the exact Hockey-stick envelope $\tau_\beta,\ \beta \geq 0$, and via Theorem A.3 the optimal information-limited certificate. □

# F Methodology Details

## F.1 Training

**Datasets** We train models on MNIST, CIFAR10, as well as binary classification tasks derived from them. MNIST-17 uses the MNIST digits 1 and 7, and CIFAR10-02 uses the CIFAR10 classes 0 and 2 (airplane and bird). The training and test sets have 60,000 and 10,000 examples on MNIST, 13,007 and 2,163 on MNIST-17, 50,000 and 10,000 on CIFAR10, and 10,000 and 2,000 on CIFAR10-02. We report certified accuracy on the full test set. Pixel values are divided by 255, which puts every feature in $[0, 1]$, and we apply no further normalization or standardization. $d = 784$ on MNIST and $d = 3072$ on CIFAR10.

**Architectures** On MNIST datasets, the base learner is a small convolutional network consisting of two $5 \times 5$ convolutions with 16 and 32 channels, respectively, each followed by ReLU activation and $2 \times 2$ average pooling, then a hidden layer of width 64 with ReLU, and finally a linear output layer. We use dropout of 0.25 and 0.5 before each linear layer. On the CIFAR10 datasets, we use the ResNet-20 of Jia et al. (2021) modified to use group normalization (at most eight groups) and weight-standardized convolutions. These modifications are due to our parallel training implementation which does not support the running statistics of batch normalization. Networks in all experiments are trained from random initializations.

**Optimization** All training uses the Adam optimizer with learning rate $10^{-3}$ and runs for 200 epochs with batch size 16. We use weight decay of $2 \cdot 10^{-3}$ on MNIST and $10^{-4}$ on CIFAR10. On CIFAR10, the learning rate is multiplied by 0.1, 0.01, 0.001, and 0.0005 from epochs 80, 120, 160, and 180, following BagFlip (Zhang et al., 2022). Models trained on the full training set (for the RAB baseline) use 20 epochs with batch size 128 on MNIST and MNIST-17, and 10 epochs with batch size 64 on CIFAR10 and CIFAR10-02.

**Augmentation** On MNIST, we use random translations of up to 2 pixels and $8 \times 8$ cutout. On CIFAR10, we use random translations of up to 4 pixels and horizontal flips. In addition, each training batch receives Gaussian noise whose standard deviation is drawn uniformly from [0, 1], motivated by the augmentation of Zhang et al. (2022), and disabled for plain bagging experiments on CIFAR10 and CIFAR10-02.

## F.2 Mechanisms and configurations

**Mechanisms** All ensembles use bagging with replacement. The feature noise scale $\sigma$ is the per-coordinate standard deviation: Gaussian noise $\mathcal{N}(0, \sigma^2)$, Laplace noise $\mathrm{Lap}(0, \sigma/\sqrt{2})$, and uniform noise $\mathrm{Unif}([-\sqrt{3}\sigma, \sqrt{3}\sigma])$. The feature noise is Gaussian unless stated otherwise. In the joint feature and label experiments, label noise keeps the original label with probability $1 - \gamma + \gamma/C = 0.9$,

with $\gamma = \frac{1}{9}$ on MNIST and CIFAR10 and $\gamma = \frac{1}{5}$ on MNIST-17 and CIFAR10-02. The backdoor experiments use no label noise.

**Ensembles** Every ensemble has $m = 10{,}000$ base models on MNIST and MNIST-17, and $m = 1{,}000$ base models on CIFAR10 and CIFAR10-02.

**$\ell_p$ feature adversaries** The experiments of Section 5.1 use $k = 100$ and $\sigma = 0.5$ on MNIST and MNIST-17, $k = 1000$ and $\sigma = 0.35$ on CIFAR10, and $k = 200$ and $\sigma = 0.35$ on CIFAR10-02. In Figure 1, the Gaussian, Laplace and uniform curves use this configuration, the bagging curve uses $k = 100$ with no noise, and RAB uses $\sigma = 0.5$ on the full training set. Figure 2 uses the configuration of each dataset.

**Label flipping and backdoor adversaries** Table 7 lists the bag sizes and noise scales of Tables 2 and 3. In Table 2, bagging uses the same $k$ as our method. In Table 3, RAB uses the $\sigma_a$ and $\sigma_b$ of our method on the full training set.

Table 7: Bag sizes and noise scales of Table 2 and Table 3.

| | **Joint feature & label** | | | **Backdoor** | | |
|---|---|---|---|---|---|---|
| Dataset | $k$ | $\sigma_a$ | $\sigma_b$ | $k$ | $\sigma_a$ | $\sigma_b$ |
| MNIST | 30 | 0.25 | 0.5 | 300 | 0.75 | 1.0 |
| MNIST-17 | 30 | 0.5 | 1.0 | 100 | 0.75 | 1.0 |
| CIFAR10 | 200 | 0.2 | 0.35 | 1000 | 0.5 | 0.75 |
| CIFAR10-02 | 25 | 0.35 | 0.5 | 100 | 0.35 | 0.5 |

### F.3 Confidence intervals

Certification in our framework requires high-confidence bounds on teh vote probabilities of the smoothed classifier. We estimate them from $N$ votes of an ensemble using the following confidence interval.

**Lemma F.1** (Clopper–Pearson Interval (Clopper & Pearson, 1934))**.** *Let $Z_1, \ldots, Z_N$ be independent, identically distributed Bernoulli random variables with mean $p \in [0, 1]$, and let $S_N \coloneqq \sum_{i=1}^{N} Z_i$. Fix a failure probability $\zeta \in (0, 1)$, and let $Q(q; \alpha, \beta)$ denote the q-quantile of the Beta distribution with shape parameters $\alpha, \beta$. Define*

$$\underline{p} \coloneqq Q\Big(\tfrac{\zeta}{2};\, S,\, N - S + 1\Big)\,, \tag{36}$$

$$\overline{p} \coloneqq Q\Big(1 - \tfrac{\zeta}{2};\, S + 1,\, N - S\Big)\,, \tag{37}$$

*with the conventions $\underline{p} \coloneqq 0$ when $S_N = 0$ and $\overline{p} \coloneqq 1$ when $S_N = N$. Then, with probability at least $1 - \zeta$, $p \in [\underline{p}, \overline{p}]$. In particular, the one-sided upper confidence bound $\overline{p}^{+} \coloneqq Q(1 - \zeta;\, S + 1,\, N - S)$ satisfies $\Pr[\, p \le \overline{p}^{+}\,] \ge 1 - \zeta$.*

More precisely, we use SimuEM as in Jia et al. (2021), which applies a Bonferroni correction over all $C$ classes and the $e$ test points. Each bound thus has failure probability $\alpha/(eC)$ and the full test-set failure probability is bounded by $\alpha$ via the union bound.

### F.4 Algorithms

Algorithms 1 and 2 follow Jia et al. (2021, Algorithm 1) and Weber et al. (2023, Algorithms 1 and 2). We perform certification following Dvijotham et al. (2020, Algorithm 2) with hockey-stick constraints of all orders.

**Algorithm 1** TRAINUNDERMECHANISM

**Input:** Training set $S = ((x_j, y_j))_{j=1}^n$, learning algorithm $\mathcal{A}$, mechanism $(\Pi, \phi_\mathcal{X}, \phi_\mathcal{Y})$, number of models $N$
**Output:** Base models $g_1, \ldots, g_N$
1: **for** $i = 1, \ldots, N$ **do**
2: Sample $U_i = (u_1, \ldots, u_k) \sim \Pi$
3: Sample $\widetilde{S}_i \leftarrow \big((\widetilde{x}_j, \widetilde{y}_j)\big)_{j=1}^k$ with $\widetilde{x}_j \sim \phi_\mathcal{X}(\cdot \mid x_{u_j})$ and $\widetilde{y}_j \sim \phi_\mathcal{Y}(\cdot \mid y_{u_j})$
4: $g_i \leftarrow \mathcal{A}(\widetilde{S}_i)$
5: **end for**
6: **return** $g_1, \ldots, g_N$

**Algorithm 2** CERTIFY

**Input:** Base models $g_1, \ldots, g_N$, test examples $x_1, \ldots, x_e$, feature noise mechanism $\phi_\mathcal{X}$, per-example budgets $(\varepsilon_\mathrm{f}, \varepsilon_\mathrm{l}, \varepsilon_\mathrm{te})$, confidence level $1 - \alpha$
**Output:** Predicted label and certified number of corrupted examples for each test example
1: **for** $x$ in $x_1, \ldots, x_e$ **do**
2: Sample $\widetilde{x}^{(i)} \sim \phi_\mathcal{X}(\cdot \mid x)$ for $i = 1, \ldots, N$ ▷ $\widetilde{x}^{(i)} = x$ if the test input is not noised
3: $\mathrm{counts}[y] \leftarrow \sum_{i=1}^N \mathbb{1}[g_i(\widetilde{x}^{(i)}) = y]$ for $y \in [C]$
4: $c, c' \leftarrow$ top two indices in counts
5: $\theta_a, \theta_b \leftarrow \textsc{SimuEM}(\mathrm{counts}, \alpha/e)$
6: $m^\star \leftarrow \textsc{BinarySearch}(\theta_a, \theta_b)$ ▷ find largest $m_\mathrm{p}$ with $\underline{p} > \overline{p}$
7: Output $c$ and $m^\star$, or ABSTAIN if no $m_\mathrm{p} \geq 0$ is certified
8: **end for**

## G BOUNDING DIVERGENCES

The core technical challenge of our method is to derive the divergence bounds

$$\tau_f(w, \rho) \geq \tau_f^\star(w, \rho) = \sup_{w' \in \mathcal{B}_\rho(w)} D_f(\mathcal{M}(w') \| \mathcal{M}(w)) \tag{38}$$

for various inputs $w$, smoothing mechanisms $\mathcal{M}$, threat models $\rho$, and generators $f$. Via Theorem 4.1, the problem reduces to bounding divergences between the paired components of the mixture induced by $\mathcal{M}$. Fixing mixture component $U$, this decomposes into two steps.

The first step is to argue that the maximum attainable divergence is exhibited by some worst-case perturbation of the subsampled input $x = S_U$, so that we may write

$$\sup_{x' \in \mathcal{B}(x)} D_f(\phi_\# x' \| \phi_\# x) = D_f(\mu^\star \| \nu^\star) \tag{39}$$

for some pair of probability measures $\mu^\star, \nu^\star$, where $\phi$ is the joint feature-and-label noising mechanism. [4] Note that $x$ and $\mathcal{B}(x)$ do not necessarily correspond exactly with the joint input $w$ and constraint set $\mathcal{B}_\rho(w)$; for example, $\mathcal{B}(x)$ likely depends on $U$. For reasons we will describe later, the two main desiderata are that the same pair $\mu^\star, \nu^\star$ works for *every* $f$, and that such pairs compose across the factors of $\phi$. We handle this step in Appendix G.1.

The second step is to compute the full set of divergences for $\mu^\star, \nu^\star$. This is typically straightforward once $\mu^\star, \nu^\star$ are known. We handle this step in Appendix G.2, where we work out closed-form divergences for Uniform, Flipping, Gaussian, and Laplace measures by hand. We also show how to construct product measures using these measures as building blocks, and reduce divergence computation to a one-dimensional numerical computation.

Together, the above steps are sufficient to derive bounds for Uniform, Gaussian, and Flipping measures, as well as product combinations of them, under all ($\ell_p$, $\ell_0$) feature-label constraints. In some cases the procedure does not work so cleanly. Informally speaking, if the smoothing geometry does not mesh with the perturbation geometry, no single pair $\mu^\star, \nu^\star$ need exist. This arises for Laplace feature noise under general $\ell_p$ norms with $p \neq 1$, which we treat separately in Appendix G.4.

[4] This is similar to the concept of a *dominating pair* (Zhu et al., 2022) in the privacy literature.

### G.1 Finding worst-case adjacent distributions

**Adjacency parameters** The left side of Equation (38) optimizes over a high-dimensional set. In many cases the divergence depends only on some lower-dimensional "sufficient statistic" computable from $x, x'$, motivating the following definition.

**Definition G.1** (Adjacency parameter)**.** *Let $\phi$ be a noising mechanism over space $\mathcal{X}$. An* adjacency parameter *is a map $\theta\colon \mathcal{X} \times \mathcal{X} \to \Theta$ such that, for every generator $f$, there exists a function $\delta_f\colon \Theta \to \mathbb{R}^+$ such that $D_f(\phi_\# x' \| \phi_\# x) = \delta_f(\theta(x, x'))$ for all $x, x' \in \mathcal{X}$.*

Writing $\Theta_{\mathcal{B}}(x) \coloneqq \{\theta(x, x') \ : \ x' \in \mathcal{B}(x)\}$, Equation (38) becomes

$$\sup_{x' \in \mathcal{B}(x)} D_f(\phi_\# x' \| \phi_\# x) = \sup_{\theta \in \Theta_{\mathcal{B}}(x)} \delta_f(\theta), \tag{40}$$

which is often simpler when $\Theta$ is low-dimensional. How low-dimensional $\Theta$ can be is governed by symmetry of the constraint set under the noise $\phi$. By Lemma B.3, $D_f$ is invariant to invertible transformations. So, if $T$ is a transform with $T_\# \phi(x) = \phi(Tx)$ for every $x \in \mathcal{X}$, the divergence can only depend on those qualities of the pair $x, x'$ invariant under $T$. Taking Gaussian measures as an example, $T$ can be taken to be any Euclidean isometry, leaving $\theta(x, x') = \|x - x'\|_2$.

**Worst-case pairs** Equation (40) may have a different maximizer for each $f$. Ideally, however, this is not the case. Define the binary relation $(\mu', \nu') \preceq (\mu, \nu)$ if and only if, for every generator $f$, $D_f(\mu' \| \nu') \le D_f(\mu \| \nu)$. Similarly, call $\theta^\star$ a *worst-case* for $\Theta' \subseteq \Theta$ if $\sup_{\theta \in \Theta'} \delta_f(\theta) = \delta_f(\theta^\star)$ for every $f$.

It can be convenient to characterize worst-case measures through their likelihood ratios. We have the following lemmata, which will become useful later.

**Lemma G.1** (Hockey-stick characterization of $\preceq$)**.** *Let $\mu \ll \nu$ and $\mu' \ll \nu'$ be measures on the same space. Then*

$$(\mu', \nu') \preceq (\mu, \nu) \quad \iff \quad D_{\mathrm{HS},\beta}(\mu' \| \nu') \le D_{\mathrm{HS},\beta}(\mu \| \nu) \ \textit{for every } \beta > 0. \tag{41}$$

*Proof.* Each $f_\beta$ is itself a generator, which gives the forward implication. Conversely, $\rho$ is nonnegative in Lemma B.5, so a uniform bound on $D_{\mathrm{HS},\beta}(\mu' \| \nu')$ is preserved into $D_f(\mu' \| \nu')$. □

**Lemma G.2** (Product composition of worst-case pairs)**.** *For each $j = 1, \dots, k$, let $\mu_j, \mu_j', \nu_j$ be measures on $(\Omega_j, \mathcal{A}_j)$ with $\mu_j, \mu_j' \ll \nu_j$ and $(\mu_j', \nu_j) \preceq (\mu_j, \nu_j)$. Then $\left(\bigotimes_j \mu', \bigotimes_j \nu_j\right) \preceq \left(\bigotimes_j \mu, \bigotimes_j \nu_j\right)$. In particular, if $\theta_j^\star$ is a worst case parameter for each $j$, then $(\theta_1^\star, \dots, \theta_k^\star)$ is worst case for the product measure.*

*Proof.* Replacing $\mu_j'$ by $\mu_j$ one index at a time forms a sequence from $\mu'$ to $\mu$. Thus, by transitivity of $\preceq$, it suffices to prove the case where the two products differ only in a single factor (WLOG, the first): $\mu_j' = \mu_j$ for all $j \neq 1$.

Grouping the factors $\mu_{>1} \coloneqq \bigotimes_{j \ge 2} \mu_j$ and $\nu_{>1} \coloneqq \bigotimes_{j \ge 2} \nu_j$, and let $\Lambda_{>1}$ denote their likelihood ratio. Applying Lemma G.3 with $\mu_{>1}$ as the first factor and using Equation (41), we have

$$\begin{aligned} \widetilde{D}_{\mathrm{HS},\beta}(\mu' \otimes \mu_{>1} \| \nu_1 \otimes \nu_{>1}) &= \mathop{\mathbb{E}}_{\mu_{>1}} \left[ \widetilde{D}_{\mathrm{HS},\beta/\Lambda_{>1}}(\mu_1' \| \nu_1) \right] \\ &\le \mathop{\mathbb{E}}_{\mu_{>1}} \left[ \widetilde{D}_{\mathrm{HS},\beta/\Lambda_{>1}}(\mu_1 \| \nu_1) \right] = \widetilde{D}_{\mathrm{HS},\beta}(\mu_1 \otimes \mu_{>1} \| \nu_1 \otimes \nu_{>1}). \end{aligned}$$

Since $\beta > 0$ was arbitrary, the claim follows from Lemma G.1.

□

### G.2 Bounds between adjacent smoothing measures

In this appendix we derive Hockey-stick and Rényi divergences for the distributions arising in our analysis.

**General Strategy** We show how to compute $D_f(\mu\|\nu)$ when $\mu$ and $\nu$ are adjacent measures. In these computations, $\nu$ may be interpreted as the clean distribution, and $\nu$ the corrupted one. Each computation proceeds by writing down the distribution of $\Lambda = p/q$ under $\nu$ and evaluating Equation (14).

**Lemma G.3** (Hockey-stick for product measures)**.** *Let $\mu = \mu_1 \otimes \mu_2$ and $\nu = \nu_1 \otimes \nu_2$ and write $\Lambda_1$ and $\Lambda_2$ for the likelihood ratios under $(\mu_1, \nu_1)$ and $(\mu_2, \nu_2)$, respectively. Then, for every $\beta > 0$,*

$$\widetilde{D}_{\mathrm{HS},\beta}(\mu\|\nu) = \mathop{\mathbb{E}}_{\nu_1}\left[\Lambda_1 \widetilde{D}_{\mathrm{HS},\beta/\Lambda_1}(\mu_2\|\nu_2)\right] = \mathop{\mathbb{E}}_{\mu_1}\left[\widetilde{D}_{\mathrm{HS},\beta/\Lambda_1}(\mu_2\|\nu_2)\right]. \tag{42}$$

*Proof.* Using $(\lambda u - \beta)_+ = \lambda(u - \beta/\lambda)$ for $\lambda > 0$, we have

$$\widetilde{D}_{\mathrm{HS},\beta}(\mu\|\nu) = \mathop{\mathbb{E}}_{\nu}\left[(\Lambda - \beta)_+\right] = \mathop{\mathbb{E}}_{\nu_1}\mathop{\mathbb{E}}_{\nu_2}\left[(\Lambda_1\Lambda_2 - \beta)_+\right] = \mathop{\mathbb{E}}_{\nu_1}\left[\Lambda_1 \widetilde{D}_{\mathrm{HS},\beta/\Lambda_1}(\mu_2\|\nu_2)\right].$$

The final equality holds since $\mathbb{E}_{\nu_1}[\Lambda_1 h(\Lambda_1)] = \mathbb{E}_{\mu_1}[h(\Lambda_1)]$. □

**Corollary G.1.** *Suppose $\mu = \bigotimes_j \mu_j$, $\nu = \bigotimes_j \nu_j$, and that $\Lambda_j$ takes finitely many values for $j < k$. Let $\mathcal{V}$ be set of values of $\Lambda_{<k} \coloneqq \prod_{j<k} \Lambda_j$, and let $P_\lambda \coloneqq \mathbb{P}_{\mu_1\otimes\cdots\otimes\mu_{k-1}}[\Lambda_{<k} = \lambda]$. Then for every $\beta > 0$,*

$$\widetilde{D}_{\mathrm{HS},\beta}\beta(\mu\|\nu) = \sum_{\lambda\in\mathcal{V}} P_\lambda \widetilde{D}_{\mathrm{HS},\beta/\lambda}(\mu_k\|\nu_k). \tag{43}$$

*Moreover, if $\mu_k, \nu_k$ is also discrete, then $\widetilde{D}_{\mathrm{HS},\beta}(\mu\|\nu) = \sum_{\lambda>\beta}(1 - \beta/\lambda)P_\lambda$.*

*Proof.* Apply Lemma G.3 with the first factor $\mu_1 \otimes \cdots \otimes \mu_{k-1}$ and take the sum over the values of $\Lambda_{<k}$. The final claim holds by taking $\mu_k = \nu_k$ to be a point mass, so $\widetilde{D}_{\mathrm{HS},\beta/\lambda} = (1 - \beta/\lambda)_+$. □

Corollary G.1 allows us to fully understand the hockey stick profile of any measure constructed from discrete factors combined with a factor whose entire profile is known. The relevant application below is to construct joint feature-and-label noises which combine discrete label flipping with continuous feature noise (e.g., Definition G.6)).

The convolution in Equation (43) in general can be computationally expensive if many factors are tensored together. The number of discrete likelihood ratio values is upper-bounded by the product $\prod_j |\mathcal{V}_j|$. Thankfully, for the noises we consider, the likelihood ratio values collapse onto a lattice, greatly simplifying the computation.

**Smoothing measures** We now catalog the smoothing measures used in our work and their associated divergences.

**Definition G.2** (Uniform measure)**.** *Let $m \in \mathbb{N}$ and $\gamma \in \mathbb{R}^+$. Given $x \in \mathbb{R}^m$, the associated* Uniform *measure is the product measure*

$$\mu \coloneqq \bigotimes_{i=1}^{m} \mathrm{Unif}(x_i - \gamma, x_i + \gamma). \tag{44}$$

*We say that two Uniform measures on $\mathbb{R}^m$ with the same $\gamma$ and data $x, x'$ are $\theta$-adjacent if*

$$\theta = 1 - \prod_{i=1}^{m}\left(1 - \frac{|x_i - x_i'|}{2\gamma}\right)_+. \tag{45}$$

**Theorem G.1** (Divergences for adjacent uniform measures)**.** *Let $\mu, \nu$ be $\theta$-adjacent Uniform measures on $\mathbb{R}^m$. Then*

$$D_{\mathrm{R},\alpha}(\mu\|\nu) = \tfrac{1}{\alpha-1}\log(1-\theta), \text{ for } \alpha \in (0,1), \text{ else } \infty, \tag{46}$$

$$D_{\mathrm{KL}}(\mu\|\nu) = \infty, \tag{47}$$

$$D_{\mathrm{HS},\beta}(\mu\|\nu) = \min(1,\beta)\theta, \tag{48}$$

$$D_{\mathrm{TV}}(\mu\|\nu) = \theta. \tag{49}$$

*Proof.* $\mu$ and $\nu$ have densities $p, q$ which evaluate to $(2\gamma)^{-m}$ on their supports and 0 everywhere else. Thus the likelihood ratio $\Lambda$ satisfies the hypotheses of Lemma B.8, so we may write $D_f(\mu\|\nu) = (f(0) + f'(\infty))D_{\mathrm{TV}}(\mu\|\nu)$ for any generator $f$. The overlap region defining $\Lambda = 1$ can be seen to be of length $(2\gamma - |x_i - x_i'|)_+$ on each coordinate, which has mass under $\nu$ of $(1 - \frac{|x_i - x_i'|}{2\gamma})_+$. Thus $\Lambda = 1$ with probability $1 - D_{\mathrm{TV}}(\mu\|\nu) = 1 - \theta$, completing the proof. □

**Definition G.3** (Gaussian measure)**.** *Let $m \in \mathbb{N}$ and $\sigma \in \mathbb{R}^+$. Given $x \in \mathbb{R}^m$, the associated* Gaussian *measure is the standard isotropic Gaussian measure:*

$$\mu \coloneqq \mathcal{N}(x, \sigma^2 I_m). \tag{50}$$

*We say that two Gaussian measures on the same space $\mathbb{R}^m$ and same parameter $\sigma$ with data $x$ and $x'$, respectively, are $r$-adjacent if $\frac{\|x - x'\|_2}{\sigma} = r$.*

**Theorem G.2** (Divergences for adjacent Gaussian measures)**.** *Let $\mu = \mathcal{N}(x, \sigma^2 I_m)$ and $\nu = \mathcal{N}(x', \sigma^2 I_m)$ be $r$-adjacent Gaussian measures on $\mathbb{R}^m$. Then*

$$D_{\mathrm{R},\alpha}(\mu\|\nu) = \frac{\alpha r^2}{2} \tag{51}$$

$$D_{\mathrm{KL}}(\mu\|\nu) = \frac{r^2}{2} \tag{52}$$

$$D_{\mathrm{HS},\beta}(\mu\|\nu) = \Phi\left(\tfrac{r}{2} - \tfrac{\log\beta}{r}\right) - \beta\Phi\left(-\tfrac{r}{2} - \tfrac{\log\beta}{r}\right) - (1-\beta)_+ \tag{53}$$

$$D_{\mathrm{TV}}(\mu\|\nu) = 2\Phi\left(\tfrac{r}{2}\right) - 1 \tag{54}$$

*Proof.* Equations (51) to (54) are well-known; see Mironov (2017) for Equation (51) and Balle & Wang (2018) for Equation (53). Equation (52) follows by taking the limit $\alpha \to 1$ in Equation (51), and Equation (54) follows by taking $\beta = 1$ in Equation (53) and applying the identity $-\Phi(-x) = 1 - \Phi(x)$. □

**Definition G.4** (Laplace measure)**.** *Let $m \in \mathbb{N}$ and $b \in \mathbb{R}^+$. Given $x \in \mathbb{R}^m$, the associated* Laplace *measure is the product measure*

$$\mu \coloneqq \mathrm{Lap}(x, b)^{\otimes m}, \qquad \frac{d\mu}{dz}(z) = (2b)^{-m}\exp(-\|z - x\|_1/b). \tag{55}$$

*We say that two Laplace measures on the same space $\mathbb{R}^m$ and same parameter $b$ with data $x$ and $x'$, respectively, are $\mathbf{t}$-adjacent if*

$$\mathbf{t} = (|x_1 - x_1'|, \ldots, |x_m - x_m'|)/b \in \mathbb{R}^m_{\geq 0}, \tag{56}$$

*taken in nonincreasing order.*

**Theorem G.3** (Divergences for adjacent Laplace measures)**.** *Let $\mu, \nu$ be $\mathbf{t}$-adjacent Laplace measures on $\mathbb{R}^m$. For notational convenience, for $\alpha \in (0,1) \cup (1,\infty)$ and $t \geq 0$ define the quantities*

$$\zeta_\alpha(t) \coloneqq \tfrac{\alpha}{2\alpha-1}e^{(\alpha-1)t} + \tfrac{\alpha-1}{2\alpha-1}e^{-\alpha t}, \qquad \kappa_\alpha(t) \coloneqq \tfrac{1}{\alpha-1}\log\zeta_a(t), \tag{57}$$

*where we define by the limit $\zeta_{1/2}(t) = e^{-t/2}(1 + t/2)$. Under $\nu$, the likelihood ratio factorizes as $\Lambda = \prod_{j=1}^m \Lambda_j = e^{-\|\mathbf{t}\|_1}\prod_{j=1}^m V_{\mathbf{t}_j}$ where each $V_{\mathbf{t}_j}$ is independent and*

$$\mathbb{P}[V_t = 1] = \frac{1}{2}, \qquad \left.\tfrac{d\mathbb{P}}{dv}\right|_{(1,e^{2t})} = \frac{1}{4}v^{-3/2}, \qquad \mathbb{P}[V_t = e^{2t}] = \frac{1}{2}e^{-t}. \tag{58}$$

*Then, we have that*

$$D_{\mathrm{R},\alpha}(\mu\|\nu) = \textstyle\sum_j \kappa_\alpha(\mathbf{t}_j) \tag{59}$$

$$D_{\mathrm{KL}}(\mu\|\nu) = \textstyle\sum_j(\mathbf{t}_j + e^{-\mathbf{t}_j} - 1). \tag{60}$$

*If $\mathbf{t} = te_1$ has a single nonzero coordinate, then for all $\beta > 0$,*

$$D_{\mathrm{HS},\beta}(\mu\|\nu) = \left[(1 - \sqrt{\beta}e^{-t/2})_+ - (1-\beta)_+\right]_+, \tag{61}$$

$$D_{\mathrm{TV}}(\mu\|\nu) = 1 - e^{-t/2}. \tag{62}$$

*For general $\mathbf{t}$, the Hockey-stick divergence can be obtained numerically by convolving the $m$ laws of Equation* (58)*.*

*Proof.* We first exhibit the likelihood ratio structure from Equation (58). That the likelihood ratio factorizes into $\Lambda_j$ is clear from the product measure structure of $\mu, \nu$. WLOG we can assume $b = 1$, $x = 0$, and $x'_j = \mathbf{t}_j \geq 0$. Write $p_j, q_j$ for the densities of $\mu_j, \nu_j$, respectively. Each coordinate $j$ has

$$\Lambda_j(x) \;=\; \frac{p_j(x)}{q_j(x)} = \exp(-|x - \mathbf{t}_j| + |x|) \;=\; \begin{cases} \exp(-\mathbf{t}_j), & x \leq 0, \\ \exp(2x - \mathbf{t}_j), & 0 \leq x \leq \mathbf{t}_j, \\ \exp(\mathbf{t}_j), & x \geq \mathbf{t}_j. \end{cases}$$

We would like to understand the distribution of the variable $V_{\mathbf{t}_j} \coloneqq \exp(\mathbf{t}_j)\Lambda_j$ under $\nu$. On the endpoints we see that

$$\mathbb{P}_\nu[V_{\mathbf{t}_j} = 1] = \int_{\infty}^{0} 1/2 \exp(-|x|) = 1/2$$
$$\mathbb{P}_\nu[V_{\mathbf{t}_j} = e^{2t_j}] = \int_{t}^{\infty} 1/2 \exp(-|x|) = e^{-t}/2,$$

recovering the first and last identities in Equation (58). For the points $x \in (1, e^{2t_j})$, we note that the map $f(x) = e^{2x}$ is differentiable with differentiable inverse $f^{-1}(v) = \frac{1}{2}\ln(v)$, so we may apply the change of coordinates formula to arrive at the density

$$\left.\frac{d\mathbb{P}}{dv}\right|_{(1,e^{2\mathbf{t}_j})} \;=\; \frac{q(f^{-1}(v))}{|f'(f^{-1}(v))|} \;=\; \frac{\frac{1}{2}v^{-1/2}}{2v} \;=\; \tfrac{1}{4}v^{-3/2}.$$

Noting that $\prod_j e^{\mathbf{t}_j} = e^{\|\mathbf{t}\|_1}$, we recover $\Lambda = e^{-\|\mathbf{t}\|_1} \prod_{j=1}^m V_{\mathbf{t}_j}$.

**Rényi and KL divergences** Evaluating the Hellinger integral $Z_\alpha$ for a single coordinate via Equation (58) gives

$$Z_\alpha(\mu_j \| \nu_j) \;=\; \mathop{\mathbb{E}}_{\nu}[\lambda_j^\alpha] \;=\; e^{-\alpha \mathbf{t}_j} \mathop{\mathbb{E}}_{\nu}[V_{\mathbf{t}_j}^\alpha].$$

Computing the latter expectation gives

$$\begin{aligned} \mathop{\mathbb{E}}_{\nu}[V_{\mathbf{t}_j}^\alpha] &= \tfrac{1}{2} + \int_1^{e^{2t}} \tfrac{1}{4} v^{\alpha - 3/2}\, dv + \tfrac{1}{2} e^{-\mathbf{t}_j + 2\alpha \mathbf{t}_j} \\ &= \tfrac{1}{2} + \tfrac{1}{4\alpha - 2}(e^{(2\alpha-1)t} - 1) + \tfrac{1}{2} e^{-\mathbf{t}_j + 2\alpha \mathbf{t}_j} \\ &= (\tfrac{1}{2} - \tfrac{1}{2(2\alpha-1)}) + (\tfrac{1}{2} + \tfrac{1}{2(2\alpha-1)}) e^{(2\alpha-1)\mathbf{t}_j} \\ &= (\tfrac{\alpha - 1}{2\alpha - 1}) + (\tfrac{\alpha}{2\alpha-1}) e^{(2\alpha-1)\mathbf{t}_j}. \end{aligned}$$

Multiplying through by $e^{-\alpha \mathbf{t}_j}$ recovers the expression $Z_\alpha(\mu_j \| \nu_j) = \zeta_\alpha(t)$. The case for $\alpha = 1/2$ follows by recomputing the integral directly and noting $\int v^{-1} = \ln(v)$. Finally, taking $D_{\mathrm{R},\alpha}(\mu_j \| \nu_j) = \frac{1}{\alpha - 1} \log Z_\alpha(\mu_j \| \nu_j)$ and summing over coordinates (applying Lemma B.7) yields Equation (60). It can be verified that taking the limit $\alpha \to 1$ yields Equation (60).

**Hockey-stick divergence** Suppose $\mathbf{t} = te_1$. Cancelling identical factors (Lemma B.2(iii)), we need only compute $D_{\mathrm{HS},\beta}(\mu_1 \| \nu_1)$. By Equation (58) for $\Lambda_1$, we see that $x \mapsto \Lambda_1(x)$ is nondecreasing in $x$ with range $[e^{-t}, e^t]$. Hence, for $\beta \in (e^{-t}, e^t)$, we have the set

$$E_\beta \coloneqq \{\Lambda_1 \geq \beta\} = (x_\beta, \infty), \qquad x_\beta \coloneqq \tfrac{1}{2}(t + \log \beta) \in (0, t).$$

Integrating over these regions gives

$$\nu(E_\beta) = \textstyle\int_{x_\beta}^{\infty} \frac{1}{2} \exp(-|x|)\, dx = \frac{1}{2} e^{-x_\beta},$$
$$\mu(E_\beta) = \textstyle\int_{x_\beta}^{\infty} \frac{1}{2} \exp(-|x - t|)\, dx = 1 - \frac{1}{2} e^{-(t - x_\beta)}.$$

Using $e^{x_\beta} = \sqrt{\beta} e^{t/2}$, we compute the Hockey-stick directly as

$$D_{\mathrm{HS},\beta}(\mu \| \nu) \;=\; \mu(E_\beta) - \beta \nu(E_\beta) = 1 - \tfrac{1}{2}\sqrt{\beta} e^{-t/2} - \tfrac{1}{2}\sqrt{\beta} e^{-t/2} \;=\; 1 - \sqrt{\beta} e^{-t/2}.$$

When $\beta \geq e^t$, $\Lambda_1 \leq \beta$ everywhere and $D_{\mathrm{HS},\beta}(\mu \| \nu) = 0$. When $\beta \leq e^{-t}$, we have $\Lambda_1 \geq \beta$ everywhere and $D_{\mathrm{HS},\beta}(\mu \| \nu) = 1 - \beta$. Combining all three cases yields Equation (61). Taking $\beta = 1$ yields Equation (62). □

**Definition G.5** (Flipping Measure)**.** *Let $C \geq 2$ be an integer alphabet size, $y = (y_1, \ldots, y_m) \in [C]^m$, and $\gamma \in [0, 1]$. We define the* flipping measure

$$\phi_\gamma(y) \coloneqq \bigotimes_{i=1}^{m} \left((1-\gamma)e_{y_i} + \gamma \operatorname{Unif}([C])\right). \tag{63}$$

*We say that two flipping measures on the same space $[C]^m$ are $s$-adjacent if $\|y - y'\|_0 = s$*

**Theorem G.4** (Divergences for Adjacent Flipping Measures)**.** *Let*

$$\mu = \phi_\gamma(y) \qquad \textit{and} \qquad \nu = \phi_\gamma(y')$$

*be $s$-adjacent flipping measures on $[C]^m$. For notational convenience define the quantities*

$$p_+ \coloneqq 1 - \gamma + \tfrac{\gamma}{C}, \qquad p_- \coloneqq \tfrac{\gamma}{C}, \qquad p_0 \coloneqq \tfrac{(C-2)}{C}\gamma \tag{64}$$

*and*

$$d_\beta \coloneqq \left\lfloor \frac{\log \beta}{\log \frac{p_+}{p_-}} \right\rfloor + 1 \qquad \beta > 1 \tag{65}$$

*Then*

$$D_{\mathrm{R},\alpha}(\mu\|\nu) = \left(\tfrac{s}{\alpha-1}\right) \log(p_+^\alpha p_-^{1-\alpha} + p_+^{1-\alpha} p_-^\alpha + p_0) \tag{66}$$

$$D_{\mathrm{KL}}(\mu\|\nu) = s(1-\gamma) \log \tfrac{p_+}{p_-} \tag{67}$$

$$D_{\mathrm{HS},\beta}(\mu\|\nu) = \sum_{d=d_\beta}^{s} \left(1 - \beta(\tfrac{p_+}{p_-})^{-d}\right) P_d \tag{68}$$

$$D_{\mathrm{TV}}(\mu\|\nu) = \sum_{d=1}^{s} \left(1 - (\tfrac{p_+}{p_-})^{-d}\right) P_d \tag{69}$$

*where $P_d$ is the law*

$$P_d \coloneqq \mathbb{P}_{\widetilde{y}\sim\mu}[\|\widetilde{y} - y'\|_0 - \|\widetilde{y} - y\|_0 = d] \tag{70}$$

$$= \sum_{\substack{n_+ + n_- + n_0 = s \\ n_+ - n_- = d}} \binom{s}{n_+, n_-, n_0} p_+^{n_+} p_-^{n_-} p_0^{n_0} \tag{71}$$

$$= \sum_{l=\max(0,-d)}^{\lfloor (s-d)/2 \rfloor} \binom{s}{l+d, l, s-2l-d} p_+^{l+d} p_-^{l} p_0^{s-2l-d} \tag{72}$$

*Proof.* Without loss of generality, we may relabel coordinates so that $y_i \neq y'_i$ for $i \in [s]$, and $y_i = y'_i$ otherwise. Write $\mu = \bigotimes_{i=1}^m \mu_i = \bigotimes_{i=1}^m \phi_\gamma(y_i)$ and $\nu = \bigotimes_{i=1}^m \nu_i = \bigotimes_{i=1}^m \phi_\gamma(y'_i)$ By definition, $\mu_i$ assigns mass $p_+ = 1 - \gamma + \gamma/C$ to $y_i$ and mass $p_- = \gamma/C$ to every other symbol. $\nu_i$ is defined similarly, swapping $y_i$ for $y'_i$.

**Tensorizing Divergences** First, we point out that the Rényi divergence Equation (66) follows from the Tensorization Lemma Lemma B.7 since $\mu$ and $\nu$ decompose as product measures. In particular, we need only understand the component-wise divergences $D_{\mathrm{R},\alpha}(\mu_i\|\nu_i)$ for $1 \leq i \leq s$, and we may recover $D_{\mathrm{R},\alpha}(\mu\|\nu) = \sum_{i=1}^s D_{\mathrm{R},\alpha}(\mu_i\|\nu_i)$. For $1 \leq i \leq s$, $y_i \neq y'_i$, and so writing down the likelihood ratio for a single component takes one of three values:

$$\frac{\mu_i(\widetilde{y})}{\nu_i(\widetilde{y})} = \begin{cases} p_+ \,/\, p_- & \widetilde{y}_i = y_i \\ p_- \,/\, p_+ & \widetilde{y}_i = y'_i \\ 1 & \text{otherwise.} \end{cases} \tag{73}$$

We can then compute the expectation over $\widetilde{y}_i \sim \mu_i$:

$$\begin{aligned} \mathop{\mathbb{E}}_{\widetilde{y}_i\sim\mu_i} \left[\left(\frac{\mu_i(\widetilde{y})}{\nu_i(\widetilde{y})}\right)^{\alpha-1}\right] &= p_+ \left(\tfrac{p_+}{p_-}\right)^{\alpha-1} + p_- \left(\tfrac{p_-}{p_+}\right)^{\alpha-1} + p_0 \\ &= p_+^\alpha p_-^{1-\alpha} + p_-^\alpha p_+^{1-\alpha} + p_0 \end{aligned}$$

from which it follows

$$\begin{aligned} D_{\mathrm{R},\alpha}(\mu_i\|\nu_i) &= \left(\tfrac{1}{\alpha-1}\right)\log\left(\mathop{\mathbb{E}}_{\widetilde{y}_i\sim\nu_i}\left[\left(\tfrac{\mu_i(\widetilde{y})}{\nu_i(\widetilde{y})}\right)^{\alpha-1}\right]\right) \\ &= \left(\tfrac{1}{\alpha-1}\right)\log(p_+^{\alpha}p_-^{1-\alpha}+p_+^{1-\alpha}p_-^{\alpha}+p_0) \end{aligned}$$

allowing us to conclude

$$\begin{aligned} D_{\mathrm{R},\alpha}(\mu\|\nu) &= \sum_{i=1}^{s} D_{\mathrm{R},\alpha}(\mu_i\|\nu_i) \\ &= \left(\tfrac{s}{\alpha-1}\right)\log(p_+^{\alpha}p_-^{1-\alpha}+p_+^{1-\alpha}p_-^{\alpha}+p_0). \end{aligned}$$

Computing $D_{\mathrm{KL}}(\mu\|\nu)$ proceeds similarly by using Equation (73) and evaluating

$$\begin{aligned} D_{\mathrm{KL}}(\mu_i\|\nu_i) &= \mathop{\mathbb{E}}_{\widetilde{y}\sim\mu}\left[\log\tfrac{\mu_i(\widetilde{y})}{\nu_i(\widetilde{y})}\right] \\ &= p_+\log\tfrac{p_+}{p_-}+p_-\log\tfrac{p_-}{p_+} \\ &= (p_+-p_-)\log\tfrac{p_+}{p_-} \\ &= (1-\gamma)\log\tfrac{p_+}{p_-}. \end{aligned}$$

Using tensorization of KL gives the identify from Equation (67).

**Hockey Stick Divergence** To compute the Hockey Stick Divergence we must understand the entire likelihood ratio distribution $\mu/\nu$ holistically. For $\widetilde{y}\sim\nu$, define the random distances

$$\delta \coloneqq \|\widetilde{y}-y\|_0 \qquad \delta' \coloneqq \|\widetilde{y}-y'\|_0, \tag{74}$$

$$\delta_s \coloneqq \|\widetilde{y}^{(s)}-y^{(s)}\|_0 \qquad \delta' \coloneqq \|\widetilde{y}^{(s)}-y'^{(s)}\|_0 \tag{75}$$

where $\widetilde{y}^{(s)}$ and $y^{(s)}$ are the respective projections onto the first $s$ coordinates. Define the signed random difference $d \coloneqq \delta'-\delta$. Notably, since $y$ and $y'$ only disagree in the first $s$ coordinates, the difference $d$ is determined entirely by the projection of $y$, $y'$, and $\widetilde{y}$ onto the first $s$ coordinates. That is, $d=\delta'-\delta=\delta'^{(s)}-\delta^{(s)}$.

For any given $\widetilde{y}\in[C]^m$, we can write

$$\begin{aligned} \mu(\widetilde{y}) &= \prod_{i=1}^{n} p_+^{\mathbb{1}[\widetilde{y}_i=y_i]}p_-^{\mathbb{1}[\widetilde{y}\neq y_i]} = p_+^{n-\delta}p_-^{\delta} \\ \nu(\widetilde{y}) &= \prod_{i=1}^{n} p_+^{\mathbb{1}[\widetilde{y}_i=y'_i]}p_-^{\mathbb{1}[\widetilde{y}\neq y'_i]} = p_+^{n-\delta'}p_-^{\delta'} \end{aligned}$$

Thus the likelihood ratio $\mu(\widetilde{y})/\nu(\widetilde{y})$ is given by

$$\frac{\mu(\widetilde{y})}{\nu(\widetilde{y})} = p_+^{n-\delta}p_-^{\delta}p_+^{\delta'-n}p_-^{-\delta'} \;=\; p_+^{\delta'-\delta}p_-^{\delta-\delta'} \;=\; \left(\tfrac{p_+}{p_-}\right)^{d}. \tag{76}$$

$d$ takes values in $\{-s,\dots,0,\dots,s\}$, and so there are exactly $2s+1=\mathcal{O}(s)$ distinct likelihood ratio value regions. Now, from Equation (76), we would like to understand the law for $d$ under $\nu$. We start by expressing $d$ as the result of a biased random walk:

$$d \overset{d}{=} \sum_i d_i, \tag{77}$$

$$d_i \overset{\text{i.i.d.}}{\sim} \begin{cases} 1 & \text{with prob. } p_+ \\ 0 & \text{with prob. } p_0 \\ -1 & \text{with prob. } p_-. \end{cases} \tag{78}$$

We can understand the law for $d$ using the theory of probability generating functions (PGFs). If $X$ is a discrete integer-values random variable taking values in $\{-B, \dots, B\}$ with mass function $p(x)$, the PGF for $X$ is defined as

$$G_X(z) \coloneqq \mathop{\mathbb{E}}_{x \sim X}[z^x] = \sum_{x=-b}^{b} p(x) z^x.$$

The PGF for $d_i$ is thus $G_i(z) = p_- z^{-1} + p_0 + p_+ z$. The PGF for a sum of independent random variables is simply a product the PGFs for the variables, and thus the PGF for $d$ is given by

$$G(z) = \prod_{i=1}^{s} G_i(z) = (p_- z^{-1} + p_0 + p_+ z)^s. \tag{79}$$

We recover the probability $P_d = \mathbb{P}_{\widetilde{y} \sim \nu}[d = \widetilde{d}]$ by extracting the coefficient of the $z^{\widetilde{d}}$ term, in this case computed as

$$[z^{\widetilde{d}}]G(z) = \sum_{\substack{n_+ + n_- + n_0 = s \\ n_+ - n_- = d}} \binom{s}{n_+, n_-, n_0} p_+^{n_+} p_-^{n_-} p_0^{n_0} \tag{80}$$

where the inner probability corresponds to choosing $n_+$ choices of $p_+$, $n_-$ choices of $p_-$, and $n_0$ choices of $p_0$ in the random walk.

Knowing the law for $d$ under $\nu$, we recover $\mathcal{O}(s)$ regions of constant likelihood ratio, one for each $d$, and each of which has likelihood ratio $(p_+/p_-)^d$. To compute the Hockey Stick divergence, we must also understand the law for $d$ under $\mu$, defined by

$$Q_d \coloneqq \mathbb{P}_{\widetilde{y} \sim \mu}[d = \widetilde{d}]. \tag{81}$$

Luckily, moving from $\nu$ to $\mu$ under a given $\widetilde{y}$ is straightforward to understand. Using the same random walk formulation gives inverted step probabilities, and pushing through the PGF $H$ gives $H(z) = G(1/z)$. Extracting coefficients, we see that

$$Q_d = P_{-d}. \tag{82}$$

Moreover, $P_d$ itself admits a certain multiplicative symmetry simplifying the expression further. For a fixed $n_+, n_-$ satisfying $n_+ - n_- = -d$, we have

$$\begin{aligned} p_+^{n_+} p_-^{n_-} p_0^{n_0} &= \left(\tfrac{p_+}{p_-} p_-\right)^{n_+} \left(\tfrac{p_-}{p_+} p_+\right)^{n_-} p_0^{n_0} && (83) \\ &= \left(\tfrac{p_+}{p_-}\right)^{n_+ - n_-} p_-^{n_+} p_+^{n_-} p_0^{n_0} && (84) \\ &= \left(\tfrac{p_+}{p_-}\right)^{-d} p_-^{n_+} p_+^{n_-} p_0^{n_0}. && (85) \end{aligned}$$

From Equation (85) we can simply swap the labels of $n_-$ and $n_+$ to get the constraint $n_+ - n_- = d$, and plugging into Equation (71) and combining Equation (82) gives us the identity

$$\begin{aligned} Q_d &= P_{-d} && (86) \\ &= \left(\tfrac{p_+}{p_-}\right)^{-d} \sum_{\substack{n_+ + n_- + n_0 = s \\ n_+ - n_- = d}} \binom{s}{n_+, n_-, n_0} p_+^{n_+} p_-^{n_-} p_0^{n_0} && (87) \\ &= \left(\tfrac{p_+}{p_-}\right)^{-d} P_d. && (88) \end{aligned}$$

All that remains is to integrate the divergence loss $\mu - \beta\nu$ over the region $\Omega_\beta \coloneqq \{\widetilde{y} \in [C]^n : \mu(\widetilde{y}) > \beta\nu(\widetilde{y})\}$. Defining the sets $\Omega_d \coloneqq \{\widetilde{y} \in [C]^n : \delta' - \delta = d\}$ for each $-s \le d \le s$, we see that $\{\Omega_d\}_{d=-s}^{s}$ partitions the outcome space, and in particular integrating $\mu - \beta\nu$ over $\Omega_\beta$ can be done by integrating over each $\Omega_\beta \cap \Omega_d$ and taking the sum. Since the likelihood ratio is constant on each $\Omega_d$, this amounts to isolating the $d$ satisfying $(p_+/p_-)^d \ge \beta$ and summing the corresponding masses. We see that

$$\begin{aligned} (p_+/p_-)^d > \beta &\iff d > \frac{\log(\beta)}{\log(p_+/p_-)} \\ &\iff d \ge \left\lfloor \frac{\log(\beta)}{\log(p_+/p_-)} \right\rfloor + 1. \end{aligned}$$

Defining $d_\beta \coloneqq \lfloor \log(\beta)/\log(p_+/p_-) \rfloor + 1$, we conclude

$$D_{\mathrm{HS},\beta}(\mu\|\nu) = \sum_{d=d_\beta}^{s} (P_d - \beta Q_d) \tag{89}$$

$$= \sum_{d=d_\beta}^{s} \left(1 - \beta(\tfrac{p_+}{p_-})^{-d}\right) P_d \tag{90}$$

**TV Distance** The TV distance follows directly from setting $\beta = 1$ in Equation (68) and observing that $d_\beta = 1$. □

The Flipping Measure of Definition G.5 has an interesting history. It is essentially a $C$-ary randomized response (Warner, 1965), well-studied in both the DP and randomized smoothing literatures (Mironov, 2017; Lee et al., 2019; Bojchevski et al., 2020; Dvijotham et al., 2020; Zhang et al., 2022). Mironov (2017) computes the $s = 1, C = 2$ case; see Table II in that work. The general Rényi divergence Equation (66) appears in Appendix A.10 of Dvijotham et al. (2020), but they do not derive the corresponding $D_{\mathrm{HS},\beta}$ divergence which can yield tighter certificates via Theorem A.3. Lee et al. (2019) and Zhang et al. (2022) both operate implicitly with the Hockey Stick divergences by way of the Neyman-Pearson Lemma characterization of optimal certificates, although these works elude an efficient closed-form expression and instead compute with a runtime-suboptimal partitioning algorithm with runtime $\mathcal{O}(d^3 + k^2 d^2)$ (the practical impacts of which are discussed in (Dvijotham et al., 2020)). Bojchevski et al. (2020) consider flipping measures and identify the $\mathcal{O}(s)$ likelihood ratio partition structure that allows us to obtain the closed-form Equations (68) and (69), but do not explicitly write down the closed-form divergence.

**Definition G.6** (Gaussian–Flipping Measure)**.** *Let $m, n \in \mathbb{N}$, an integer alphabet size $C \geq 2$, $\sigma \in \mathbb{R}^+$, and $\gamma \in [0, 1]$. For $j \in [C]$, define the flipping measure*

$$\phi_\gamma(j) \coloneqq (1 - \gamma)e_j + \gamma \,\mathrm{Unif}([C]) \tag{91}$$

*on $[C]$. Given $x \in \mathbb{R}^m$ and $y \in [C]^n$, the associated* Gaussian–Flip measure *is the product measure on $\mathbb{R}^m \times [C]^n$*

$$\mu \coloneqq \mathcal{N}(x, \sigma^2 I_m) \otimes \bigotimes_{i=1}^{n} \phi_\gamma(y_i). \tag{92}$$

*We say that two Gaussian–Flip measures on the same space $\mathbb{R}^m \times [C]^n$ and same parameters $\sigma, \gamma$, with data $(x, y)$ and $(x', y')$, respectively, are $(r, s)$-*adjacent *if*

$$\tfrac{\|x - x'\|_2}{\sigma} = r \quad \textit{and} \quad \|y - y'\|_0 = s. \tag{93}$$

The Gaussian–Flipping measure of Definition G.6 has two main applications within our framework. First, it allows joint smoothing of features and labels for dirty-label data poisoning adversaries. Second, it allows smoothing over feature spaces that combine continuous feature representations with discrete ones, for example of the form $\mathcal{X} = \mathbb{R}^m \times K^n$ for some discrete space $K$. In such a setting a data sample can be noised by applying continuous noise (e.g., Gaussian) to the continuous features and discrete noise (e.g., flipping) to the categorical features.

**Theorem G.5** (Divergences for Adjacent Gaussian–Flipping Measures)**.** *Let*

$$\mu = \mathcal{N}(\delta, \sigma^2 I_m) \otimes \bigotimes_{i=1}^{n} \phi_\gamma(y_i), \qquad \nu = \mathcal{N}(\delta', \sigma^2 I_m) \otimes \bigotimes_{i=1}^{n} \phi_\gamma(y_i')$$

*be $(r, s)$-adjacent Gaussian–Flip measures on $\mathbb{R}^m \times [C]^n$. Define $p_+ \coloneqq 1 - \gamma + \frac{\gamma}{C}$, $p_- \coloneqq \frac{\gamma}{C}$, $p_0 \coloneqq \frac{(C-2)}{C}\gamma$. For $-s \leq d \leq s$, let $P_d$ be the law of Equation* (71) *and define the* tilted threshold

$$\beta_d \coloneqq \beta \left(\tfrac{p_+}{p_-}\right)^{-d}, \qquad \beta > 0. \tag{94}$$

*Then*

$$D_{\mathrm{R},\alpha}(\mu\|\nu) = \frac{\alpha r^2}{2} + \frac{s}{\alpha-1}\log(p_+^\alpha p_-^{1-\alpha} + p_+^{1-\alpha}p_-^\alpha + p_0) \tag{95}$$

$$D_{\mathrm{KL}}(\mu\|\nu) = \frac{r^2}{2} + s(1-\gamma)\log\tfrac{p_+}{p_-} \tag{96}$$

$$D_{\mathrm{HS},\beta}(\mu\|\nu) = \sum_{d=-s}^{s} P_d\Big(\Phi\left(\tfrac{r}{2} - \tfrac{\log\beta_d}{r}\right) - \beta_d\Phi\left(-\tfrac{r}{2} - \tfrac{\log\beta_d}{r}\right)\Big) - (1-\beta)_+ \tag{97}$$

$$D_{\mathrm{TV}}(\mu\|\nu) = \sum_{d=-s}^{s} P_d\Big(\Phi\left(\tfrac{r}{2} + \tfrac{d}{r}\log\tfrac{p_+}{p_-}\right) - \left(\tfrac{p_+}{p_-}\right)^{-d}\Phi\left(-\tfrac{r}{2} + \tfrac{d}{r}\log\tfrac{p_+}{p_-}\right)\Big) \tag{98}$$

*Proof.* **Rényi and KL divergences.** Applying tensorization (Lemma B.7), the result follows by adding the Gaussian values of Theorem G.2 and the flipping values of Theorem G.4.

**Hockey-stick divergence.** Apply Lemma G.3 with the flipping block as the first factor. Under the flipping factor of $\mu$, the likelihood ratio equals $(p_+/p_-)^d$ with probability $P_d$, so $\widetilde{D}_{\mathrm{HS},\beta}(\mu\|\nu) = \sum_d P_d\,\widetilde{D}_{\mathrm{HS},\beta_d}\big(\mathcal{N}(x,\sigma^2 I_m)\|\mathcal{N}(x',\sigma^2 I_m)\big)$, where the Gaussian term is $\Phi(\frac{r}{2} - \frac{\log\beta_d}{r}) - \beta_d\Phi(-\frac{r}{2} - \frac{\log\beta_d}{r})$. Subtracting $(1-\beta)_+$ gives Equation (97).

**TV Divergence.** The TV Divergence bound follows from taking $\beta = 1$ in Equation (97). □

### G.3 Worst-case Uniform parameters

In this appendix we derive the worst-case Uniform noise adjacency parameter under different $\ell_p$ constraints. Luckily, Uniform noise admits a low-dimensional adjacency parameter $\theta = 1 - \prod_i(1 - \frac{|x_i - x_i'|}{2\gamma})_+$ (Definition G.2). By the proof of Theorem G.1, $D_f(\mu\|\nu) = (f(0) + f'(\infty))\theta$ for every generator $f$, which is nondecreasing in $\theta$. Hence, the optimization problem Equation (40) is always maximized at the maximum attainable $\theta$. This reduces the problem to maximizing $\theta$ under an $\ell_p$ constraint.

Write $\Delta \coloneqq x - x'$. For a budget $\|\Delta\|_p \le \varepsilon$, write $t_i \coloneqq |\delta_i|/(2\gamma)$ and $\eta = \varepsilon/(2\gamma)$. Define the quantity

$$g(t) \coloneqq -\log(1-t) \;\text{ for } t < 1, \qquad g(t) \coloneqq \infty \;\text{ for } t \ge 1. \tag{99}$$

so that $-\log(1-\theta) = \sum_{i=1}^m g(t_i)$. Evidently $\theta$ is nondecreasing in $\sum_{i=1}^m g(t_i)$, so we seek to maximize $\sum_{i=1}^m g(t_i)$ subject to the constraint $\|\mathbf{t}\|_p \le \eta$. If $\eta \ge 1$, the adversary can shift one coordinate by $2\gamma$, in which case the supports are disjoint, in which case $\theta = 1$ for every $p$. We therefore assume $\eta < 1$.

**Theorem G.6** (Worst-case Uniform parameter)**.** *Let $\eta < 1$, and write $\theta_p^\star \coloneqq \sup\{\theta(x,x') \colon \|x - x'\|_p \le \varepsilon\}$ for Uniform measures on $\mathbb{R}^m$ with radius $\gamma$. Define $\Phi_p \coloneqq -\log(1-\theta_p^\star)$. Then $\theta_p^\star = 1 - e^{-\Phi_p}$, where*

1. *$\Phi_1 = g(\eta)$, attained at $\mathbf{t} = \eta e_1$;*
2. *$\Phi_\infty = mg(\eta)$, attained at $\mathbf{t} = \eta\mathbf{1}$;*
3. *for $1 < p < \infty$ and $m \ge 2$, we have*

$$m^{1-1/p}\eta \le \Phi_p \le m^{1-1/p}\eta + m^{\max(0,1-2/p)}\frac{\eta^2}{2(1-\eta)}. \tag{100}$$

*Proof.* We start with the case $p = 1$. Since $g$ is convex with $g(0) = 0$, for any $a, c$ with $a + c < 1$, we have $g(a) \le \frac{a}{a+c}g(a+c)$ and $g(c) \le \frac{c}{a+c}g(a+c)$, so that $g(a) + g(c) \le g(a+c)$. In particular, if $\mathbf{t}$ is nonzero in more than one coordinate, we can obtain a larger objective $sum_i g(t_i)$ by merging coordinates. Via inductive argument, the objective is maximized at $\mathbf{t} = \eta e_1$. The corresponding maximum is $g(\eta)$.

Now, the case when $p = \infty$. This case is easy, since $g$ is increasing in each coordinate, we need only maximize each coordinate independently to obtain $g(\mathbf{t}) = mg(\eta)$.

Finally, the case when $1 < p < \infty$. For $0 \le t < 1$, we have the series representation $g(t) = \sum_{j\ge1} t^j/j$, which gives the bound $g(t) \le t + \frac{1}{2}\sum_{j\ge2} t^j = t + t^2/(2(1-t))$. Every coordinate satisfies $t_i \le \|\mathbf{t}\| \le \eta$, so

$$\sum_{i=1}^{m} g(t_i) \le \|\mathbf{t}\|_1 + \frac{\|\mathbf{t}\|_2^2}{2(1-\eta)}.$$

By Hölder's inequality, $\|\mathbf{t}\|_1 \le m^{1-1/p}\|\mathbf{t}\|_p$ and $\|\mathbf{t}\|_2 \le m^{\max(0,1/2-1/p)}\|\mathbf{t}\|_p$, giving the upper bound. For the lower bound, we can take $t_i = \eta m^{-1/p}$ so that $\|\mathbf{t}\|_p = \eta$. We conclude that

$$\sum_{i=1}^{m} g(t_i) = mg(\eta m^{-1/p}) \ge m^{1-1/p}\eta,$$

where the final inequality uses that $g(s) \ge s$. □

### G.4 Worst-case Laplace parameters

To bound worst-case divergences for Laplace noise, we must solve Equation (40) applied to Theorem G.3. Unlike other noises like Gaussian or Uniform, the adjacency parameter for Laplace noise is the list-valued $\mathbf{t}$ in Equation (56). Concretely, in this appendix we attempt to solve (or bound)

$$\tau_f^\star(\varepsilon, p) = \sup\left\{D_f(\mu_{\mathbf{t}}\|\nu) \;:\; \mathbf{t} \in \mathbb{R}^m_{\ge0},\ \|\mathbf{t}\|_p \le \eta\right\}, \qquad \eta \coloneqq \varepsilon/b. \tag{101}$$

Throughout this appendix we write $\nu \coloneqq \mathrm{Lap}(0,b)^{\otimes m}$ and $\mu_{\mathbf{t}}$ for the measure centered at $b\mathbf{t}$, and take $b = 1$.

We summarize our findings here. The $\ell_1$ and $\ell_\infty$ cases are solvable exactly for all $f$-divergences. For $\ell_\infty$, the argument is straightforward: every divergence is monotone in each coordinate $\mathbf{t}$, so is maximized at $\mathbf{t} = \eta\mathbf{1}$. The result for $\ell_\infty$ is stated in Theorem G.7. For $\ell_1$, the divergence is maximized when the adversary concentrates their perturbation in a single coordinate. The result for $\ell_1$ is stated in Theorem G.8.

**Technical lemmas** We first present some technical lemmas which will help us understand the structure of Laplace noise.

**Lemma G.4** (Coordinate-wise monotonicity)**.** *For $0 \le t \le t'$, $(\mathrm{Lap}(t,1), \mathrm{Lap}(0,1)) \preceq (\mathrm{Lap}(t',1), \mathrm{Lap}(0,1))$. In particular, $(\mu_{\mathbf{t}}, \nu) \preceq (\mu_{\mathbf{t}'}, \nu)$ whenever $\mathbf{t} \le \mathbf{t}'$ coordinate-wise.*

*Proof.* By Equation (41) it suffices to compare the Hockey-stick profiles at every $\beta > 0$, which are determined by Equation (61) as $D_{\mathrm{HS},\beta}(\mathrm{Lap}(t,1)\|\,\mathrm{Lap}(0,1)) = [(1-\sqrt{\beta}e^{-t/2})_+ - (1-\beta)_+]_+$. By inspection, this quantity is nondecreasing in $t$ for every $\beta$, which gives the first claim.

The second claim follows from the first by applying Lemma G.2 for each coordinate. □

The following technical lemma characterizes the Hockey-stick behavior under an $\ell_1$ constraint when the adversary splits the perturbation between two coordinates.

**Lemma G.5.** *Let $s, t > 0$ and write $\eta \coloneqq s + t$. Let $\mu_{(s,t)}, \nu$ be $(s,t)$-adjacent Laplace measures on $\mathbb{R}^2$. Then, for every $\beta \in (e^{-\eta}, e^{\eta})$,*

$$\widetilde{D}_{\mathrm{HS},\beta}(\mu_{(s,t)}\|\nu) = 1 - \sqrt{\beta}e^{-\eta/2}(1+\Delta), \qquad \Delta \coloneqq \tfrac{1}{2}\min\{\ell, s, t, \eta-\ell\}, \quad \ell \coloneqq \tfrac{\eta+\log\beta}{2}, \tag{102}$$

*while $\widetilde{D}_{\mathrm{HS},\beta}(\mu_{(s,t)}\|\nu) = 1-\beta$ for $\beta < e^{-\eta}$ and 0 for $\beta \ge e^{\eta}$. In particular, $\Delta \ge 0$, so $\widetilde{D}_{\mathrm{HS},\beta}(\mu_{(s,t)}\|\nu) \le \widetilde{D}_{\mathrm{HS},\beta}(\mu_{\eta e_1}\|\nu)$ for every $\beta > 0$, with equality if and only if $\beta \notin (e^{-\eta}, e^{\eta})$.*

*Proof.* First, when $\beta \notin (e^{-\eta}, e^{\eta})$, the corresponding results hold since $|\log\Lambda| \le \eta$ almost surely by Theorem G.3. Fix $\beta \in (e^{-\eta}, e^{\eta})$, and write $\mu_s \coloneqq \mathrm{Lap}(s,1)$ and $\mu_t \coloneqq \mathrm{Lap}(t,1)$ for the marginals of $\mu_{(s,t)}$, $\nu_0 \coloneqq \mathrm{Lap}(0,1)$ for the common marginal of $\nu$, and $\Lambda_t$ for the likelihood ratio of $(\mu_t, \nu_0)$.

Applying Lemma G.3 to $\mu_{(s,t)}, \nu$ (with $\mu_t$ as the first factor), we have

$$\widetilde{D}_{\mathrm{HS},\beta}(\mu_{(s,t)}\|\nu) = \mathop{\mathbb{E}}_{\mu_t}[\widetilde{D}_{\mathrm{HS},\beta/\Lambda_t}(\mu_s\|\nu_0)],$$

where by Equation (61),

$$\widetilde{D}_{\mathrm{HS},\beta/\Lambda_t}(\mu_s\|\nu_0) = \begin{cases} 0, & \Lambda_t \le \beta e^{-s}, \\ 1 - \sqrt{\beta/\Lambda_t}e^{-s/2}, & \beta e^{-s} \le \Lambda_t \le \beta e^{s}, \\ 1 - \beta/\Lambda_t, & \Lambda_t \ge \beta e^{s}. \end{cases} \tag{103}$$

Since $\Lambda_t \in [e^{-t}, e^t]$, the branches of Equation (103) entering into the expectation depend on $s$ and $t$. The first line applies exactly when $\beta e^{-s} \ge e^{-t}$; that is, $\ell \ge s$, and the third exactly when $\beta e^{s} \le e^{t}$; that is, $\ell \le t$.

By Theorem G.3, we have $\Lambda_t = e^{-t}V_t$, so under $\nu_0$, it follows that $\Lambda_t$ is distributed with mass $\frac{1}{2}$ at $e^{-t}$, mass $\frac{1}{2}e^{-t}$ at $e^t$, and density $\frac{1}{4}e^{-t/2}\lambda^{-3/2}$ on $(e^{-t}, e^t)$. Integrating the measure gives, for $x \in (e^{-t}, e^t]$,

$$\begin{aligned} \mathbb{P}_{\nu_0}[\Lambda_t \ge x] &= \tfrac{1}{2}e^{-t/2}x^{-1/2}, \\ \mathop{\mathbb{E}}_{\nu_0}[\Lambda_t; \Lambda_t \ge x] &= 1 - \tfrac{1}{2}e^{-t/2}x^{1/2}, \\ \mathop{\mathbb{E}}_{\nu_0}[\Lambda_t^{1/2}; \Lambda_t < x] &= e^{-t/2}(\tfrac{1}{2} + \tfrac{1}{4}\log(xe^t)), \\ \mathop{\mathbb{E}}_{\nu_0}[\Lambda_t^{1/2}] &= Z_{1/2}(\mu_t\|\nu_0) = \zeta_{1/2}(t). \end{aligned}$$

We have four meaningful cases depending on the ordering of $s, t, \ell$.

**Case** $t \le \ell \le s$ In this case only the second line of Equation (103) applies, so

$$\widetilde{D}_{\mathrm{HS},\beta}(\mu_{(s,t)}\|\nu) = 1 - \sqrt{\beta}e^{-s/2}\mathop{\mathbb{E}}_{\nu_0}[\Lambda_t^{1/2}] = 1 - \sqrt{\beta}e^{-s/2}e^{-t/2}(1 + \tfrac{t}{2}) = 1 - \sqrt{\beta}e^{-\eta/2}(1 + \tfrac{t}{2}).$$

In this case $\min\{\ell, s, t, \eta - \ell\} = t$.

**Case** $\ell \ge \max(s, t)$ In this case the first line of Equation (103) applies on $\{\Lambda_t \le \beta e^{-s}\}$ and the second on the rest, so

$$\begin{aligned} \widetilde{D}_{\mathrm{HS},\beta}(\mu_{(s,t)}\|\nu) &= \mathop{\mathbb{E}}_{\nu_0}[\Lambda_t; \Lambda_t \ge \beta e^{-s}] - \sqrt{\beta}e^{-s/2}\mathop{\mathbb{E}}_{\nu_0}[\Lambda_t^{1/2}; \Lambda_t \ge \beta e^{-s}] \\ &= 1 - \sqrt{\beta}e^{-\eta/2}(1 + \tfrac{1}{2}(\eta - \ell)). \end{aligned}$$

In this case $\min\{\ell, s, t, \eta - \ell\} = \eta - \ell$.

**Case** $\ell \le \min(s, t)$ In this case the third line of Equation (103) applies on $\{\Lambda_t \ge \beta e^{s}\}$ and the second on the rest, so

$$\begin{aligned} \widetilde{D}_{\mathrm{HS},\beta}(\mu_{(s,t)}\|\nu) &= 1 - \sqrt{\beta}e^{-s/2}\mathop{\mathbb{E}}_{\nu_0}[\Lambda_t^{1/2}; \Lambda_t < \beta e^{s}] - \beta\mathbb{P}[\Lambda_t \ge \beta e^{s}] \\ &= 1 - \sqrt{\beta}e^{-\eta/2}(1 + \tfrac{\ell}{2}). \end{aligned}$$

In this case $\min\{\ell, s, t, \eta - \ell\} = \ell$.

**Case** $s \le \ell \le t$ In this case $\mu_{(s,t)}$ and $\mu_{(t,s)}$ are related by swapping coordinates. By Lemma B.3 both sides of Equation (102) are symmetric in $s$ and $t$, so we can reduce to the case $t \le \ell \le s$ handled above.

We have that $\Delta \ge 0$ in all four cases. Finally, taking $s = \eta$ and $t = 0$ gives the final comparison. □

The property that ultimately allows us to characterize Laplace noise for $\ell_1$ adversaries is *Schur convexity*. Recall that we say $\mathbf{t}$ is *majorized* by $\mathbf{t}'$, written $\mathbf{t} \prec \mathbf{t}'$, if the two have equal sums and every prefix sum of the coordinates of $\mathbf{t}$ when listed in nonincreasing order is at most the corresponding sum for $\mathbf{t}$. If the prefix-sum condition holds but the sums of $\mathbf{t}$ and $\mathbf{t}'$ may differ, we say $\mathbf{t}$ is *weakly submajorized* by $\mathbf{t}'$, written $\mathbf{t} \prec_w \mathbf{t}'$.

**Lemma G.6** (Schur convexity in Laplace parameter)**.** *For every generator $f$, the map $\mathbf{t} \mapsto D_f(\mu_{\mathbf{t}} \| \nu)$ is Schur convex. In particular, if $\mathbf{t} \prec \mathbf{t}'$, then $(\mu_{\mathbf{t}}, \nu) \prec (\mu_{\mathbf{t}'}, \nu)$.*

*Proof.* Suppose $\mathbf{t} \prec \mathbf{t}'$. As a property of the majorization relation, it holds that there exists a sequence $\mathbf{t} = \mathbf{t}^{(0)} \prec \mathbf{t}^{(1)} \prec \cdots \prec \mathbf{t}^{(n)} = \mathbf{t}'$ such that each adjacent pair differs only a single pair of coordinates $i, j$ and with $\mathbf{t}_i^{(k)} + \mathbf{t}_j^{(k)} = \mathbf{t}_i^{(k+1)} + \mathbf{t}_j^{(k+1)}$. Thus, it is sufficient to prove the result only for the case when $\mathbf{t}, \mathbf{t}'$ differ in a single coordinate pair with $\mathbf{t}_i + \mathbf{t}_j = \mathbf{t}'_i + \mathbf{t}'_j$ and $\min(\mathbf{t}_i, \mathbf{t}_j) \geq \min(\mathbf{t}'_i, \mathbf{t}'_j)$.

Applying Equation (102) on the two swapped coordinates for $\mathbf{t}$ and $\mathbf{t}'$, with $\eta = \mathbf{t}_i + \mathbf{t}_j$ and fixed $\beta$, we see that $\Delta \geq \Delta'$ because $\Delta$ is nondecreasing in $\min(\cdot, \cdot)$. In particular, the pair $(\mathbf{t}_i, \mathbf{t}_j)$ generates a smaller Hockey-stick divergence under Lemma G.5.

Finally, applying Lemma G.2, we conclude that $(\mu_{\mathbf{t}}, \nu) \preceq (\mu_{\mathbf{t}'}, \nu)$. □

### $\ell_\infty$ constraints

**Theorem G.7** (Worst-case $\ell_\infty$ Laplace parameter)**.** *Let $\mathbf{t} \in \mathbb{R}^m_{\geq 0}$ with $\|\mathbf{t}\|_\infty \leq \eta$. Then $(\mu_{\mathbf{t}}, \nu) \preceq (\mu_{\eta \mathbf{1}}, \nu)$, and the $\mathbf{t}$ maximizing Equation (101) is $\mathbf{t}^\star = \eta \mathbf{1}$ and $\tau_f^\star(\varepsilon, \infty) = D_f(\mu_{\eta \mathbf{1}} \| \nu)$ for every generator $f$. In particular*

$$\tau^\star_{\mathrm{R},\alpha}(\varepsilon, \infty) = m \kappa_\alpha(\eta), \tag{104}$$

*and $\tau^\star_{\mathrm{HS},\beta}(\varepsilon, \infty)$ is obtained by convolving $m$ copies of the distribution in Equation (58) at $t = \eta$ and evaluating Equation (14).*

*Proof.* The constraint $\|\mathbf{t}\|_\infty \leq \eta$ is exactly $\mathbf{t} \leq \eta \mathbf{1}$ coordinate-wise. By Lemma G.4, we have $(\mu_{\mathbf{t}}, \nu) \preceq (\mu_{\eta \mathbf{1}}, \nu)$. Equation (104) follows by evaluating Equation (59) at $\mathbf{t} = \eta \mathbf{1}$. □

### $\ell_1$ constraints

**Theorem G.8** (Worst-case $\ell_1$ Laplace parameter)**.** *Let $\mathbf{t} \in \mathbb{R}^m_{\geq 0}$ with $\|\mathbf{t}\|_1 \leq \eta$. Then, $(\mu_{\mathbf{t}}, \nu) \preceq (\mathrm{Lap}(b\eta, b), \mathrm{Lap}(0, b))$. In particular, the $\mathbf{t}$ maximizing Equation (101) is $\mathbf{t}^\star = \eta e_1$ and, for every $\beta \geq 1$,*

$$\tau^\star_{\mathrm{HS},\beta}(\varepsilon, 1) = \left(1 - \sqrt{\beta} e^{-\varepsilon/2b}\right)_+, \qquad \tau^\star_{\mathrm{R},\alpha}(\varepsilon, 1) = \kappa_\alpha(\eta), \tag{105}$$

*with $\beta < 1$ determined by Lemma B.4.*

*Proof.* Suppose first that $\|\mathbf{t}\|_1 = \eta$ exactly. Then $\mathbf{t} \prec \eta e_1$, since the two have equal sums and the largest partial sum of $\eta e_1$ is $\eta$, which dominates every partial sum of $\mathbf{t}$. Lemma G.6 therefore gives $(\mu_{\mathbf{t}}, \nu) \preceq (\mu_{\eta e_1}, \nu)$. If instead $\|\mathbf{t}\|_1 = \eta' < \eta$, apply the above to obtain $(\mu_{\mathbf{t}}, \nu) \preceq (\mu_{\eta' e_1}, \nu)$ and then Lemma G.4 in the single coordinate to obtain $(\mu_{\eta' e_1}, \nu) \preceq (\mu_{\eta e_1}, \nu)$. Since $\mu_{\eta e_1}$ and $\nu$ agree in all coordinates but the first, canceling common factors (Lemma B.2(iii)) identifies the pair $(\mu_{\eta e_1}, \nu)$ with the scalar pair $(\mathrm{Lap}(b\eta, b), \mathrm{Lap}(0, b))$, proving the first claim. Equation (105) then follows from Equations (59) and (61) evaluated at $\mathbf{t} = \eta e_1$. □

**$\ell_p$ constraints** For $1 < p < \infty$, no single $\mathbf{t}$ maximizes the Hockey-stick divergences of all orders simultaneously. We instead bound every admissible $\mathbf{t}$ by a dominating parameter, which concentrates the perturbation into as few coordinates as the budget allows.

**Theorem G.9** (Dominating $\ell_p$ Laplace parameter)**.** *Let $1 < p < \infty$ and $\mathbf{t} \in \mathbb{R}^m_{\geq 0}$ with $\|\mathbf{t}\|_p \leq \eta$. Let $K \coloneqq \lceil m^{1-1/p} \rceil$ and $\mathbf{t}^\star \coloneqq \eta \, (\mathbf{1}_K, \mathbf{0}_{m-K})$. Then $(\mu_{\mathbf{t}}, \nu) \preceq (\mu_{\mathbf{t}^\star}, \nu)$. In particular, $\tau_f^\star(\varepsilon, p) \leq D_f(\mu_{\mathbf{t}^\star} \| \nu)$ for every generator $f$, and the right-hand side is obtained by convolving $K$ copies of the distribution in Equation (58) at $t = \eta$. For $p = 2$, $K = \lceil \sqrt{m} \rceil$.*

*Proof.* Every coordinate satisfies $\mathbf{t}_j \leq \|\mathbf{t}\|_p \leq \eta$, and by Hölder's inequality $\|\mathbf{t}\|_1 \leq m^{1-1/p} \|\mathbf{t}\|_p \leq K\eta$. Listing $\mathbf{t}$ in nonincreasing order, the sum of its $j$ largest coordinates is therefore at most $j\eta$ for $j \leq K$ and at most $K\eta$ for $j > K$. These are exactly the prefix sums of $\mathbf{t}^\star$, so $\mathbf{t} \prec_w \mathbf{t}^\star$. As a property of weak majorization, there exists $\mathbf{u}$ with $\mathbf{t} \leq \mathbf{u}$ coordinate-wise and $\mathbf{u} \prec \mathbf{t}^\star$. Lemma G.4 gives $(\mu_{\mathbf{t}}, \nu) \preceq (\mu_{\mathbf{u}}, \nu)$, and Lemma G.6 gives $(\mu_{\mathbf{u}}, \nu) \preceq (\mu_{\mathbf{t}^\star}, \nu)$. □

The bound of Theorem G.9 is sound but loose, since $\mathbf{t}^\star$ has $\ell_p$ norm $K^{1/p}\eta$, which exceeds $\eta$ whenever $K > 1$.