How Many Traces Suffice? PAC Guarantees for Profiled and Non-profiled Side-Channel Analysis

Seyedmohammad Nouraniboosjin and Fatemeh Ganji

Worcester Polytechnic Institute

Abstract. Side-channel analysis (SCA) is commonly evaluated by reporting the number of traces required to reduce the rank of the correct key. Still, such evaluations remain empirical and do not explain how many traces suffice for reliable recovery, how profiling and attack data contribute separately, or when additional traces cannot overcome weak key distinguishability. We address these questions through a Probably Approximately Correct (PAC) formulation of profiled and non-profiled SCA. Our framework treats candidate-key scores as the common cryptanalytic object and separates finite-sample estimation from the intrinsic separation between the correct key and competing hypotheses. This distinction enables confidence guarantees for key rank and helps determine whether an attack failure is due to insufficient data or an inherently weak attack score. We instantiate the framework with representative profiled and non-profiled attacks chosen for their analytical tractability. Experiments on ASCAD-f and ASCAD-r show that this analytical tractability does not come at the cost of impractical attack performance. The profiled attack achieves exact recovery with tens of attack traces, whereas the non-profiled single-attack rank certificate guarantees exact recovery with about 1,000 traces. These results are competitive with recent ASCAD attacks and, in the non-profiled setting, substantially below the smallest trace counts identified in prior studies, while additionally providing finite-sample guarantees on key rank. More generally, the same finite-sample rank analysis can be adapted to other learners and distinguishers by deriving the corresponding score-gap guarantees. Overall, the framework turns trace complexity from an empirical attack observation into a reusable finite-sample criterion for key recovery.

Keywords: Side Channel Analysis · Probably Approximately Correct Learning · Non-profiled Attacks · Profiled Attacks.

1 Introduction

Physical implementations of cryptographic algorithms disclose secret-dependent information through observable behavior, including power consumption and electromagnetic emanations [1, 2]. Side-channel analysis (SCA) employs such measurements to distinguish candidate values of a secret key [1–3]. In a profiled attack, the adversary first learns a statistical leakage model from labeled measurements collected under known keys [2, 3]. In a non-profiled attack, the adversary

operates without a labeled profiling set and directly evaluates key-dependent leakage predictions, relations, or partitions on measurements collected from the target device [2, 4]. Despite this difference, both settings ultimately assign scores to candidate keys and rank the correct key according to the resulting score vector [2, 5].

Current evaluations report this outcome through key rank, guessing entropy (average rank over the attacks), success rate, and/or the number of traces required to reach a selected metric value [2, 5, 6]. These metrics summarize attacks' performance on sampled profiling and attack sets, and their empirical estimates may vary across repeated experiments [5, 6]. By themselves, they do not provide an accuracy-confidence statement specifying how many traces suffice to preserve a key ordering with probability at least $1 - \delta$ [7, 8]. The distinction is particularly important in profiled SCA. A classifier is trained from N_p profiling traces, while the resulting probability estimates are accumulated over a separate set of N_a attack traces to rank key hypotheses [3, 7]. Predictive performance on individual traces, as well as the confidence or uncertainty of the resulting probabilities, need not coincide with GE or other cryptanalytic metrics computed after score accumulation [9–12]. Consequently, a useful leakage model may still fail to recover the key in a particular execution when the online attack set is too small to stabilize the accumulated scores [2, 5]. In the non-profiled setting, the same finite-sample issue arises without a trained classifier: a key-dependent distinguisher is estimated from finitely many attack traces, and its finite-sample value may deviate from its true value [4, 13].

Recent profiled SCA has developed this score-based pipeline along several complementary directions [14, 15]. Early NN-enhanced profiling studies established deep models as practical alternatives to classical templates, while ASCAD supplied masked-AES datasets and a reproducible posterior-based key-ranking workflow [16–18]. Subsequent studies examined convolutional NN (CNN) architecture design and replication, showing that attack performance depends on architecture, preprocessing, and experimental choices rather than on a universally dominant network [19–21]. Denoising and differentiable preprocessing, ensembles, explicit feature selection, and regularization have been investigated to improve robustness or diagnose overfitting [22–26]. Other works have aligned learning more closely with cryptanalytic evaluation through ranking losses, attack-oriented hyperparameter search, leakage-assessment workflows, learned features for template attacks, label-aware models, and interpretable neural rules [27–33]. [11] provided the first dedicated treatment of post-hoc probability calibration in NN-assisted SCA through temperature scaling, and later extended this line to predictive-uncertainty estimation and its relationship with key rank [12].

Non-profiled learning-assisted SCA follows a different direction from classical DPA and CPA [1, 34]. Timon [4] trains a model under each key hypothesis, SCAUL [35] learns an unsupervised representation before statistical discrimination, and horizontal approaches create learning samples from repeated operations within one or a few traces [36–38]. Deep-learning-assisted collision analysis instead uses a learned relation detector within a non-profiled collision attack [39].

Surveys spanning both settings document the dominance of profiled AES experiments and treat acquisition, preprocessing, model construction, and key-recovery evaluation as parts of one attack pipeline [14, 15]. Taken together, this literature broadens attack design and evaluation, while the finite-sample accuracy-confidence question addressed here remains distinct from architecture, preprocessing, calibration, uncertainty, or attack-metric design [7, 14, 15, 11, 12].

Probably Approximately Correct (PAC) learning provides a natural language for finite-sample statements of this form [40, 41]. It makes the data distribution, hypothesis class, accuracy parameter, confidence parameter, and sample size explicit [40, 41]. Nevertheless, statistical learning theory has been used only selectively in analyses of side-channel key recovery [7, 42, 8, 43]. Among the prior works discussed here, the closest direct use appears in Masure et al. [7]. That work recalls a standard consistency result and states that, for a hypothesis class of finite Vapnik–Chervonenkis (VC) dimension, the finite-sample negative log-likelihood converges uniformly in probability to the true cross-entropy as the number of profiling traces tends to infinity [7]. Its SCA-specific contribution then relates cross-entropy to perceived information [7]. The result justifies negative-log-likelihood minimization asymptotically, but it does not derive an explicit finite-N_p accuracy-confidence bound for the complete data-dependent profiling pipeline or translate leakage-model error into a probability of correct-key rank recovery [7]. Zaid later develops VC-based uniform-convergence arguments in the broader study of deep profiled SCA, but does not formulate the complete key-recovery procedure as a PAC key-rank problem [43].

Other theoretical studies address related but different questions. Bronchain et al. bound errors between estimated leakage models and information-theoretic leakage quantities [44]. Masure et al. analyze information bounds and convergence rates for profiled side-channel evaluators and identify parameters that influence profiling complexity [8]. Picek et al. use an accuracy-confidence viewpoint based on expectation estimation to study the robustness of profiled attacks under perturbations, but their framework does not learn over a key-indexed hypothesis class or derive a finite-sample key-rank guarantee [42]. Van der Valk and Picek adapt bias-variance analysis to GE and emphasize that classifier behavior and cryptanalytic performance need not coincide [10]. Taken together, these works provide information bounds, convergence analyses, robustness bounds, or diagnostic decompositions [44, 8, 42, 10]. They do not establish the complete finite-sample proof chain considered here, from a side-channel learning or estimation problem to a high-probability bound on the rank of the correct key [7, 44, 8, 42, 10, 43].

The missing step is therefore not another isolated application of a VC-dimension or concentration bound [7, 8]. The missing step is a finite-sample formulation whose output is the cryptanalytic quantity used by the attack [2, 5]. More precisely, the problem is to determine when finite profiling and attack samples are sufficient to control the ordering of candidate-key scores and to express the result as an average-rank, top-r, or exact-recovery guarantee [2, 6]. This requires separating finite-sample variability from the score separation

induced by the leakage distribution, targeted intermediate, and selected attack method. More traces can reduce finite-sample variability, but they cannot make an attack informative when the correct-key score is not favorably separated from competing keys.

Key-score formulation. We make the candidate-key score the common cryptanalytic objective of the analysis, while using the statistical gap appropriate to each attack regime. The profiled analysis compares the correct-key log score with each wrong-key log score across independent attack traces, whereas the non-profiled analysis uses complete MMD score gaps and, for the single-attack top-r certificate, their ordered rank-boundary gaps. The common high-level path is therefore finite traces $\rightarrow$ candidate-key scores $\rightarrow$ score gaps $\rightarrow$ rank guarantees, without forcing the two attack settings into the same concentration argument.

Our approach. For profiled SCA, we instantiate the framework with Random-Forest feature selection followed by a regularized Gaussian-kernel Support Vector Classifier [45–49]. The SVC is not selected because it must empirically outperform every alternative classifier, but because its bounded kernel, explicit RKHS norm, margin regularization, and convex training provide a direct route to finite-sample capacity control [48, 50, 49]. Feature selection, kernel selection, SVC training, and probability calibration are treated as one data-dependent profiling pipeline rather than as a fixed model chosen independently of the profiling data. A uniform clipped-log-loss bound controls the complete post-selection calibrated predictor class, and strict propriety of the log loss transfers this guarantee to the expected pairwise key-score margins [50–53]. Conditioned on the learned model, the attack-stage score difference for each wrong key is the average of independent, bounded per-trace contributions. Hence, a variance-sensitive pairwise Bernstein inequality gives an individual wrong-key crossing probability [54]. Summing these pairwise probabilities yields the guessing-entropy bound, while the same probabilities give a high-probability rank bound for a single attack realization. Section 3 therefore separates the profiling budget N_p from the attack budget N_a and connects the complete learning pipeline to key recovery.

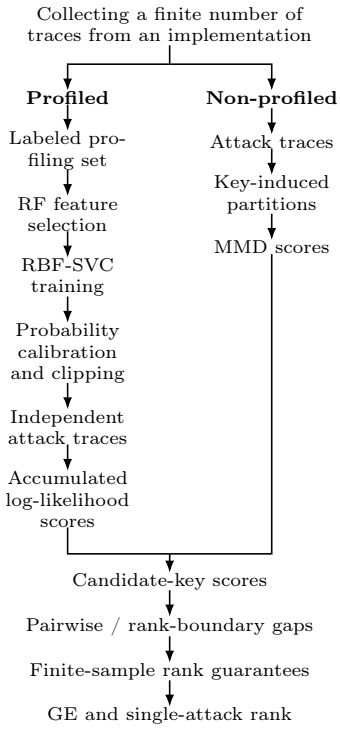


Fig. 1: High-level flow of the profiled and non-profiled analyses.

For non-profiled SCA, no labeled profiling set or learned leakage predictor is available [2, 4]. To the best of our knowledge, [55] first introduced MMD into SCA as a distributional-separability metric for explaining attack hardness in profiled NN-assisted analysis. We use MMD in a different role, where each candidate

key induces hypothetical leakage classes in the attack traces, and the resulting distributional discrepancy serves as the candidate-key score, as formalized in Section 4. MMD provides a nonparametric comparison of these key-induced distributions, but its empirical form is a quadratic U/V-statistic rather than an average of independent per-trace terms [13]. Accordingly, the independent-sum Bernstein argument used for the profiled attack is not assumption-matched to the complete MMD score, and U-statistic Bernstein bounds require kernel- or projection-level variance information and depend on whether the statistic is in a non-degenerate or canonical regime [56–58, 13]. A Gaussian model for the complete MMD gap would likewise impose an additional distributional approximation that need not hold uniformly across key comparisons [13]. We therefore apply Cantelli's one-sided inequality directly to the complete pairwise MMD gap, which requires only its finite mean and variance and does not assume Gaussianity or a particular statistic regime [59, 60]. The individual pairwise probabilities yield an upper bound on GE, while the single-attack top-r certificate is handled at the rank boundary by applying Cantelli to the ordered MMD gap. This makes the choice of the Cantelli bound assumption-driven rather than selected for numerical fit.

The two instantiations also clarify the scope of generalization across attack methods. The proof structure is reusable, but the resulting trace requirement is not independent of the selected learner or distinguisher. Replacing the SVC requires a new capacity or generalization bound for the resulting probabilistic predictor class and a corresponding transfer from probability-estimation error to key scores [50, 52, 53]. A neural-network instantiation is possible in principle, but available generalization bounds depend on quantities such as network architecture, products of layer spectral norms, margins, and finite-sample error [61]. A probabilistic neural key-ranking instantiation would additionally require a calibration-to-score analysis analogous to the one developed in Section 3. Replacing MMD requires finite-sample control of the score-gap fluctuations of the new key-dependent distinguisher. Thus, SVC and MMD do not provide universally smaller trace requirements; they provide analytically tractable instantiations in which the path from finite traces to key rank can be stated explicitly.

Our contributions. We introduce a common PAC formulation of key-ranking SCA across profiled and non-profiled attack settings; see Table 1 for a glance over bounds. The formulation uses candidate-key scores as the shared cryptanalytic object but allows each attack regime to use the score gap and probabilistic tool matched to its estimator. This separates finite-sample variability, correct-key separation, and the resulting rank without imposing one concentration argument on both settings. Notably, the finite-sample guarantees do not appear to come at the cost of an impractical attack. In fact, the trace requirements are competitive with, and, in the non-profiled experiments, substantially below, the smallest trace counts identified in prior studies on the ASCAD dataset [18], a commonly applied benchmark, while additionally providing explicit confidence guarantees on key rank.

– We distinguish average-rank and single-attack rank PAC guarantees explicitly. GE is bounded from the individual key-crossing probabilities, whereas single-attack rank is treated through the rank-tail event appropriate to each setting. This distinction keeps the theoretical quantities aligned with the cryptanalytic metrics reported in the evaluation.
– We provide a learning-theoretic analysis of profiled SCA that separates the contribution of profiling data from that of attack traces, accounts for the full data-dependent learning pipeline, and transfers predictor generalization to correct-key versus wrong-key score margins. This yields finite-sample guarantees on both guessing entropy and high-probability key rank, rather than stopping at classification accuracy alone.
– We derive PAC-style finite-sample guarantees for a non-profiled MMD attack. We derive finite-sample guarantees for both average and single-attack top-r key rank directly from MMD score gaps, without relying on Gaussian or regime-specific statistic assumptions. The construction avoids Gaussian gap assumptions and does not require selecting a non-degenerate or degenerate U-statistic approximation.
– We evaluate the framework by varying profiling and attack budgets separately and by reporting both average-rank and single-attack rank behavior on ASCAD-f [62] and ASCAD-r [63]. The experiments systematically analyze the score separations and rank quantities that appear in the corresponding profiled and non-profiled bounds.

2 Background

Notation. The trace, public-input, leakage-class, and key spaces are denoted by $\mathcal{X}$, $\mathcal{P}$, $\mathcal{Y}$, and $\mathcal{K}$, respectively, and $M = |\mathcal{K}|$. Uppercase letters denote random variables and lowercase letters denote their realizations. Thus, $X \in \mathcal{X} \subseteq \mathbb{R}^d$ is a side-channel trace, $P \in \mathcal{P}$ is the corresponding public input, and $Y \in \mathcal{Y}$ is the targeted leakage class. Candidate keys are denoted by $k \in \mathcal{K}$ and the correct key by $k^\star$. For each key hypothesis, $c_k : \mathcal{P} \to \mathcal{Y}$ denotes the leakage class induced by the public input under k. The exact form of c_k is attack-dependent and is specified locally in the profiled and non-profiled methodologies. The expectation operator is denoted by $\mathbb{E}[\cdot]$. A positive-definite kernel is denoted by κ, its reproducing kernel Hilbert space by $\mathcal{H}_\kappa$, and the kernel mean embedding of a distribution ν by μ_ν. Method-specific symbols are introduced only where they are needed.

2.1 Side-Channel Analysis and Key Rank

Side-channel analysis (SCA) exploits physical observations collected during the execution of a cryptographic implementation to recover secret-dependent information [18]. The observations may arise from power consumption, electromagnetic emanations, timing, or other physical effects. In a profiled attack, the adversary first uses labeled traces from a profiling device to learn a model of the

leakage and then applies the learned model to traces from the target device [18]. In the supervised formulation used in this work, the learned model estimates the class posterior $\eta_c(x) = \Pr[Y = c \mid X = x]$ for $c \in \mathcal{Y}$, and a candidate key selects the class $c_k(p)$ associated with each public input. In a non-profiled attack, no classifier is learned from a separate profiling set, and candidate keys are scored directly from the attack traces through a key-dependent distinguisher.

Both settings eventually assign a score $S_{N_a}(k)$ to every candidate key after N_a attack traces. Higher scores are taken to indicate more likely key hypotheses. Using a one-based pessimistic convention, the correct-key rank is

$$R_{N_a} = 1 + \sum_{k \neq k^\star} \mathbf{1}[S_{N_a}(k) \geq S_{N_a}(k^\star)] \,. \tag{1}$$

Ties are therefore placed before the correct key, and $R_{N_a} = 1$ indicates exact recovery. The guessing entropy is the expected correct-key rank at a fixed attack budget, $\mathrm{GE}(N_a) = \mathbb{E}[R_{N_a}]$, and is commonly estimated by averaging ranks over independent attacks performed at the same N_a [2, 16]. Averaging internal score components or averaging ranks across different attack budgets does not define GE.

2.2 Probably Approximately Correct Learning

Probably Approximately Correct (PAC) learning studies how well a hypothesis learned from finitely many independent samples generalizes to unseen data [41, 64]. Let $\mathcal{F}$ denote a hypothesis class and let $\mathcal{R}(h)$ denote the true risk of $h \in \mathcal{F}$ under the relevant loss. A PAC guarantee controls this risk, or its excess over an appropriate reference predictor, with probability at least $1 - \delta$, where δ is the allowed failure probability and ε denotes the desired accuracy. The number of samples needed to obtain a prescribed (ε, δ) guarantee is the sample complexity, and it depends on the complexity of the hypothesis class in addition to the target accuracy and confidence [41, 64]. For key recovery, ϵ represents the tolerated deviation of the finite-sample correct-key rank R_{N_a} from its true rank $R^\star$, while δ bounds the probability of exceeding this deviation: $\Pr\left[R_{N_a} - R^\star \leq \epsilon\right] \geq 1 - \delta$. In Section 3, these principles are used to control the learned profiled predictor and its induced key scores. In Section 4, the same finite-sample viewpoint is applied directly to the random MMD key-score gaps rather than to a learned classifier.

2.3 Maximum Mean Discrepancy

Maximum Mean Discrepancy (MMD) is a kernel-based measure of the difference between probability distributions [13, 65]. For a positive-definite kernel κ with RKHS $\mathcal{H}_\kappa$, a distribution ν is represented by its kernel mean embedding $\mu_\nu = \mathbb{E}_{X \sim \nu}[\kappa(X, \cdot)]$. The MMD between distributions ν and ν' is

$$\mathrm{MMD}_{\mathcal{H}_\kappa}(\nu, \nu') = \|\mu_\nu - \mu_{\nu'}\|_{\mathcal{H}_\kappa}. \tag{2}$$

For characteristic kernels, MMD is zero if and only if the two distributions coincide [13, 66]. In practice, MMD is estimated from finite samples through pairwise kernel evaluations, giving a quadratic U/V-statistic-type structure rather than an average of independent per-trace terms [13]. We use Gaussian radial-basis kernels of the form $\kappa(x, x') = \exp(-\gamma\|x - x'\|_2^2)$. Such kernels are characteristic and bounded by one, properties used by the later MMD analysis [13, 66].

3 Learning-Theoretic Guarantees for the Profiled SCA

The profiled setting is formulated as supervised learning of the leakage class Y from the trace X, rather than as direct learning of the final key rank. This preserves the standard i.i.d. learning setting, while key rank remains a downstream function of the learned class probabilities, the public inputs, the attack budget, and the competing key hypotheses. We use Random Forest (RF) feature selection, followed by a regularized Gaussian RBF Support Vector Classifier (SVC) and probability calibration. The RF stage restricts the trace representation using a supervised multivariate criterion [46, 47], while the RBF SVC provides a nonlinear hypothesis class with norm- and margin-based capacity control [48, 50, 49]. The calibrated class probabilities retain the relative evidence assigned to all leakage classes and can therefore be accumulated across attack traces. The analysis below first controls the complete data-dependent profiled learner and then transfers that guarantee to pairwise key-score margins and finite-attack rank.

3.1 Profiled Learning Model

Let $\mathcal{T}_p = \{(x_i, p_i, y_i)\}_{i=1}^{N_p}$ denote the labeled profiling dataset, where y_i is the targeted leakage class derived from the known profiling input and key. For the AES S-box target, $\mathcal{Y} = \{0, \ldots, 255\}$ and $c_k(p) = \mathrm{SBox}(p \oplus k)$. We assume that the profiling examples are i.i.d. from an unknown joint distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, with class posterior $\eta_c(x) = \Pr[Y = c \mid X = x]$ as defined in Section 2. A training subset of size N_{train} is denoted by $\mathcal{T}_{\mathrm{train}}$. All feature selection, normalization, bandwidth selection, SVC training, and probability calibration are performed using profiling data only. The attack sample is $\mathcal{T}_a = \{(x_j, p_j)\}_{j=1}^{N_a}$ and is not used in any profiling step.

3.2 Training-Only Random-Forest Feature Selection

The first stage of the learner trains a Random Forest on $\mathcal{T}_{\mathrm{train}}$ following the construction of Breiman [45]. Each tree uses the entropy criterion to select splits. Thus, at each split, the selected coordinate is chosen to reduce class entropy, equivalently to increase information gain in the sense of entropy-based decision-tree induction [67].

Let $I_{\mathcal{T}_{\mathrm{train}}}(t)$ denote the feature-importance score assigned to time sample $t \in \{1, \ldots, d\}$ by the trained Random Forest. For a fixed number of retained

samples q, the selected feature set is $J(\mathcal{T}_{\text{train}}) = \text{Top}_q\left(\{I_{\mathcal{T}_{\text{train}}}(t)\}_{t=1}^{d}\right)$. This induces the restricted trace representation $x_J = x[J(\mathcal{T}_{\text{train}})] \in \mathbb{R}^q$.

The selector is data-dependent and is computed from the same labeled sample used to train the SVC. Consequently, $J(\mathcal{T}_{\text{train}})$ and the induced kernel cannot be treated as fixed independently of the profiling sample. Let $\mathcal{J}_q = \{J \subseteq \{1, \ldots, d\} : |J| = q\}$ denote the family of all feature subsets of size q, and let $\mathcal{G}$ denote the finite set of admissible kernel bandwidth parameters. We analyze the complete learner over the union class $\mathcal{H}_{\text{sel}} = \bigcup_{J \in \mathcal{J}_q} \bigcup_{\gamma \in \mathcal{G}} \mathcal{H}_{\kappa_{J,\gamma,s}}$, This class accounts for the fact that standardization, feature and bandwidth selections, and SVC training use the same profiling dataset.

Remark 1 (Role of RF in the proof). The Random Forest is not the object for which we claim a new PAC guarantee. It is a supervised preprocessing step that defines a data-dependent restriction of the trace representation. The effect of this data-dependent choice is accounted for by the union class $\mathcal{H}_{\text{sel}}$ and the corresponding selection-complexity term in Theorem 1. A fixed-kernel bound applied only after observing the selected feature set would not by itself justify post-selection generalization.

3.3 Selected-Feature RBF Kernel Construction

On the selected feature space, we define the selected-feature Gaussian RBF kernel

$$\kappa_J(x, x') = \exp\left(-\gamma \|x_J - x'_J\|_2^2\right). \tag{3}$$

For the post-selection analysis, κ_J denotes the realized member of the family $\{\kappa_{J,\gamma} : J \in \mathcal{J}_q, \gamma \in \mathcal{G}\}$ selected from the profiling sample.

The SVC learns its class-score functions in $\mathcal{H}_{\kappa_J}$. Since $\kappa_J(x, x) = 1$ for every $x \in \mathcal{X}$, the selected-feature RBF kernel is uniformly bounded.

3.4 SVC Hypothesis Class

The SVC learns a multiclass hypothesis $h \in \mathcal{F}$, where $h : \mathcal{X} \to \{0, \ldots, 255\}$. For the analysis, the hypothesis is represented through class-score functions $f_0, \ldots, f_{255} \in \mathcal{H}_{\kappa_J}$. The predicted label is $h(x) = \arg\max_{c \in \{0,\ldots,255\}} f_c(x)$.

For a labeled example (x, y), define the multiclass margin

$$m_f(x, y) = f_y(x) - \max_{c \neq y} f_c(x). \tag{4}$$

The classification risk is $\mathcal{R}(h) = \Pr_{(X,Y)\sim\mathcal{D}}[h(X) \neq Y]$. For a margin parameter $\rho > 0$, define the finite-sample margin error $\widehat{\mathcal{R}}_\rho(h) = N_{\text{train}}^{-1} \sum_{(x_i,y_i)\in\mathcal{T}_{\text{train}}} \mathbf{1}[m_f(x_i, y_i) \leq \rho]$.

3.5 Margin-Based Generalization

We now invoke a standard margin-based generalization theorem for kernel classifiers. The statement below follows from established Rademacher-complexity risk bounds for bounded RKHS classes [50] together with their multiclass margin extensions [68].

Theorem 1 (Post-selection RKHS margin generalization). *Let $\mathcal{J}_q$ be the family of all feature subsets of size q, and let $\mathcal{G}$ be a finite set of admissible kernel bandwidth parameters. Assume that $\kappa_{J,\gamma}(x,x) \le 1$ for every $x \in \mathcal{X}$, $J \in \mathcal{J}_q$, and $\gamma \in \mathcal{G}$. Assume also that the learned class-score functions satisfy $f_c \in \mathcal{H}_{\kappa_{J,\gamma}}$ and $\|f_c\|_{\mathcal{H}_{\kappa_{J,\gamma}}} \le B$ for every $c \in \{0,\dots,255\}$. Then, with probability at least $1-\delta$ over the draw of the profiling sample, the SVC hypothesis selected using the same sample for feature selection, bandwidth selection, and SVC training satisfies a bound of the form*

$$\mathcal{R}(h) \le \widehat{\mathcal{R}}_\rho(h) + O\left(\frac{B}{\rho\sqrt{N_{\text{train}}}} + \sqrt{\frac{\log\binom{d}{q} + \log|\mathcal{G}| + \log(1/\delta)}{N_{\text{train}}}}\right), \tag{5}$$

where the hidden constant depends on the multiclass reduction or multiclass surrogate used by the SVC.

Proof. For each fixed pair $(J,\gamma) \in \mathcal{J}_q \times \mathcal{G}$, apply the Rademacher-complexity risk bound for bounded RKHS predictors established by Bartlett and Mendelson [50] with failure probability $\delta/(\binom{d}{q}|\mathcal{G}|)$. For a direct multiclass formulation, use the multiclass Rademacher-complexity margin bound of Cortes, Mohri, and Rostamizadeh [68]; the underlying direct multiclass kernel-machine formulation follows Crammer and Singer [49]. A union bound over the $\binom{d}{q}|\mathcal{G}|$ candidate classes gives an event on which the bound holds simultaneously for every admissible feature set and bandwidth. It therefore holds for the data-dependent pair selected from the same profiling sample.

Theorem 1 is the main PAC-style classification statement for the learned profiled classifier. The term $\log\binom{d}{q}$ accounts for selecting a feature subset after observing the labeled sample, while $\log|\mathcal{G}|$ accounts for selecting the bandwidth from $\mathcal{G}$. If γ is fixed in advance, the latter term is omitted; if it is selected from a continuous family, $\log|\mathcal{G}|$ must be replaced by an appropriate covering-number term. The theorem does not state that RF feature importance identifies the physical leakage points. It states that the complete post-selection kernel classifier has controlled out-of-sample classification error under the stated assumptions.

3.6 Conditional Stability View of the SVC

A complementary interpretation uses algorithmic stability. Regularized finite-sample-risk minimization with a Lipschitz convex surrogate is uniformly stable under the conditions established for Hilbert-space regularization by Bousquet and Elisseeff [69]. Therefore, the difference between finite-sample and true surrogate risk can also be controlled through stability.

Theorem 2 (Conditional stability-based generalization). *Fix the selected feature set J and the kernel κ_J. Suppose the SVC training objective is viewed as regularized finite-sample-risk minimization over $\mathcal{H}_{\kappa_J}$ with a bounded Lipschitz surrogate loss. If the learning algorithm is uniformly stable with stability parameter β, then its generalization gap is controlled with high probability by a term depending on β, N_{train}, and δ.*

Proof. This is the stability-implies-generalization result of Bousquet and Elisseeff [69], applied to the regularized learner over $\mathcal{H}_{\kappa_J}$ after the selected feature set and κ_J have been fixed.

The stability view is not used as a separate experimental claim. It provides a complementary learning-theoretic explanation for why SVC regularization controls sensitivity to individual profiling traces after J and κ_J have been fixed. It does not establish stability of the complete RF-SVC pipeline because changing one profiling trace may change the selected feature set.

3.7 Calibrated Leakage Predictor

For each class $c \in \mathcal{Y}$, let $\widehat{\eta}_c(x)$ denote the probability returned by the calibrated SVC. For candidate key k and public input p, the attack therefore uses the class probability $\widehat{\eta}_{c_k(p)}(x)$. The predictor is learned only from profiling traces and is subsequently frozen during the attack phase. We allow the profiling and attack distributions to differ, as may occur across devices or acquisition conditions, and quantify this mismatch by a shift term ρ_{sh} that captures changes in both the trace distribution and the leakage posterior. The identical-distribution setting corresponds to $\rho_{sh} = 0$.

3.8 Uniform Clipped-Log-Loss Generalization

The SVC decision functions do not directly define class probabilities. The probability calibration stage maps the collection of SVC decision scores to a probability vector in the multiclass simplex. In the implementation considered here, this stage consists of sigmoid calibration of the binary SVC scores followed by multiclass pairwise coupling [70, 71]. Because the calibration parameters are estimated from profiling data, calibration is part of the learned predictor and must be included in the generalization class.

Let $C = 256$, and let $\mathbf{f}(x)$ denote the collection of decision scores produced by the SVC. For each $(J, \gamma) \in \mathcal{J}_q \times \mathcal{G}$, let $\mathcal{F}_{J,\gamma}(B)$ denote the class of admissible multiclass SVC score vectors whose component functions belong to $\mathcal{H}_{\kappa_{J,\gamma}}$ and have RKHS norm at most B. Let $\mathcal{C}_{\text{cal}} = \{C_\theta : \theta \in \Theta\}$ denote the family of admissible calibration maps from the SVC score space to the probability simplex. Let T_τ denote a fixed probability-flooring map into the truncated simplex $\Delta_C^\tau = \{p \in [0,1]^C : \sum_{c=0}^{C-1} p_c = 1,\ p_c \geq \tau\}$, where $0 < \tau < 1/C$. The complete post-selection calibrated predictor class is

$$\mathcal{P}_{\text{sel}}^\tau = \bigcup_{J \in \mathcal{J}_q} \bigcup_{\gamma \in \mathcal{G}} \{T_\tau \circ C_\theta \circ \mathbf{f} : \mathbf{f} \in \mathcal{F}_{J,\gamma}(B), \theta \in \Theta\}. \tag{6}$$

This class contains every predictor that can be produced by RF feature selection, bandwidth selection, regularized SVC training, and probability calibration on the same profiling sample.

For $p \in \mathcal{P}^{\tau}_{\mathrm{sel}}$, define the true and finite-sample clipped log losses as $\mathcal{L}_\tau(p) = \mathbb{E}[-\log p_Y(X)]$ and $\widehat{\mathcal{L}}_\tau(p) = N_{\mathrm{train}}^{-1} \sum_{i=1}^{N_{\mathrm{train}}} -\log p_{y_i}(x_i)$. Because $p_c(x) \geq \tau$, the loss lies in $[0, B_\tau]$, where $B_\tau = \log(1/\tau)$, and is $1/\tau$-Lipschitz in the selected class probability.

Assume that every calibration map in $\mathcal{C}_{\mathrm{cal}}$ is L_{cal}-Lipschitz on the bounded SVC score domain. For $\alpha > 0$, let $\mathcal{N}_{\mathrm{cal}}(\alpha)$ denote the size of an α-cover of $\mathcal{C}_{\mathrm{cal}}$ under the uniform norm on this domain. Define

$$\Delta_{\tau, N_{\mathrm{train}}}(\delta, \alpha) = O\Bigg(\frac{L_{\mathrm{cal}} B \sqrt{C}}{\tau \sqrt{N_{\mathrm{train}}}} + B_\tau \sqrt{\frac{\log \binom{d}{q} + \log |\mathcal{G}| + \log \mathcal{N}_{\mathrm{cal}}(\alpha) + \log(1/\delta)}{N_{\mathrm{train}}}} + \frac{\alpha}{\tau} \Bigg). \tag{7}$$

Theorem 3 (Uniform post-selection clipped-log-loss bound). *Under the preceding bounded-norm, bounded-kernel, and Lipschitz-calibration assumptions, with probability at least $1 - \delta$ over the profiling sample, every $p \in \mathcal{P}^{\tau}_{\mathrm{sel}}$ satisfies $|\mathcal{L}_\tau(p) - \widehat{\mathcal{L}}_\tau(p)| \leq \Delta_{\tau, N_{\mathrm{train}}}(\delta, \alpha)$. In particular, the bound holds for the data-dependent calibrated predictor $\widehat{\eta}$ returned after RF feature selection, SVC training, and probability calibration on the same profiling sample.*

Proof. For a fixed feature set, bandwidth, and calibration map, the Rademacher complexity of the bounded multiclass RKHS score class is of order $B\sqrt{C/N_{\mathrm{train}}}$. The vector-contraction inequality transfers this bound through the L_{cal}-Lipschitz calibration map and the $1/\tau$-Lipschitz clipped log loss [51]. The Rademacher uniform-convergence theorem for bounded losses then controls the true and finite-sample clipped log losses [50]. A union bound over the $\binom{d}{q}$ admissible feature subsets, the bandwidth set $\mathcal{G}$, and an α-cover of the calibration family yields the logarithmic selection terms in Equation (7). Replacing a calibration map by its nearest element in the cover changes the clipped log loss by at most α/τ. Because the resulting event holds uniformly over the entire class in Equation (6), it also holds for the predictor selected from the same profiling sample.

The uniform theorem controls generalization of the complete calibrated predictor, but a bound relative to the Bayes conditional probabilities also contains approximation and algorithmic terms. Define the approximation error as $A_\tau = \inf_{p \in \mathcal{P}^{\tau}_{\mathrm{sel}}} \mathcal{L}_\tau(p) - \mathbb{E}[-\log \eta_Y(X)]$ and define the finite-sample algorithmic gap as $\xi_\tau = \widehat{\mathcal{L}}_\tau(\widehat{\eta}) - \inf_{p \in \mathcal{P}^{\tau}_{\mathrm{sel}}} \widehat{\mathcal{L}}_\tau(p)$. The approximation term accounts for the restriction to RF-selected, RKHS-bounded, calibrated predictors, while ξ_τ accounts for the fact that the RF-SVC-calibration procedure need not globally minimize clipped log loss over the entire class.

Corollary 1 (Uniform excess-log-loss bound). *With probability at least* $1-\delta$ *over the profiling sample, the calibrated predictor satisfies*

$$\mathcal{E}_{\log}(\widehat{\eta}) \le A_\tau + \xi_\tau + 2\Delta_{\tau,N_{\text{train}}}(\delta,\alpha). \tag{8}$$

Proof. Let $p^\star$ be a true-risk minimizer in $\mathcal{P}^\tau_{\text{sel}}$, or an arbitrarily accurate approximate minimizer. On the uniform event of Theorem 3, $\mathcal{L}_\tau(\widehat{\eta}) \le \widehat{\mathcal{L}}_\tau(\widehat{\eta}) + \Delta_{\tau,N_{\text{train}}}(\delta,\alpha)$. By the definition of ξ_τ, $\widehat{\mathcal{L}}_\tau(\widehat{\eta}) \le \widehat{\mathcal{L}}_\tau(p^\star) + \xi_\tau$, and the same uniform event gives $\widehat{\mathcal{L}}_\tau(p^\star) \le \mathcal{L}_\tau(p^\star) + \Delta_{\tau,N_{\text{train}}}(\delta,\alpha)$. Subtracting the Bayes log risk $\mathbb{E}[-\log \eta_Y(X)]$ yields Equation (8).

Corollary 1 establishes the profiling-sample premise required by the proper-loss transfer theorem. In the remainder of the section, we use $\varepsilon_{\log}(\delta,\alpha) = A_\tau + \xi_\tau + 2\Delta_{\tau,N_{\text{train}}}(\delta,\alpha)$.

3.9 Proper-Loss Control of Pairwise Key Margins

Define the cross-entropy excess risk by $\mathcal{E}_{\log}(\widehat{\eta}) = \mathbb{E}[-\log \widehat{\eta}_Y(X)] - \mathbb{E}[-\log \eta_Y(X)]$. For multiclass log loss, this quantity equals the expected conditional Kullback–Leibler divergence and is nonnegative because log loss is strictly proper [53]. Let $\psi_\tau(u) = \log \max\{u,\tau\}$ for a clipping threshold $\tau > 0$. For each key hypothesis k, define the learned and ideal expected attack scores

$$S^\tau_{\widehat{\eta}}(k) = \mathbb{E}_{P,X}\left[\psi_\tau\big(\widehat{\eta}_{c_k(P)}(X)\big)\right], \tag{9}$$

$$S^\tau_{\eta}(k) = \mathbb{E}_{P,X}\left[\psi_\tau\big(\eta_{c_k(P)}(X)\big)\right]. \tag{10}$$

For every wrong key $k \ne k^\star$, define the ideal and learned pairwise margins as $\Delta^\star_k = S^\tau_\eta(k^\star) - S^\tau_\eta(k)$ and $\Delta_k = S^\tau_{\widehat{\eta}}(k^\star) - S^\tau_{\widehat{\eta}}(k)$.

Theorem 4 (Proper-loss transfer to pairwise key margins). *Let* η_p *and* η_a *denote the profiling- and attack-domain class posteriors, respectively, and let* $\rho_{\text{sh}} = 2\,\text{TV}(\mathcal{D}^X_p, \mathcal{D}^X_a) + \mathbb{E}_{X\sim\mathcal{D}^X_a}[\|\eta_p(\cdot \mid X) - \eta_a(\cdot \mid X)\|_1]$. *Then, simultaneously for every key hypothesis* k, $|S^\tau_{\widehat{\eta}}(k) - S^\tau_{\eta_a}(k)| \le (\sqrt{2\mathcal{E}_{\log}(\widehat{\eta})} + \rho_{\text{sh}})/\tau$. *Consequently, every wrong-key pair satisfies* $\Delta_k \ge \Delta^\star_k - 2(\sqrt{2\mathcal{E}_{\log}(\widehat{\eta})} + \rho_{\text{sh}})/\tau$, *where* $\Delta^\star_k$ *is the ideal pairwise margin under the attack distribution.*

Proof. For multiclass log loss under the profiling distribution, $\mathcal{E}_{\log}(\widehat{\eta}) = \mathbb{E}_{X\sim\mathcal{D}^X_p}[\text{KL}(\eta_p(\cdot \mid X)\|\widehat{\eta}(\cdot \mid X))]$ [53]. Pinsker's inequality and Jensen's inequality therefore give $\mathbb{E}_{X\sim\mathcal{D}^X_p}[\|\widehat{\eta}(\cdot \mid X) - \eta_p(\cdot \mid X)\|_1] \le \sqrt{2\mathcal{E}_{\log}(\widehat{\eta})}$; see also [52]. Since the ℓ_1 distance between two probability vectors is at most 2, changing the trace marginal from $\mathcal{D}^X_p$ to $\mathcal{D}^X_a$ contributes at most $2\,\text{TV}(\mathcal{D}^X_p, \mathcal{D}^X_a)$. The triangle inequality then gives $\mathbb{E}_{X\sim\mathcal{D}^X_a}[\|\widehat{\eta}(\cdot \mid X) - \eta_a(\cdot \mid X)\|_1] \le \sqrt{2\mathcal{E}_{\log}(\widehat{\eta})} + \rho_{\text{sh}}$. Since ψ_τ is $1/\tau$-Lipschitz, the expected attack-domain score of every key changes by at most $(\sqrt{2\mathcal{E}_{\log}(\widehat{\eta})} + \rho_{\text{sh}})/\tau$. Subtracting the two score perturbations yields the pairwise margin bound.

The theorem gives the link that a classification-error guarantee alone does not provide. It converts probability-estimation error under a strictly proper loss into uniform control over the expected key scores and hence of every correct-key versus wrong-key margin. Corollary 1 establishes the required profiling-sample premise, while Theorem 1 remains a separate classification-level guarantee.

3.10 Attack as Usual

For each attack trace (x_j, p_j) and candidate key k, the profiled attack uses the normalized accumulated clipped log score

$$\widehat{S}^{\tau}_{N_a}(k) = \frac{1}{N_a} \sum_{j=1}^{N_a} \psi_\tau\big(\widehat{\eta}_{c_k(p_j)}(x_j)\big)\,. \tag{11}$$

Multiplication by N_a gives the usual accumulated log-likelihood score and does not change the key ordering [72]. The correct-key rank R_{N_a} is obtained from Equation (1) using $S_{N_a}(k) = \widehat{S}^{\tau}_{N_a}(k)$.

3.11 Pairwise Bernstein Bounds for the Profiled Attack

The profiling guarantees above control the learned probabilistic predictor, but classification generalization alone does not imply finite-attack key recovery because the cryptanalytic decision depends on accumulated key scores rather than on single-trace classification accuracy [72]. We therefore analyze the correct-key score against each wrong-key score directly. Throughout this subsection, we condition on the trained and calibrated model $\widehat{\eta}$ and assume that the N_a attack traces are independent.

For attack trace (X_i, P_i) and key hypothesis k, let $Q_{i,k} = \max\{\widehat{\eta}_{c_k(P_i)}(X_i), \tau\}$, so that $\widehat{S}^{\tau}_{N_a}(k) = N_a^{-1} \sum_{i=1}^{N_a} \log Q_{i,k}$ as in Equation (11). For every wrong key $k \neq k^\star$, define the per-trace pairwise score difference $Z_i^{(k)} = \log Q_{i,k^\star} - \log Q_{i,k}$. By construction, $\mathbb{E}[Z_i^{(k)}] = \Delta_k$, and let $\sigma_k^2 = \mathrm{Var}(Z_i^{(k)})$. Thus, $\Delta_k > 0$ means that the correct key has a positive expected per-trace score advantage over wrong key k. We retain the individual (Δ_k, σ_k^2) for every wrong key rather than replacing them by global extrema or representative values.

Since $\psi_\tau(u) \in [-B_\tau, 0]$, where $B_\tau = \log(1/\tau)$, we have $Z_i^{(k)} \in [-B_\tau, B_\tau]$. Let b_k be any valid one-sided range satisfying $\Delta_k - Z_i^{(k)} \leq b_k$ almost surely. Clipping guarantees a finite range, and $b_k = B_\tau + \Delta_k \leq 2B_\tau$ is valid whenever $\Delta_k > 0$.

Theorem 5 (Pairwise Bernstein crossing bound). *For every wrong key $k \neq k^\star$ with $\Delta_k > 0$, define*

$$p_k(N_a) = \exp\left(-\frac{N_a \Delta_k^2}{2\sigma_k^2 + \frac{2}{3} b_k \Delta_k}\right). \tag{12}$$

Then $\Pr[\widehat{S}^{\tau}_{N_a}(k) \geq \widehat{S}^{\tau}_{N_a}(k^\star)] \leq p_k(N_a)$. *For $\Delta_k \leq 0$, we set $p_k(N_a) = 1$.*

Proof. Wrong key k outranks or ties the correct key exactly when $N_a^{-1}\sum_{i=1}^{N_a} Z_i^{(k)} \le 0$. Equivalently, this event is $\sum_{i=1}^{N_a}(\Delta_k - Z_i^{(k)}) \ge N_a\Delta_k$. The variables $\Delta_k - Z_i^{(k)}$ are independent, have mean zero and variance σ_k^2, and are bounded above by b_k. The one-sided Bernstein inequality therefore gives Equation (12) [54, 73].

Theorem 5 does not require Gaussian score distributions and uses the variance of the correct-key versus wrong-key score difference directly [54]. Because both scores are evaluated on the same trace, their covariance is already contained in $\sigma_k^2 = \mathrm{Var}(Z_i^{(k)})$. This avoids separately concentrating the two candidate scores and then combining their deviations.

Average-Case Guessing-Entropy Bound Conditioned on the trained model, let $\mathrm{GE}(N_a \mid \widehat{\eta}) = \mathbb{E}[R_{N_a} \mid \widehat{\eta}]$, using the rank convention of Equation (1). Since $R_{N_a} = 1 + \sum_{k\neq k^\star} \mathbf{1}\{\widehat{S}^{\tau}_{N_a}(k) \ge \widehat{S}^{\tau}_{N_a}(k^\star)\}$, linearity of expectation gives the exact identity $\mathrm{GE}(N_a \mid \widehat{\eta}) = 1 + \sum_{k\neq k^\star} \Pr[\widehat{S}^{\tau}_{N_a}(k) \ge \widehat{S}^{\tau}_{N_a}(k^\star) \mid \widehat{\eta}]$. No independence among the different wrong-key crossing events is required for this identity to hold. Combining it with Theorem 5 yields

$$\mathrm{GE}(N_a \mid \widehat{\eta}) \le \min\left\{ M,\, 1 + \sum_{k\neq k^\star} p_k(N_a) \right\}. \tag{13}$$

If the profiling procedure is itself random, taking expectation over the profiling randomness preserves the inequality and gives an unconditional average-case bound.

Single-attack Top-r Rank Certificate A single attack realization produces a rank rather than a guessing-entropy value, so Equation (13) is not a pointwise upper bound on an individual realized rank. We therefore derive the Best-Rank guarantee from the same pairwise Bernstein probabilities. For a fixed N_a, order the $M-1$ pairwise crossing bounds as $p_{(1)}(N_a) \ge p_{(2)}(N_a) \ge \cdots \ge p_{(M-1)}(N_a)$. For $r \in \{1, \ldots, M-1\}$, define

$$q_r(N_a) = \min\left\{ 1,\, \sum_{j=r}^{M-1} p_{(j)}(N_a) \right\}, \qquad q_M(N_a) = 0. \tag{14}$$

Theorem 6 (Pairwise Bernstein Best-Rank guarantee). *For every $r \in \{1, \ldots, M\}$, $\Pr[R_{N_a} > r \mid \widehat{\eta}] \le q_r(N_a)$. Consequently, for any $\delta \in (0,1)$, if $r_\delta(N_a) = \min\{r \in \{1, \ldots, M\} : q_r(N_a) \le \delta\}$, then $\Pr[R_{N_a} \le r_\delta(N_a) \mid \widehat{\eta}] \ge 1-\delta$.*

Proof. Let T_{r-1} contain the $r-1$ wrong keys having the largest pairwise crossing bounds $p_k(N_a)$. If $R_{N_a} > r$, at least r wrong keys outrank or tie the correct key, so at least one such crossing must occur for a key outside T_{r-1}. A union bound over the keys outside T_{r-1} therefore gives $\Pr[R_{N_a} > r \mid \widehat{\eta}] \le \sum_{j=r}^{M-1} p_{(j)}(N_a)$, and clipping this sum at one gives Equation (14). The second statement follows directly from the definition of $r_\delta(N_a)$.

For $r = 1$, Theorem 6 gives $\Pr[R_{N_a} > 1 \mid \widehat{\eta}] \leq \min\{1, \sum_{k \neq k^\star} p_k(N_a)\}$. Hence, $\sum_{k \neq k^\star} p_k(N_a) \leq \delta$ is sufficient for exact recovery with confidence at least $1 - \delta$. The confidence parameter δ is a property of the PAC statement and is not determined by the number of repeated attacks used to estimate average performance.

3.12 Relation to the Profiling Guarantee

The pairwise Bernstein results above control the finite attack budget after the probabilistic model has been trained. The profiling budget enters through the uniform clipped-log-loss guarantee and the proper-loss transfer established earlier. Theorem 4 implies the simultaneous pairwise relation $\Delta_k \geq \Delta_k^\star - 2/\tau(\sqrt{2\varepsilon_{\log}} + \rho_{sh})$ for every $k \neq k^\star$. Thus, the profiling analysis determines whether the learned predictor preserves positive correct-key versus wrong-key margins, while Theorems 5 and 6 determine how many independent attack traces are required for those learned margins to yield average-case or Best-Rank recovery.

Corollary 2 (End-to-end profiled average-case and Best-Rank guarantees). *Let $\delta_p, \delta_a \in (0,1)$, and let ρ_{sh} denote the profiling-to-attack distribution-shift term defined in Theorem 4. With probability at least $1 - \delta_p$ over the profiling sample, Corollary 1 gives $\mathcal{E}_{\log}(\widehat{\eta}) \leq A_\tau + \xi_\tau + 2\Delta_{\tau, N_{\mathrm{train}}}(\delta_p, \alpha)$. Consequently, every learned pairwise margin satisfies $\Delta_k \geq \underline{\Delta}_k$, where*

$$\underline{\Delta}_k = \Delta_k^\star - \frac{2}{\tau}\left(\sqrt{2\left(A_\tau + \xi_\tau + 2\Delta_{\tau, N_{\mathrm{train}}}(\delta_p, \alpha)\right)} + \rho_{\mathrm{sh}}\right).$$

Let $\bar{\sigma}_k^2 \geq \sigma_k^2$ and $\bar{b}_k \geq b_k$ be valid upper bounds, and define $\bar{p}_k(N_a) = 1$ when $\underline{\Delta}_k \leq 0$ and otherwise

$$\bar{p}_k(N_a) = \exp\left(-\frac{N_a \underline{\Delta}_k^2}{2\bar{\sigma}_k^2 + \frac{2}{3}\bar{b}_k \underline{\Delta}_k}\right). \tag{15}$$

On the profiling event, the conditional guessing entropy satisfies $\mathrm{GE}(N_a \mid \widehat{\eta}) \leq \min\{M, 1 + \sum_{k \neq k^\star} \bar{p}_k(N_a)\}$. Let $\bar{q}_r(N_a)$ be obtained from the ordered $\bar{p}_k(N_a)$ as in Equation (14). If $\bar{q}_r(N_a) \leq \delta_a$, then $\Pr[R_{N_a} \leq r] \geq 1 - \delta_p - \delta_a$ over the joint profiling and attack randomness.

Proof. Corollary 1 bounds the profiling-domain excess log loss with probability at least $1 - \delta_p$ over the profiling sample. Theorem 4 transfers this guarantee to the attack domain while accounting for profiling-to-attack distribution shift through ρ_{sh}, yielding the stated lower bound $\underline{\Delta}_k$ simultaneously for all wrong keys. For positive margins, the Bernstein exponent is nondecreasing in the margin and nonincreasing in the variance and range terms, so the conservative quantities $\underline{\Delta}_k$, $\bar{\sigma}_k^2$, and $\bar{b}_k$ give valid upper bounds $\bar{p}_k(N_a)$ on all pairwise crossing probabilities. Equation (13) then gives the conditional guessing-entropy statement. Conditioned on the same profiling event, Theorem 6 gives attack failure probability at most δ_a whenever $\bar{q}_r(N_a) \leq \delta_a$. A union bound over the profiling and attack failure events yields a total failure probability of at most $\delta_p + \delta_a$.

Clipping also provides fully distribution-free fallback bounds $\sigma_k^2 \le B_\tau^2$ and $b_k \le 2B_\tau$, although sharper valid variance and range bounds preserve the principal advantage of the pairwise Bernstein analysis. The profiling and attack budgets therefore enter through different mechanisms and should not be conflated.

3.13 Attack-Trace Complexity

For an individual wrong key with $\Delta_k > 0$, requiring $p_k(N_a) \le \delta_k$ is guaranteed whenever

$$N_a \ge \frac{2\sigma_k^2 + \frac{2}{3} b_k \Delta_k}{\Delta_k^2} \log \frac{1}{\delta_k}, \tag{16}$$

by directly inverting the Bernstein tail probability in Equation (12) [54]. Consequently, the pairwise attack complexity scales as $N_a = O((\sigma_k^2/\Delta_k^2 + b_k/\Delta_k) \log(1/\delta_k))$. For exact recovery with confidence $1-\delta$, the simple allocation $\delta_k = \delta/(M-1)$ for all wrong keys gives a sufficient attack budget obtained by taking the maximum of Equation (16) over $k \ne k^\star$. For the average-case quantity, requiring $p_k(N_a) \le \varepsilon/(M-1)$ for all wrong keys is sufficient to obtain $\mathrm{GE}(N_a \mid \widehat{\eta}) \le 1+\varepsilon$. These are sufficient upper bounds on the required attack budget and do not constitute information-theoretic lower bounds.

4 PAC Guarantees for Non-Profiled SCA

We now describe the finite-sample framework for the non-profiled SCA. No classifier is learned from a separate profiling set; instead, each candidate key induces leakage classes in the attack traces and is scored by the separation between these two classes. The analysis therefore starts with the key scores and uses their pairwise gaps to determine the correct-key rank.

4.1 Attack Model

Let $(P_i, X_i)_{i=1}^{N_a}$ be independent attack observations. For the Hamming-weight (HW) attack, the common key-induced class map is instantiated as $c_k(p) = \mathrm{HW}(\mathrm{SBox}(p \oplus k))$, with $\mathcal{Y} = \{0, \dots, 8\}$. Fix two distinct classes $a, b \in \mathcal{Y}$. For each k, let $\mathcal{D}_{k,a} = \mathcal{L}(X \mid c_k(P) = a)$ and $\mathcal{D}_{k,b} = \mathcal{L}(X \mid c_k(P) = b)$, and assume both conditioning events have nonzero probability. The attack sample induces the trace sets $A_k^{(a)} = \{x_i : c_k(p_i) = a\}$ and $B_k^{(b)} = \{x_i : c_k(p_i) = b\}$, with sizes m_k and n_k.

4.2 Finite-Sample MMD Scores

Definition 1 (True MMD score). *For key hypothesis $k \in \mathcal{K}$, its true score is $s_{a,b}(k) = \mathrm{MMD}_{\mathcal{H}_\kappa}(\mathcal{D}_{k,a}, \mathcal{D}_{k,b})$, using the MMD definition in Equation* (2).

A large $s_{a,b}(k)$ means that the partition induced by k separates the selected leakage classes more strongly in the RKHS. The definition does not assume that the correct key is the unique maximizer. Let $A_k^{(a)} = \{x_{k,a,1}, \ldots, x_{k,a,m_k}\}$ and $B_k^{(b)} = \{x_{k,b,1}, \ldots, x_{k,b,n_k}\}$. The finite-sample mean embeddings are $\widehat{\mu}_{k,a} = m_k^{-1} \sum_{i=1}^{m_k} \kappa(x_{k,a,i}, \cdot)$ and $\widehat{\mu}_{k,b} = n_k^{-1} \sum_{j=1}^{n_k} \kappa(x_{k,b,j}, \cdot)$.

Definition 2 (Finite-sample MMD score). *The finite-sample score of k is $\widehat{s}_{a,b}(k) = \|\widehat{\mu}_{k,a} - \widehat{\mu}_{k,b}\|_{\mathcal{H}_\kappa}$. Equivalently,*

$$\begin{aligned}\widehat{s}_{a,b}^2(k) = \frac{1}{m_k^2} \sum_{i=1}^{m_k} \sum_{j=1}^{m_k} \kappa(x_{k,a,i}, x_{k,a,j}) + \frac{1}{n_k^2} \sum_{i=1}^{n_k} \sum_{j=1}^{n_k} \kappa(x_{k,b,i}, x_{k,b,j}) \\ - \frac{2}{m_k n_k} \sum_{i=1}^{m_k} \sum_{j=1}^{n_k} \kappa(x_{k,a,i}, x_{k,b,j}).\end{aligned} \tag{17}$$

Equation (17) is the biased quadratic MMD estimator of Gretton et al. [13, Equation (5)]. The term "biased" refers to estimation of the squared population MMD and not to the resulting key rank [13]. We assume the bounded kernel introduced in the background, so every finite-sample MMD score and every pairwise gap below has finite variance.

4.3 Why We Bound the Complete MMD Gap

The ordinary pairwise Bernstein argument used in the profiled attack is not assumption-matched to the MMD score. There, the pairwise log-score difference is an average of independent bounded per-trace contributions, whereas the empirical MMD in Equation (17) is a quadratic statistic in which each trace participates in many kernel evaluations [13]. Bernstein inequalities for U-statistics require variance quantities tied to the U-statistic kernel or its Hoeffding projections, rather than the variance of a collection of complete MMD scores inserted into the independent-sum Bernstein formula [56, 57]. The distinction is important because MMD statistics can also enter a canonical or degenerate regime, for which the fluctuation law and concentration analysis differ from the ordinary non-degenerate case [13, 58]. For example, the unbiased squared-MMD statistic is asymptotically Gaussian under a fixed alternative but has an N_a-scaled weighted chi-squared limit under equality of the two distributions [13]. A Gaussian model for the complete key-score gap is therefore a useful approximation only under additional regime-dependent assumptions and cannot provide one distribution-free PAC bound for all key comparisons.

We instead apply Cantelli's one-sided inequality directly to the complete random MMD gap that determines each key crossing. Cantelli requires only a finite mean and variance and does not require Gaussianity, independence among the internal kernel terms, or a choice between degenerate and non-degenerate U-statistic regimes [59, 60]. Moreover, Cantelli is sharp when only the first two moments are specified, so obtaining a uniformly tighter one-sided guarantee requires additional assumptions beyond those moments [60]. This makes Cantelli

a deliberate assumption-matched choice rather than a replacement selected for numerical fit.

4.4 Pairwise Cantelli Bound

For a fixed attack budget N_a and every wrong key $k \neq k^\star$, define the complete pairwise MMD gap

$$D_k(N_a) = \widehat{s}_{a,b}(k^\star) - \widehat{s}_{a,b}(k). \tag{18}$$

A wrong key outranks or ties the correct key exactly when $D_k(N_a) \leq 0$ under the pessimistic ranking convention. Because the biased finite-sample MMD score can have an N_a-dependent expectation, we define the finite-budget mean and variance of the complete gap as $\Delta_k(N_a) = \mathbb{E}[D_k(N_a)]$ and $V_k(N_a) = \mathrm{Var}(D_k(N_a))$. A positive $\Delta_k(N_a)$ means that the correct key has a positive expected finite-sample MMD advantage over wrong key k at attack budget N_a.

Cantelli's one-sided inequality states that a random variable with finite mean μ and variance v satisfies $\Pr[X - \mu \leq -t] \leq v/(v + t^2)$ for every $t > 0$ [59, 60]. Applying it to $D_k(N_a)$ with $t = \Delta_k(N_a)$ yields the following pairwise crossing probability.

Theorem 7 (Pairwise Cantelli crossing bound). *For every wrong key $k \neq k^\star$, define*

$$p_k(N_a) = \begin{cases} \dfrac{V_k(N_a)}{V_k(N_a) + \Delta_k^2(N_a)}, & \Delta_k(N_a) > 0, \\ 1, & \Delta_k(N_a) \leq 0. \end{cases} \tag{19}$$

Then $\Pr[\widehat{s}_{a,b}(k) \geq \widehat{s}_{a,b}(k^\star)] \leq p_k(N_a)$.

Proof. For $\Delta_k(N_a) > 0$, the crossing event is $D_k(N_a) \leq 0$, equivalently $D_k(N_a) - \Delta_k(N_a) \leq -\Delta_k(N_a)$. Cantelli's inequality gives $\Pr[D_k(N_a) \leq 0] \leq V_k(N_a)/(V_k(N_a) + \Delta_k^2(N_a))$. When $\Delta_k(N_a) \leq 0$, Cantelli does not provide a nontrivial upper bound for this event, so the trivial bound 1 is used.

Each wrong key retains its own $\Delta_k(N_a)$ and $V_k(N_a)$. The analysis therefore does not combine a favorable margin from one key with a favorable variance from another key. The quantity $p_k(N_a)$ is itself an upper bound on the pairwise attack failure probability for wrong key k. Hence, for a prescribed pairwise attack failure level $\delta_k \in (0, 1)$, the condition $p_k(N_a) \leq \delta_k$ can be checked directly from the same gap moments.

4.5 Average-Case Bound: Guessing Entropy

Using the common rank convention in Equation (1), the MMD rank is $R_{N_a} = 1 + \sum_{k \neq k^\star} \mathbf{1}[D_k(N_a) \leq 0]$. Taking expectations gives the exact identity $\mathrm{GE}(N_a) = 1 + \sum_{k \neq k^\star} \Pr[D_k(N_a) \leq 0]$. No independence between the $M - 1$ key-crossing events is required for this identity.

Theorem 8 (Pairwise Cantelli upper bound on GE). *Under the conditions of Theorem 7,*

$$\mathrm{GE}(N_a) \le 1 + \sum_{k \neq k^\star} p_k(N_a) \tag{20}$$

with the right-hand side clipped at the trivial maximum M when necessary.

Proof. The rank representation above and linearity of expectation give $\mathrm{GE}(N_a) = 1 + \sum_{k \neq k^\star} \Pr[D_k(N_a) \le 0]$. Applying Theorem 7 to every wrong key proves the result.

No separate attack-stage failure probability δ_a is required for the GE bound. Guessing entropy is already an expectation over the attack randomness, and each pairwise crossing probability contributes directly to that expectation. Cantelli bounds those probabilities by $p_k(N_a)$, which gives Equation (20) without introducing an additional high-probability event. For example, $\sum_{k \neq k^\star} p_k(N_a) \le \varepsilon$ is sufficient for $\mathrm{GE}(N_a) \le 1 + \varepsilon$. An attack-stage parameter δ_a is needed only when the objective is a high-probability statement about the realized rank, such as $\Pr[R_{N_a} \le r] \ge 1 - \delta_a$.

The bound is pairwise rather than uniform over the complete key space. It pays once for each event that actually increases the correct-key rank and preserves the individual mean-to-variance ratio of every wrong-key comparison.

4.6 Single-Attack Top-r Rank Certificate

A single attack realization produces a rank rather than a guessing-entropy value, so Equation (20) is not a pointwise upper bound on an individual realized rank. We therefore derive a high-probability rank guarantee for one complete attack.

For a fixed attack budget N_a, sort the $M-1$ pairwise gaps in increasing order as $D_{(1)}(N_a) \le \cdots \le D_{(M-1)}(N_a)$. Because the pessimistic rank is one plus the number of nonpositive gaps, the equivalence $R_{N_a} > r$ if and only if $D_{(r)}(N_a) \le 0$ is exact. Define $\Delta_r(N_a) = \mathbb{E}[D_{(r)}(N_a)]$ and $V_r(N_a) = \mathrm{Var}(D_{(r)}(N_a))$. Cantelli applied directly to the ordered gap gives, for $r \in \{1, \ldots, M-1\}$,

$$q_r(N_a) = \begin{cases} \dfrac{V_r(N_a)}{V_r(N_a) + \Delta_r^2(N_a)}, & \Delta_r(N_a) > 0, \\ 1, & \Delta_r(N_a) \le 0, \end{cases} \qquad q_M(N_a) = 0. \tag{21}$$

Theorem 9 (Pairwise-gap Cantelli rank guarantee). *For every $r \in \{1, \ldots, M\}$,*

$$\Pr[R_{N_a} > r] \le q_r(N_a). \tag{22}$$

Consequently, for any attack-stage failure probability $\delta_a \in (0, 1)$, define

$$r_{\delta_a}(N_a) = \min\{r \in \{1, \ldots, M\} : q_r(N_a) \le \delta_a\}. \tag{23}$$

Then

$$\Pr[R_{N_a} \le r_{\delta_a}(N_a)] \ge 1 - \delta_a. \tag{24}$$

Proof. For $r \in \{1, \ldots, M-1\}$, the event $R_{N_a} > r$ is equivalent to $D_{(r)}(N_a) \leq 0$. Applying Cantelli's inequality to $D_{(r)}(N_a)$ gives Equation (22). The case $r = M$ is trivial because $q_M(N_a) = 0$. The second statement follows directly from the definition of $r_{\delta_a}(N_a)$.

For $r = 1$, Theorem 9 gives $\Pr[R_{N_a} > 1] \leq q_1(N_a)$. Hence, $q_1(N_a) \leq \delta_a$ is sufficient for exact recovery with confidence at least $1-\delta_a$. As in the profiled case, δ_a is the attack-stage failure probability of the single-attack rank certificate; it is not needed for the GE bound.

4.7 Trace-Complexity Interpretation

For one wrong-key comparison with $\Delta_k(N_a) > 0$, requiring $p_k(N_a) \leq \delta_k$ is equivalent to

$$\Delta_k(N_a) \geq \sqrt{\frac{1-\delta_k}{\delta_k} V_k(N_a)}. \tag{25}$$

Equivalently, $\Delta_k^2(N_a)/V_k(N_a) \geq (1-\delta_k)/\delta_k$. Thus, Cantelli relates the pairwise attack failure level directly to the signal-to-variance ratio of the complete pairwise MMD gap. Unlike the independent-sum Bernstein case, a universal closed-form rate in N_a does not follow without an additional bound on how $V_k(N_a)$ decreases with the attack budget. This is deliberate: imposing an N_a^{-1} variance law would reintroduce the non-degenerate U-statistic assumption that the Cantelli construction is intended to avoid. If a valid problem-specific variance rate is available, it can be substituted into Equation (19) without changing the rank analysis.

4.8 Empirical Evaluation of the Cantelli Bounds

The quantities $\Delta_k(N_a)$ and $V_k(N_a)$ in the previous theorems are moments of the complete random MMD gap. To evaluate the bounds experimentally, suppose the complete attack is repeated independently R times at the same attack budget. Let $D_k^{(u)}(N_a)$ denote the pairwise gap for wrong key k in repetition u. We estimate the required moments by

$$\widehat{\Delta}_k(N_a) = \frac{1}{R}\sum_{u=1}^{R} D_k^{(u)}(N_a), \qquad \widehat{V}_k(N_a) = \frac{1}{R-1}\sum_{u=1}^{R}\left(D_k^{(u)}(N_a) - \widehat{\Delta}_k(N_a)\right)^2. \tag{26}$$

The numerical GE-bound curve is then obtained by evaluating Equation (19) with $\widehat{\Delta}_k(N_a)$ and $\widehat{V}_k(N_a)$ for every wrong key and substituting the resulting values into Equation (20).

To evaluate the single-attack top-r certificate, the $M-1$ gaps are first sorted within each complete attack. For every rank boundary r, the same sample mean and sample variance are computed from $D_{(r)}^{(1)}(N_a), \ldots, D_{(r)}^{(R)}(N_a)$ and substituted into Equation (21). For a chosen δ_a, the certified rank is then the smallest r satisfying $q_r(N_a) \leq \delta_a$, as in Equation (23).

5 Experimental Evaluation

We evaluate the practical feasibility of the two attack instantiations analyzed in Sections 4 and 3. The non-profiled experiments examine the Hamming-weight (HW) class construction, the separation of the correct-key MMD score, and the evolution of GE and best single-realization rank as the attack budget increases. The profiled experiments examine the RF-SVC pipeline under separate profiling and attack budgets and compare the observed ranks with the corresponding pairwise Bernstein bounds. Throughout this section, rank "1" denotes exact recovery. The theoretical statements in Sections 4 and 3 are expressed in terms of plotted bound curves drawn using the statistics available from the experiments, as described below. For the experiments on the bound on the best rank, $\delta = 0.99$, i.e., at least one of the hundred attacks will succeed. The GE/rank curves show the observed attack behavior, while the corresponding bound curves quantify when the measured correct-key separation is sufficiently large relative to finite-sample variability to support a rank guarantee.

5.1 Experimental Setup

We use the fixed-key and random-key variants of the ASCAD database, denoted ASCAD-f and ASCAD-r, respectively [18]. ASCAD-f contains traces with 700 time samples and uses a fixed key during profiling and attack. ASCAD-r contains traces with 1400 time samples, uses random keys during profiling, and uses a fixed key in the attack set. The available profiling sets contain 50,000 traces for ASCAD-f and 200,000 traces for ASCAD-r. The targeted intermediate is the AES S-box output byte. The correct target-key bytes are 224 for ASCAD-f and 34 for ASCAD-r.

The Random Forest and SVC components were implemented using scikit-learn [74]. All experiments were executed on a high-performance computing cluster using four CPU cores and 50 GB of memory per task. The cluster contains Intel Xeon Scalable Gold 6248 and AMD EPYC 7543 processors.

5.2 Profiled RF-SVC Attack

Implementation. We instantiate the profiled pipeline of Section 3 with one RF-SVC model per complete profiling experiment. In our implementation, the 256-class SVC trains one-vs-one binary classifiers, whose pairwise scores are calibrated and coupled to produce a probability distribution over the 256 classes. This yields one calibrated multiclass hypothesis in the class of all such composed one-vs-one SVC predictors obtainable from the training pipeline. For a given profiling budget N_{train}, a new training subset is sampled from the available profiling set. An entropy-based Random Forest is trained on this subset, and the 200 coordinates with the largest feature-importance values are retained. The selected coordinates are standardized using feature-wise means and standard deviations computed from the same training subset. These parameters are

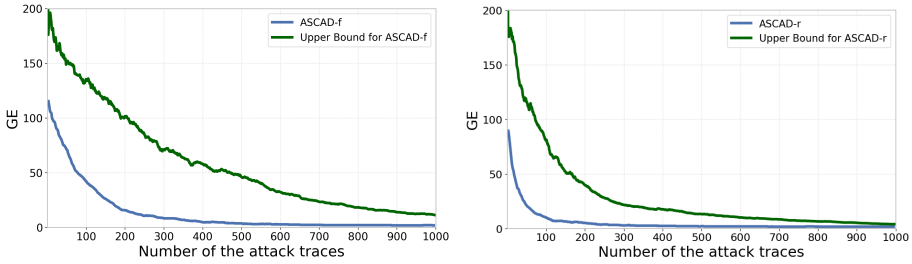
then reused for the attack traces, and no attack trace is used to fit the feature selector, normalization, bandwidth, SVC, or calibration stage.

On the selected representation, we train an RBF-SVC with $C = 1$. The default scale is $\gamma_{\text{scale}} = 1/q$ with $q = 200$. The admissible bandwidth set is $\{0.5\gamma_{\text{scale}}, \gamma_{\text{scale}}, 2\gamma_{\text{scale}}\}$, and bandwidth selection is performed using profiling data only. The selected bandwidth is then fixed for the corresponding trained model and its attack evaluation. The SVC decision scores are converted to multiclass probabilities by sigmoid calibration and pairwise coupling, as described in Section 3.8. Before accumulating the key log-scores, the class probabilities are floored at the fixed threshold τ used in Equation (11).

For the average-case attack-budget experiment, the complete profiling and attack procedure is repeated 30 times. Each repetition resamples the profiling subset, repeats RF selection and SVC training, evaluates the resulting model on attack traces, and records the one-based correct-key rank at each fixed attack budget. The mean rank over these complete repetitions is the GE. The best single-realization rank experiment is kept separate from these repeated attacks and is not averaged into GE.

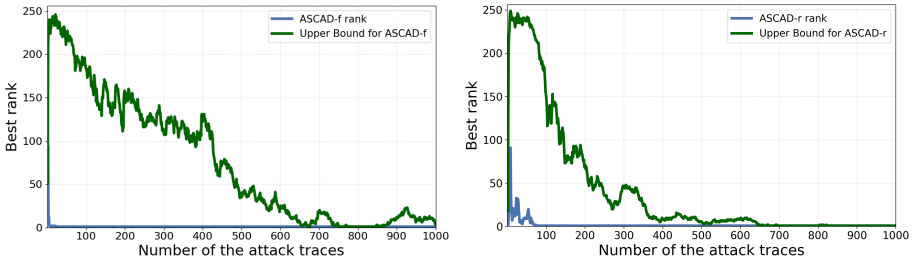
For the plotted pairwise Bernstein curves, the pairwise statistics are recomputed for each attack budget using only the attack traces that contribute to the accumulated score at that budget. For every wrong key, the margin, variance, and one-sided range term are kept pair-specific and are inserted into Equation (12). The average-case curve uses the GE construction of Equation (13), whereas the single-realization curve uses the rank-tail construction of Equation (14) rather than the GE formula. No smoothing or substitution of favorable statistics across different wrong-key pairs is applied. As in the non-profiled plots, these numerical curves evaluate population-bound expressions using experimental estimates of the required pairwise quantities.

Effect of the Attack Budget. Figure 2 reports GE and its pairwise Bernstein upper-bound evaluation as the number of attack traces increases to $N_a = 1{,}000$. The models are trained using the RF-selected 200-feature representations described above. For both datasets, accumulating calibrated class evidence across additional attack traces lowers GE and eventually yields exact recovery. ASCAD-r reaches low rank with fewer attack traces in the reported average-case experiment. Because ASCAD-r provides a larger profiling set than ASCAD-f and differs in both acquisition window and profiling-key regime, this comparison should be interpreted as dataset-specific rather than as an isolated effect of any single design choice. The difference between ASCAD-f and ASCAD-r should also be interpreted in light of their distinct acquisition campaigns (see also Section 5.4). The datasets differ in sampling rate, acquisition window, and filtering conditions, which can directly affect both correct-key score separation and score variability and therefore the resulting finite-sample bounds [18]. The decreasing Bernstein bounds therefore show that, as more attack traces are accumulated, the correct-key pairwise margins increasingly dominate their score variability, reducing the probability that wrong keys overtake the correct key.



(a) ASCAD-f (b) ASCAD-r

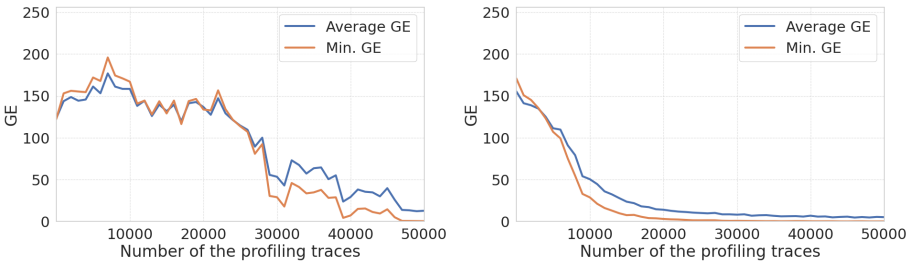
Fig. 2: GE and the pairwise Bernstein GE upper-bound evaluation for the profiled RF-SVC attack as functions of the number of attack traces. Each model uses 200 RF-selected coordinates, calibrated SVC probabilities, and the clipped log-score of Equation (11).



(a) ASCAD-f (b) ASCAD-r

Fig. 3: Best Rank and the corresponding pairwise Bernstein rank upper-bound evaluation for the profiled RF-SVC attack. Rank one denotes exact recovery. ASCAD-f reaches stable rank one after 22 attack traces, whereas ASCAD-r requires 74 traces.

Single-Realization Best Rank. The profiled experiment uses a single complete attack realization rather than an average over the 30 average-case repetitions. Figure 3 therefore plots the realized one-based rank together with the pairwise Bernstein rank bound obtained from Equation (14) at the selected confidence level. The ASCAD-f trajectory reaches rank one and remains there after 22 attack traces, whereas ASCAD-r requires 74 attack traces over the remainder of the evaluated range. For the profiled setting, the observed trace requirements are competitive with the best results reported on ASCAD [55, 33, 24, 30, 75], despite using a comparatively simple analytically tractable learner rather than heavily optimized deep architectures. The gap between the observed rank and its bound reflects the conservatism of the finite-sample guarantee, i.e., empirical recovery may occur earlier, whereas the bound identifies a sufficient trace budget for the stated rank confidence.



(a) ASCAD-f (b) ASCAD-r

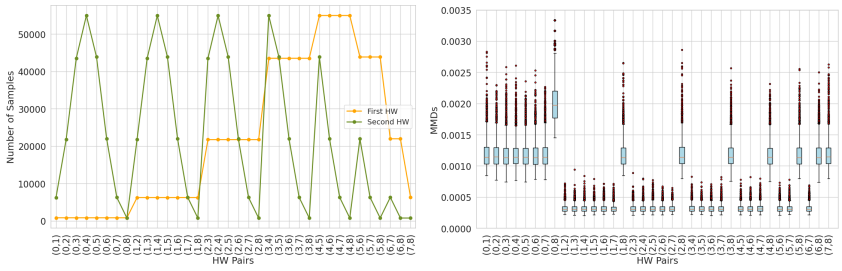
Fig. 4: Average one-based correct-key ranks of the profiled RF-SVC attack as functions of the profiling budget N_{train}, with the attack budget fixed at $N_a = 1{,}000$. Each point summarizes 30 complete repetitions for ASCAD-f and 4 complete repetitions for ASCAD-r.

Effect of the Profiling Budget. To isolate the profiling component of the two-budget analysis, we vary N_{train} while fixing the attack budget at $N_a = 1{,}000$. For each training size, the complete RF selection, normalization, bandwidth selection, SVC training, calibration, and attack procedure are repeated across the available independent profiling experiments. Figure 4 reports the resulting average one-based correct-key rank, i.e., GE at the fixed attack budget. Increasing the profiling budget generally lowers this average rank, indicating that the learned probability scores and their induced key ordering become more reliable as more labeled profiling traces are available. This experiment isolates the profiling-budget effect and does not reuse the attack-budget Bernstein curve as a numerical bound on N_{train}. This behavior is the empirical counterpart of the profiling side of the theory that increasing N_{train} improves the learned probability model and thereby the key-score margins on which the subsequent attack-stage guarantee depends.

5.3 Non-Profiled MMD Attack

Implementation. We instantiate the attack model of Section 4.1 with the HW labels induced by each candidate key and use the biased finite-sample MMD score in Definition 2. The experiments reported here use the single HW pair $(2, 6)$. MMD is computed with the Gaussian kernel $\kappa(x, x') = \exp(-\gamma\|x - x'\|_2^2)$. The Gaussian RBF kernel is consistent with additive-noise leakage models commonly used in SCA [76], but its use here does not impose a Gaussian model on the resulting MMD gaps. The scale parameter is set according to $\gamma = 1/(d\,\mathrm{Var}(X))$, where d is the dimension of the trace representation and the variance is computed for that representation. The same trace representation and bandwidth rule are used for all 256 key hypotheses, so the kernel is not tuned separately for favorable keys.

For ASCAD-f, MMD is computed on all 700 time samples of each trace. For ASCAD-r, the 1400-sample traces are reduced to a fixed set of 200 coordinates



(a) Frequency of HW labels

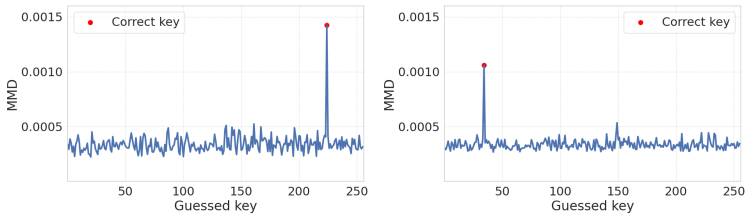
(b) Variability of pairwise MMD estimates

Fig. 5: selection of the HW pair on ASCAD-r [18]. (a) Observed class counts show that the extreme HW classes provide substantially fewer samples. (b) Repeated MMD estimates show greater stability for sufficiently populated class pairs. Based on these two criteria, the subsequent attack uses the pair $(2, 6)$.

before the MMD sweep to reduce the quadratic cost of the kernel computations. The selected ASCAD-r representation is fixed before the evaluated attack subsets and reused for every candidate key. For the score-separation experiment, we use 5,000 traces selected from each selected HW class and repeat the complete scoring procedure 10 times on independently resampled subsets. For every repetition, all 256 MMD scores are computed from the same attack subset before evaluating the correct-key rank.

Selection of the HW Pair. Figure 5 reports the observed HW frequencies and the variability of MMD estimates across candidate class pairs on ASCAD-r. The extreme classes contain substantially fewer traces and exhibit less stable estimates. The pair $(2, 6)$ provides two sufficiently populated classes while retaining visible distributional separation. We therefore fix this pair before evaluating the candidate keys. This is an experimental design choice intended to avoid unstable MMD estimates from sparsely populated classes and is not an additional assumption in the pairwise Cantelli analysis.

Correct-Key Score Separation. For each key candidate, we partition the same attack traces using its hypothetical HW labels and compute the MMD between the induced classes 2 and 6. Figure 6 shows the resulting score vectors. For ASCAD-f, the correct-key score is approximately 0.0016, whereas the strongest wrong-key score is approximately 0.00059. For ASCAD-r, the corresponding values are approximately 0.0013 and 0.00063. Thus, the correct key attains the largest finite-sample MMD score in both datasets and is separated from the strongest wrong-key competitor in these experiments. These observed score separations are the counterparts of the pairwise MMD gaps $D_k(N_a)$ analyzed in Section 4. This separation is the empirical prerequisite for a nontrivial rank guarantee. The Cantelli bound becomes informative only when the correct key has a positive expected score advantage over its competitors.



(a) ASCAD-f (b) ASCAD-r

Fig. 6: Finite-sample MMD scores for all 256 key hypotheses using the HW pair $(2, 6)$ and 5,000 traces from each induced class. The marked point denotes the correct key. ASCAD-f uses the full trace representation, whereas ASCAD-r uses the fixed 200-feature representation described in Section 5.3.

Effect of the Attack Budget. We next vary the number of attack traces from 100 to 5,000. For each attack budget N_a, we draw 10 independent attack subsets, recompute all 256 MMD scores, and evaluate the one-based correct-key rank separately in every repetition. The GE at a fixed N_a is the mean of these 10 ranks. The single-attack rank experiment is kept separate from GE and is evaluated from one complete attack realization at each N_a.

For the numerical GE-bound curve, we form the pairwise gaps $D_k(N_a)$ for each complete attack and use the across-repetition mean and variance to evaluate the pairwise Cantelli expression in Equation (19), after which the 255 pairwise terms are summed according to Equation (20). For the single-attack top-r certificate, the 255 gaps are first sorted within each complete attack, and the mean and variance of each ordered gap are evaluated across repeated complete attacks and substituted into Equation (21). For a chosen attack-stage failure probability δ_a, the certified rank is the smallest r satisfying $q_r(N_a) \leq \delta_a$, as in Equation (23). No smoothing, monotone correction, or substitution of statistics across different key pairs or attack budgets is applied.

Figure 7 compares the GE with its pairwise Cantelli upper-bound evaluation. Both datasets show a decrease in GE as the attack budget grows, while the Cantelli curve remains conservative because it uses only the first two moments of each complete MMD gap. Exact recovery is achieved with substantially fewer traces than the lowest trace counts identified in prior ASCAD studies, e.g., [77–79], although the reported rank statistics are not always directly comparable across works. In terms of the theory, the improving bound reflects an increasing gap-to-variance ratio, $\Delta_k^2(N_a)/V_k(N_a)$, which is precisely the quantity through which the Cantelli bound characterizes trace sufficiency.

Figure 8 shows the best single-realization key rank and the corresponding ordered-gap Cantelli rank bound. The rank reaches stable exact recovery after 400 attack traces for ASCAD-f and after 500 traces for ASCAD-r. For the bound, the pairwise MMD gaps are sorted within each complete attack, and the mean and variance of each ordered gap are estimated across the 100 repetitions. The

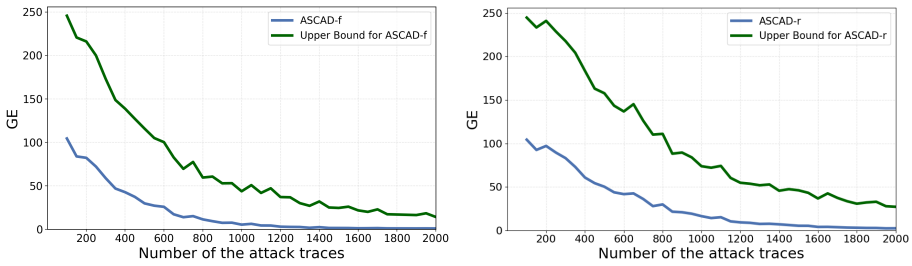


Fig. 7: GE and the pairwise Cantelli GE upper-bound evaluation for the non-profiled MMD attack as functions of the attack budget. At each N_a, GE is the mean one-based correct-key rank over 10 independent complete attacks. A lower rank is better, and rank 1 denotes exact recovery.

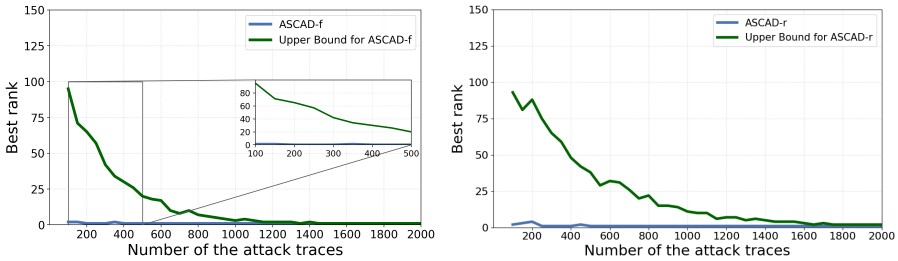


Fig. 8: Single-realization correct-key rank and the pairwise-gap Cantelli single-attack top-r certificate for the non-profiled MMD attack. For a chosen δ_a, the certified rank is the smallest r satisfying $q_r(N_a) \leq \delta_a$.

certified rank is then the smallest rank boundary satisfying the Cantelli condition $q_r(N_a) \leq \delta_a$ for the selected value of $\delta_a = 0.99$.

5.4 Summary of the Results

The experiments support two feasibility claims. First, the single-pair MMD attack produces a correct-key score that is distinct from all wrong-key scores across both ASCAD variants, and GE improves as the attack budget increases. The single-attack rank experiment is treated separately through the ordered-gap Cantelli rank-tail construction and the corresponding top-r certificate. The pairwise Cantelli evaluations remain conservative but use the same score-gap quantities as the non-profiled theory, without imposing a Gaussian-gap or non-degenerate U-statistic model. Second, the RF-SVC pipeline recovers the correct key on both datasets, while the separate attack-budget and profiling-budget experiments expose the two finite-sample effects isolated by the profiled analysis. The pairwise Bernstein overlays distinguish the GE bound from the single-realization rank

bound and retain the individual correct-key versus wrong-key score statistics required by the profiled theory.

The two datasets were collected under materially different acquisition conditions. ASCAD-f uses a 700-sample cropped window and appears to have been low-pass filtered during acquisition, whereas ASCAD-r uses a 1400-sample window, was collected without that filtering, and was sampled at 500 MS/s rather than 200 MS/s. This corresponds to roughly 125 samples per processor clock cycle versus 50. The paper argues that the higher sampling rate in ASCAD-r can improve temporal resolution and help separate leakage from overlapping operations, thereby reducing data-related uncertainty despite the randomized-key profiling regime. The bounds are ultimately controlled by score separation relative to variability: (a) for profiled Bernstein, by the pairwise margins Δ_k and variances σ_k^2; (b) for non-profiled Cantelli, by the MMD gap-to-variance ratios Δ_k^2/V_k. The acquisition setup can affect both sides of these ratios. Better temporal resolution may increase useful key-dependent separation or reduce effective leakage variability, while random-key profiling in ASCAD-r makes the learned profiled predictor less transferable to the fixed-key attack domain. Accordingly, differences between the ASCAD-f and ASCAD-r bounds reflect not only the learning or distinguishing procedure, but also the statistical quality of the leakage induced by the respective acquisition settings.

To conclude, the experiments demonstrate the intended connection between theory and practice, namely, additional traces improve rank because they either improve the learned key-score margins or make those margins large relative to their finite-sample variability. The bounds translate this empirical separation into a quantitative statement about rank reliability, thereby turning an observed trace count into a finite-sample recovery criterion rather than merely another attack-performance point.

6 Discussion

6.1 Rationale for the Selected Approaches

The SVC was selected in the profiled setting because it provides a comparatively transparent route from a learned predictor class to a finite-sample guarantee. With a bounded kernel and an explicit RKHS-norm constraint, its capacity can be controlled through standard Rademacher-complexity arguments [50]. SVC training is also based on a convex regularized optimization problem, which makes its optimization and capacity terms easier to characterize than those of a general NN [80]. The Gaussian-kernel SVC nevertheless defines a nonlinear and nonparametric predictor class rather than imposing an explicit parametric leakage distribution. A NN could provide better empirical prediction, but available finite-sample bounds generally depend on additional architecture- and training-dependent quantities such as depth, layer norms, spectral norms, and achieved margins [61]. The advantage of the SVC in this work is therefore not a claim of universally superior predictive accuracy, but rather that its bounded-RKHS

structure provides a comparatively direct and auditable route from profiling data to probability-estimation control and then to key-score margins.

MMD was selected in the non-profiled setting because it compares key-induced trace distributions without explicit density estimation and, with a characteristic kernel, can detect general distributional differences [13, 66]. This is useful in SCA because a wrong key and the correct key can induce distributions that differ in more than their means or variances. The bounded Gaussian kernel used here also guarantees finite moments of the finite-sample MMD score and of the resulting key-score gaps. However, the empirical MMD score is a quadratic U/V-statistic-type quantity rather than an average of independent per-trace terms [13]. For this reason, the independent-sum Bernstein argument used in the profiled attack does not transfer to the non-profiled attack. U-statistic Bernstein inequalities require kernel- or projection-level variance information and depend on the non-degenerate or canonical structure of the statistic [56–58]. Likewise, a Gaussian model for the complete MMD gap would introduce an additional regime-dependent approximation because MMD statistics need not have the same limiting law in degenerate and non-degenerate regimes [13]. We therefore use Cantelli's one-sided inequality on the complete MMD gap, since it requires only a finite mean and variance and does not require Gaussianity or an internal independence approximation [59, 60]. This choice is assumption-driven rather than selected for a particular numerical fit.

Together, the two instantiations follow the same design principle without forcing the same concentration theorem on two different statistical objects. The profiled attack uses pairwise Bernstein because its conditioned attack score is an average of independent bounded per-trace differences. The non-profiled attack uses pairwise Cantelli because the complete MMD gap is the directly relevant random quantity and only its first two moments are required.

6.2 Scope and Generalizability of the Trace Bounds

The bounds should not be interpreted as universal numerical trace counts that remain unchanged after replacing the learner, score, leakage model, or dataset. What generalizes is the finite-sample route from candidate-key score separation to a rank statement, whereas the quantities controlling that separation depend on the method.

The profiled analysis contains two distinct trace budgets. The profiling budget controls the quality of the selected and calibrated probabilistic predictor, while the attack budget controls the finite-sample ordering of the accumulated key scores produced by that fixed predictor. The profiling side is governed by the complexity of the complete data-dependent predictor class and by the excess clipped-log-loss guarantee. The attack side is governed pairwise by the learned expected margin Δ_k, the variance σ_k^2 of the per-trace correct-key versus wrong-key score difference, and a valid one-sided range b_k. Consequently, two learners with similar classification accuracy can still require different attack budgets if they induce different probability estimates, calibration quality, pairwise key margins, or score variances.

The non-profiled analysis has no separate profiling budget. At each attack budget N_a, the relevant objects are the complete MMD gaps $D_k(N_a)$ and, for a rank boundary, the ordered gap $D_{(r)}(N_a)$. Their means and variances may themselves depend on N_a, so Cantelli does not imply a universal closed-form trace rate unless an additional justified variance-scaling law is available. This is intentional because imposing a generic $1/N_a$ variance law would reintroduce assumptions about the U-statistic regime that the complete-gap analysis is designed to avoid.

The average-rank and single-attack rank guarantees should also remain distinct. In both settings, guessing entropy is the expected rank at a fixed attack budget. For the profiled attack, the Best-Rank theorem gives a high-probability bound on the rank of a single attack realization from the pairwise Bernstein crossing probabilities. For the non-profiled attack, the single-attack top-r certificate is obtained directly from the ordered MMD gap $D_{(r)}(N_a)$, since the event $R_{N_a} > r$ is exactly equivalent to $D_{(r)}(N_a) \leq 0$; Cantelli then bounds this rank-tail event without introducing a repetition count.

6.3 Replacing the Profiled Learner

The attack-stage Bernstein argument is conditionally independent of the specific profiled learner, but not of the predictor produced by the algorithm. Suppose the SVC is replaced by a NN, random forest, logistic regression model, template model, or another probabilistic learner. If the replacement produces a fixed, calibrated predictor after profiling, uses the same clipped logarithmic key score, and is evaluated on independent attack traces, then the same pairwise attack-stage construction applies. For each wrong key, the replacement learner induces its own pairwise margin Δ_k, variance σ_k^2, and one-sided range b_k, and these quantities are inserted into the pairwise Bernstein bound of Section 3.11. Thus, the form of the attack-stage theorem is portable, but the resulting trace requirement is not invariant to the learner.

The profiling-side guarantee is more learner-dependent. The current analysis controls the complete post-selection predictor class containing RF-selected feature subsets, selected kernel bandwidths, bounded-RKHS SVC score functions, and probability-calibration maps. Replacing the SVC requires replacing the RKHS-based complexity calculation with a bound appropriate to the new predictor class. For NNs, for example, available bounds may depend on depth, layer norms, margins, spectral complexity, PAC-Bayes quantities, or algorithmic stability [61, 81, 69]. For a Gaussian template attack, the corresponding analysis would instead require finite-sample control of the estimated class means and covariance matrices under the assumptions used by that model.

A replacement learner must also provide the type of guarantee required by the key-score proof. A hard-label predictor or a classification-error guarantee alone does not control the probabilities accumulated by the attack. The present analysis uses the strict propriety of log loss to transfer excess probability-estimation risk to perturbations of the expected key scores and then to every pairwise key margin [82, 53, 52]. A replacement can inherit this part of the proof when

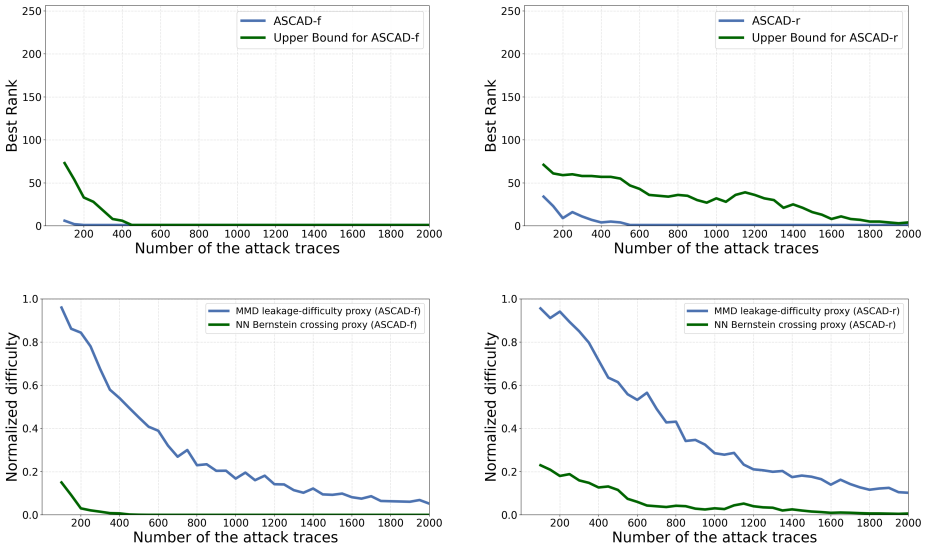


Fig. 9: Rank analysis for NN distinguishers, namely MLP and CNN trained on ASCAD-f/r, respectively [83] and temperature calibrated following [11]. The bottom row compares the empirical MMD-based leakage-difficulty proxy with the difficulty induced by the NN score gaps. The MMD quantity is used as an empirical diagnostic of leakage difficulty, implicitly embedded in the Bernstein rank bound.

it provides probabilistic outputs together with a suitable proper-loss guarantee; otherwise, a new link between its learning objective and the key score is required.

The role of the SVC is therefore not that PAC analysis applies uniquely to SVCs. Its role is to provide a nonlinear predictor class with explicit capacity control while avoiding an explicit parametric leakage-distribution assumption. The overall profiled proof structure is portable, but the profiling complexity term and the induced pairwise attack quantities must be rederived for each replacement learner.

NN-based Distinguisher. To illustrate this portability, we instantiate the same attack-stage analysis with calibrated NN distinguishers. Softmax outputs are not necessarily calibrated, so we temperature-calibrate the logits following [11] before using the resulting probabilities for key ranking. Using the notation of Section 3.11, the calibrated NN outputs determine $Q_{i,k}$ and hence $Z_i^{(k)}$, from which the prefix-wise Δ_k, σ_k^2, and b_k are evaluated and substituted into Equation (12). The GE and single-attack top-r constructions therefore follow without modification, while their numerical values depend on the score separation and variability induced by the NN.

To separate this learned-distinguisher behavior from the underlying leakage difficulty, we also reuse the repeated non-profiled MMD experiments as an

empirical leakage-difficulty proxy. For each attack budget N_a, the methodology already defines the complete correct-versus-wrong MMD gap as $D_k(N_a) = \widehat{s}_{a,b}(k^\star) - \widehat{s}_{a,b}(k)$, with finite-budget mean $\Delta_k(N_a) = \mathbb{E}[D_k(N_a)]$ and variance $V_k(N_a) = \mathrm{Var}(D_k(N_a))$. Across the repeated complete attacks, we estimate these quantities by $\widehat{\Delta}_k(N_a)$ and $\widehat{V}_k(N_a)$ as in Equation (26), and evaluate the non-profiled Cantelli pairwise crossing quantity $p_k(N_a)$ by substituting these empirical moments into Equation (19). Because $p_k(N_a) = V_k(N_a)/(V_k(N_a) + \Delta_k^2(N_a))$ when $\Delta_k(N_a) > 0$, values close to one indicate weak or unstable correct-key separation, whereas values close to zero indicate strong and stable separation.

For a dataset-level diagnostic, we use the normalized average $\mathcal{D}_{\mathrm{leak}}(N_a) = \frac{1}{M-1}\sum_{k\neq k^\star} p_k(N_a)$, with the numerical $p_k(N_a)$ evaluated from $\widehat{\Delta}_k(N_a)$ and $\widehat{V}_k(N_a)$. This MMD quantity is not inserted algebraically into the NN Bernstein bound because the present analysis does not establish that the MMD gap equals, or bounds, the ideal profiled margin $\Delta_k^\star$. Instead, the MMD analysis serves as an empirical diagnostic of leakage difficulty, while Δ_k, σ_k^2, and b_k characterize the score separation actually produced by the trained NN distinguisher.

In our experiments, the NNs [83] may not be the best-performing model among a broader architecture or hyperparameter search. Still, they serve as a representative profiled neural distinguisher to show how the proposed finite-sample rank bounds can be instantiated beyond the SVC. The NN Bernstein bound already reflects the combined practical effect of the underlying leakage, learned representation, probability estimation, and calibration because all of these factors determine the observed $Z_i^{(k)}$ and hence Δ_k, σ_k^2, and b_k. As shown in Figure 9, ASCAD-r remains more difficult than ASCAD-f at moderate and large attack budgets. At the same time, the difference between the NN curves is substantially larger than the corresponding difference between the MMD leakage-difficulty curves. This suggests that the difference between the NN attacks cannot be attributed solely to the leakage and that profiling, representation learning, preprocessing, or calibration can additionally affect how efficiently the available leakage is converted into key-score separation. Conversely, an NN difficulty curve below the MMD proxy does not imply that the NN exceeds the leakage available in the traces, because the MMD quantity approximates leakage difficulty only with respect to the selected trace representation, kernel, and leakage-class pair and therefore need not capture every form of separability that the NN can exploit.

6.4 Replacing the Non-Profiled Distinguisher

The Cantelli construction is more general than MMD because it is applied to the complete score gap rather than to the internal form of the distinguisher. Let a replacement distinguisher assign a finite-sample score $\widehat{S}_{\mathsf{D}}(k)$ to every candidate key, with larger scores preferred, and define the complete pairwise gap $D_k^{\mathsf{D}} = \widehat{S}_{\mathsf{D}}(k^\star) - \widehat{S}_{\mathsf{D}}(k)$. Whenever D_k^{D} has a finite mean and variance, Cantelli gives a one-sided bound on the probability that the wrong key k crosses the correct key. Likewise, ordering the complete gaps produces the rank-boundary

statistic needed for the single-attack top-r construction. Therefore, the pairwise and ordered-gap rank logic can be reused without assuming that the replacement score is an MMD statistic.

What changes is the separation achieved by the distinguisher and the moments of its finite-sample gaps. A bounded mean-difference or correlation score with an independent-sum structure may admit a sharper Hoeffding or Bernstein analysis than Cantelli [84]. An energy-distance score is closely related to MMD under suitable negative-type semimetric conditions [85]. A Wasserstein score's empirical convergence behavior can depend strongly on the dimension and moment assumptions [86]. Mutual-information and KL-divergence scores generally involve density, mass function, entropy, or discretization estimation and therefore require different finite-sample assumptions. A classifier two-sample distinguisher introduces an additional learned hypothesis class whose capacity and training procedure must also be controlled [87].

MMD is therefore not claimed to empirically dominate these alternatives. Its value here is that it is nonparametric; characteristic-kernel MMD is sensitive to general distributional differences, and the finite-sample score is well defined without density estimation [13]. Cantelli then supplies a distribution-free one-sided bound for the complete MMD gap without forcing a possibly incorrect Gaussian or non-degenerate U-statistic approximation. If additional valid structural information about a replacement distinguisher is available, a sharper concentration inequality can be substituted without changing the key-ranking objective.

6.5 Role of the Concept and Hypothesis Classes

The concept and hypothesis classes remain useful because they separate the cryptographic target from the statistical method used to attack it. In the profiled setting, the target leakage labels are fixed by the cryptographic intermediate, while the predictor class determines which conditional probability functions can be learned from profiling traces. The complete predictor class must include every data-dependent choice made during profiling, including feature selection, kernel selection, SVC training, and calibration, because these choices affect the profiling-sample guarantee. This separation also distinguishes approximation error from finite-sample estimation and optimization error.

On the attack side, the candidate-key set $\mathcal{K}$ determines the collection of wrong-key comparisons that can increase the rank of the correct key. The final GE bounds therefore retain one pairwise crossing probability for each wrong key rather than replacing all comparisons by a single worst-case margin. For the non-profiled single-attack top-r result, the relevant object is instead the ordered collection of those pairwise gaps because the event $R_{N_a} > r$ is exactly a rank-boundary event.

MMD additionally uses the unit ball of an RKHS as its witness-function class [13]. This RKHS class defines the distributional discrepancy itself and should not be conflated with the finite candidate-key set. Changing the distinguisher can preserve the same cryptographic key hypotheses while replacing the statistical function class used to score them. This distinction identifies which

assumptions arise from the key-recovery problem and which arise from the selected statistical method.

6.6 What Generalizes Across the Two Settings

The common contribution is not a single score-concentration theorem or one numerical trace formula. The profiled and non-profiled attacks use different random objects and therefore require different finite-sample tools. What they share is a score-gap view of key recovery: the correct-key rank changes only when wrong-key score gaps cross zero, and average-rank or single-attack rank guarantees are obtained by controlling the crossing events appropriate to the reported rank quantity.

In the profiled setting, learning theory first controls the calibrated predictor, proper-loss analysis transfers this control to pairwise expected key-score margins, and pairwise Bernstein controls the finite attack. In the non-profiled setting, the analysis starts directly from the complete MMD score gaps, pairwise Cantelli controls average-rank crossings, and Cantelli applied to the ordered gap $D_{(r)}(N_a)$ controls the top-r rank boundary for a single attack realization. This separation preserves a common finite-sample perspective while matching the probabilistic argument to the statistical structure.

References

1. Kocher, P., Jaffe, J., Jun, B.: Differential power analysis. In: Annual international cryptology conference. pp. 388–397. Springer (1999)
2. Standaert, F.X., Malkin, T.G., Yung, M.: A unified framework for the analysis of side-channel key recovery attacks. In: Annual international conference on the theory and applications of cryptographic techniques. pp. 443–461. Springer (2009)
3. Chari, S., Rao, J.R., Rohatgi, P.: Template attacks. In: International workshop on cryptographic hardware and embedded systems. pp. 13–28. Springer (2002)
4. Timon, B.: Non-profiled deep learning-based side-channel attacks with sensitivity analysis. IACR transactions on cryptographic hardware and embedded systems pp. 107–131 (2019)
5. Papagiannopoulos, K., Glamočanin, O., Azouaoui, M., Ros, D., Regazzoni, F., Stojilović, M.: The side-channel metrics cheat sheet. ACM Computing Surveys **55**(10), 1–38 (2023)
6. Béguinot, J., Cheng, W., Guilley, S., Rioul, O.: Be my guess: Guessing entropy vs. success rate for evaluating side-channel attacks of secure chips. In: 2022 25th Euromicro Conference on Digital System Design (DSD). pp. 496–503. IEEE (2022)
7. Masure, L., Dumas, C., Prouff, E.: A comprehensive study of deep learning for side-channel analysis. IACR Transactions on Cryptographic Hardware and Embedded Systems pp. 348–375 (2020)
8. Masure, L., Cassiers, G., Hendrickx, J., Standaert, F.X.: Information bounds and convergence rates for side-channel security evaluators. IACR Transactions on Cryptographic Hardware and Embedded Systems **2023**(3), 522–569 (2023)
9. Picek, S., Heuser, A., Jovic, A., Bhasin, S., Regazzoni, F.: The curse of class imbalance and conflicting metrics with machine learning for side-channel evaluations.

IACR Trans. on Cryptographic Hardware and Embedded Systems **2019**(1), 1–29 (2019)
10. van der Valk, D., Picek, S.: Bias-variance decomposition in machine learning-based side-channel analysis. Cryptology ePrint Archive, Paper 2019/570 (2019), https://eprint.iacr.org/2019/570
11. Nouraniboosjin, S., Ganji, F.: Too hot to be true: Temperature calibration for higher confidence in NN-assisted side-channel analysis. In: Constructive Approaches for SeCurity Analysis and Design of Embedded systems (CASCADE) (2026), https://eprint.iacr.org/2024/071
12. Nouraniboosjin, S., Ganji, F.: Uncertainty estimation in neural network-enabled side-channel analysis and links to explainability. IACR Transactions on Cryptographic Hardware and Embedded Systems **2026**(2), 764–807 (2026)
13. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. The Journal of Machine Learning Research **13**(1), 723–773 (2012)
14. Panoff, M., Yu, H., Shan, H., Jin, Y.: A review and comparison of ai-enhanced side channel analysis. ACM Journal on Emerging Technologies in Computing Systems (JETC) **18**(3), 1–20 (2022)
15. Picek, S., Perin, G., Mariot, L., Wu, L., Batina, L.: Sok: Deep learning-based physical side-channel analysis. ACM Computing Surveys **55**(11), 1–35 (2023)
16. Maghrebi, H., Portigliatti, T., Prouff, E.: Breaking cryptographic implementations using deep learning techniques. In: Intrl. Conf. on Security, Privacy, and Applied Cryptography Engineering. pp. 3–26. Springer (2016)
17. Cagli, E., Dumas, C., Prouff, E.: Convolutional neural networks with data augmentation against jitter-based countermeasures. In: Intrl. Conf. on Cryptographic Hardware and Embedded Systems. pp. 45–68. Springer (2017)
18. Benadjila, R., Prouff, E., Strullu, R., Cagli, E., Dumas, C.: Study of deep learning techniques for side-channel analysis and introduction to ascad database (2018)
19. Picek, S., Samiotis, I.P., Kim, J., Heuser, A., Bhasin, S., Legay, A.: On the performance of convolutional neural networks for side-channel analysis. In: Intrl. Conf. on Security, Privacy, and Applied Cryptography Engineering. pp. 157–176. Springer (2018)
20. Zaid, G., Bossuet, L., Habrard, A., Venelli, A.: Methodology for efficient cnn architectures in profiling attacks. IACR Trans. on Cryptographic Hardware and Embedded Systems **2020**(1), 1–36 (2020)
21. Wouters, L., Arribas, V., Gierlichs, B., Preneel, B.: Revisiting a methodology for efficient cnn architectures in profiling attacks. IACR Transactions on Cryptographic Hardware and Embedded Systems pp. 147–168 (2020)
22. Wu, L., Weissbart, L., Krč, M., Li, H., Perin, G., Batina, L., Picek, S., et al.: On the attack evaluation and the generalization ability in profiling side-channel analysis. IACR Cryptol. ePrint Arch., 2020/899 (2020)
23. Won, Y.S., Hou, X., Jap, D., Breier, J., Bhasin, S.: Back to the basics: Seamless integration of side-channel pre-processing in deep neural networks. IEEE Transactions on Information Forensics and Security **16**, 3215–3227 (2021)
24. Perin, G., Chmielewski, Ł., Picek, S.: Strength in numbers: Improving generalization with ensembles in machine learning-based profiled side-channel analysis. IACR Transactions on Cryptographic Hardware and Embedded Systems pp. 337–364 (2020)
25. Perin, G., Wu, L., Picek, S.: Exploring feature selection scenarios for deep learning-based side-channel analysis. Cryptology ePrint Archive (2021)

26. Rezaeezade, A., Batina, L.: Regularizers to the rescue: fighting overfitting in deep learning-based side-channel analysis. Journal of Cryptographic Engineering **14**(4), 609–629 (2024)
27. Zaid, G., Bossuet, L., Dassance, F., Habrard, A., Venelli, A.: Ranking loss: Maximizing the success rate in deep learning side-channel analysis. IACR Transactions on Cryptographic Hardware and Embedded Systems pp. 25–55 (2021)
28. Rijsdijk, J., Wu, L., Perin, G., Picek, S.: Reinforcement learning for hyperparameter tuning in deep learning-based side-channel analysis. Cryptol. ePrint Arch., Report 2021/071 (2021)
29. Moos, T., Wegener, F., Moradi, A.: Dl-la: Deep learning leakage assessment: A modern roadmap for sca evaluations. IACR Transactions on Cryptographic Hardware and Embedded Systems pp. 552–598 (2021)
30. Wu, L., Perin, G., Picek, S.: The best of two worlds: Deep learning-assisted template attack. IACR Cryptol. ePrint Arch., 2021/959 (2021)
31. Wu, L., Weissbart, L., Krček, M., Li, H., Perin, G., Batina, L., Picek, S.: Label correlation in deep learning-based side-channel analysis. Ieee transactions on information forensics and security **18**, 3849–3861 (2023)
32. Yap, T., Benamira, A., Bhasin, S., Peyrin, T.: Peek into the black-box: Interpretable neural network using sat equations in side-channel analysis. IACR Transactions on Cryptographic Hardware and Embedded Systems (2023)
33. Acharya, R.Y., Ganji, F., Forte, D.: Information theory-based evolution of neural networks for side-channel analysis. IACR Transactions on Cryptographic Hardware and Embedded Systems pp. 401–437 (2023)
34. Brier, E., Clavier, C., Olivier, F.: Correlation power analysis with a leakage model. In: International workshop on cryptographic hardware and embedded systems. pp. 16–29. Springer (2004)
35. Ramezanpour, K., Ampadu, P., Diehl, W.: SCAUL: Power side-channel analysis with unsupervised learning. IEEE Transactions on Computers **69**(11), 1626–1638 (2020)
36. Perin, G., Chmielewski, Ł., Batina, L., Picek, S.: Keep it unsupervised: Horizontal attacks meet deep learning. IACR Transactions on Cryptographic Hardware and Embedded Systems pp. 343–372 (2021)
37. Weissbart, L., Picek, S., Batina, L.: One trace is all it takes: Machine learning-based side-channel attack on eddsa. In: Intrl. Conf. on Security, Privacy, and Applied Cryptography Engineering. pp. 86–105. Springer (2019)
38. Weissbart, L., Chmielewski, Ł., Picek, S., Batina, L.: Systematic side-channel analysis of curve25519 with machine learning. Journal of Hardware and Systems Security **4**(4), 314–328 (2020)
39. Staib, M., Moradi, A.: Deep learning side-channel collision attack. IACR Transactions on Cryptographic Hardware and Embedded Systems **2023**(3), 422–444 (2023)
40. Valiant, L.G.: A theory of the learnable. Communications of the ACM **27**(11), 1134–1142 (1984)
41. Blumer, A., Ehrenfeucht, A., Haussler, D., Warmuth, M.K.: Learnability and the vapnik-chervonenkis dimension. Journal of the ACM (JACM) **36**(4), 929–965 (1989)
42. Picek, S., Heuser, A., Wu, L., Alippi, C., Regazzoni, F.: When theory meets practice: A framework for robust profiled side-channel analysis. Cryptology ePrint Archive, Paper 2018/1123 (2018), https://eprint.iacr.org/2018/1123
43. Zaid, G.: Bridging Deep Learning and Classical Profiled Side-Channel Attacks. Ph.D. thesis, University of Lyon (2021), https://theses.hal.science/tel-03722660

44. Bronchain, O., Hendrickx, J.M., Massart, C., Olshevsky, A., Standaert, F.X.: Leakage certification revisited: Bounding model errors in side-channel security evaluations. In: Annual International Cryptology Conference. pp. 713–737. Springer (2019)
45. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
46. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of machine learning research **3**(Mar), 1157–1182 (2003)
47. Louppe, G., Wehenkel, L., Sutera, A., Geurts, P.: Understanding variable importances in forests of randomized trees. Advances in neural information processing systems **26** (2013)
48. Steinwart, I.: On the influence of the kernel on the consistency of support vector machines. Journal of machine learning research **2**(Nov), 67–93 (2001)
49. Crammer, K., Singer, Y.: On the algorithmic implementation of multiclass kernel-based vector machines. Journal of machine learning research **2**(Dec), 265–292 (2001)
50. Bartlett, P.L., Mendelson, S.: Rademacher and gaussian complexities: Risk bounds and structural results. Journal of machine learning research **3**(Nov), 463–482 (2002)
51. Maurer, A.: A vector-contraction inequality for rademacher complexities. In: International Conference on Algorithmic Learning Theory. pp. 3–17. Springer (2016)
52. Reid, M.D., Williamson, R.C.: Information, divergence and risk for binary experiments. Journal of Machine Learning Research **12**(22), 731–817 (2011)
53. Williamson, R.C., Vernet, E., Reid, M.D.: Composite multiclass losses. Journal of Machine Learning Research **17**(222), 1–52 (2016)
54. Maurer, A.: A bernstein-type inequality for functions of bounded interaction. Bernoulli **25**(2), 1451–1471 (2019). https://doi.org/10.3150/18-BEJ1026
55. Nouraniboosjin, S., Ganji, F.: Rock and a hard place: Attack hardness in neural network-assisted side channel analysis. Cryptology ePrint Archive, Paper 2025/903 (2025)
56. Arcones, M.A.: A bernstein-type inequality for u-statistics and u-processes. Statistics & probability letters **22**(3), 239–247 (1995)
57. Peel, T., Anthoine, S., Ralaivola, L.: Empirical bernstein inequalities for u-statistics. Advances in Neural Information Processing Systems **23** (2010)
58. Giné, E., Latała, R., Zinn, J.: Exponential and moment inequalities for u-statistics. In: High dimensional probability II, pp. 13–38. Springer (2000)
59. Cantelli, F.P.: Sui confini della probabilita. In: Atti del Congresso Internazionale dei Matematici: Bologna del 3 al 10 de settembre di 1928. pp. 47–60 (1929)
60. Ion, R.A., Klaassen, C.A., Heuvel, E.R.v.d.: Sharp inequalities of bienaymé–chebyshev and gauß type for possibly asymmetric intervals around the mean: Ra ion et al. Test **32**(2), 566–601 (2023)
61. Bartlett, P.L., Foster, D.J., Telgarsky, M.J.: Spectrally-normalized margin bounds for neural networks. Advances in neural information processing systems **30** (2017)
62. Benadjila, R., Prouff, E., Strullu, R., Cagli, E., Dumas, C.: ASCAD: the ATMega8515 SCA traces databases (fixed key). [Online] https://github.com/ANSSI-FR/ASCAD/tree/master/ATMEGA_AES_v1/ATM_AES_v1_fixed_key [Accessed: May.19, 2025] (2017)
63. Benadjila, R., Prouff, E., Strullu, R., Cagli, E., Dumas, C.: ASCAD: the ATMega8515 SCA traces databases (variable key). [Online] https://github.com/ANSSI-FR/ASCAD/tree/master/ATMEGA_AES_v1/ATM_AES_v1_variable_key [Accessed: May.19, 2025] (2017)

64. Shalev-Shwartz, S., Ben-David, S.: Understanding machine learning: From theory to algorithms. Cambridge university press (2014)
65. Fukumizu, K., Gretton, A., Lanckriet, G., Schölkopf, B., Sriperumbudur, B.K.: Kernel choice and classifiability for rkhs embeddings of probability distributions. Advances in neural information processing systems **22** (2009)
66. Sriperumbudur, B.K., Fukumizu, K., Lanckriet, G.R.: Universality, characteristic kernels and rkhs embedding of measures. Journal of Machine Learning Research **12**(7) (2011)
67. Quinlan, J.R.: Induction of decision trees. Machine learning **1**(1), 81–106 (1986)
68. Cortes, C., Mohri, M., Rostamizadeh, A.: Multi-class classification with maximum margin multiple kernel. In: International conference on machine learning. pp. 46–54. PMLR (2013)
69. Bousquet, O., Elisseeff, A.: Stability and generalization. Journal of machine learning research **2**(Mar), 499–526 (2002)
70. Platt, J.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers **10**(3), 61–74 (1999)
71. Wu, T.F., Lin, C.J., Weng, R.C.: Probability estimates for multi-class classification by pairwise coupling. Journal of Machine Learning Research **5**(Aug), 975–1005 (2004)
72. Zhang, J., Zheng, M., Nan, J., Hu, H., Yu, N.: A novel evaluation metric for deep learning-based side channel analysis and its extended application to imbalanced data. IACR Transactions on Cryptographic Hardware and Embedded Systems pp. 73–96 (2020)
73. Boucheron, S., Lugosi, G., Massart, P.: Concentration Inequalities: A Nonasymptotic Theory of Independence. Oxford University Press (2013)
74. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
75. Wu, L., Perin, G., Picek, S.: On the evaluation of deep learning-based side-channel analysis. In: International Workshop on Constructive Side-Channel Analysis and Secure Design. pp. 49–71. Springer (2022)
76. Prouff, E., Rivain, M.: Theoretical and practical aspects of mutual information based side channel analysis. In: Applied Cryptography and Network Security: 7th International Conference, ACNS 2009, Paris-Rocquencourt, France, June 2-5, 2009. Proceedings 7. pp. 499–518. Springer (2009)
77. Kwon, D., Hong, S., Kim, H.: Optimizing implementations of non-profiled deep learning-based side-channel attacks. IEEE Access **10**, 5957–5967 (2022)
78. Wu, L., Perin, G., Picek, S.: Weakly profiling side-channel analysis. IACR Transactions on Cryptographic Hardware and Embedded Systems **2024**(3), 707–730 (2024)
79. Li, D., Gong, Z., Tang, Y., Li, C.: Ds-nsca: Differentiable search for non-profiling side-channel analysis. IACR Transactions on Cryptographic Hardware and Embedded Systems **2026**(2), 659–685 (2026)
80. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**(3), 273–297 (1995)
81. McAllester, D.A.: Pac-bayesian model averaging. In: Proceedings of the twelfth annual conference on Computational learning theory. pp. 164–170 (1999)
82. Gneiting, T., Raftery, A.E.: Strictly proper scoring rules, prediction, and estimation. Journal of the American statistical Association **102**(477), 359–378 (2007)

83. Wu, L., Perin, G., Picek, S.: I choose you: Automated hyperparameter tuning for deep learning-based side-channel analysis. IEEE Transactions on Emerging Topics in Computing (2022)
84. Hoeffding, W.: Probability inequalities for sums of bounded random variables. Journal of the American statistical association **58**(301), 13–30 (1963)
85. Sejdinovic, D., Sriperumbudur, B., Gretton, A., Fukumizu, K.: Equivalence of distance-based and rkhs-based statistics in hypothesis testing. The annals of statistics pp. 2263–2291 (2013)
86. Fournier, N., Guillin, A.: On the rate of convergence in wasserstein distance of the empirical measure. Probability theory and related fields **162**(3), 707–738 (2015)
87. Lopez-Paz, D., Oquab, M.: Revisiting classifier two-sample tests. arXiv preprint arXiv:1610.06545 (2016)

Table 1: Summary and interpretation of the PAC bounds for non-/profiled SCA.

Quantity	Bound	Interpretation	Practical implication
Profiled SCA			
GE	Bernstein, pairwise $p_k(N_a) = \exp\left(-\frac{N_a \Delta_k^2}{2\sigma_k^2 + \frac{2}{3} b_k \Delta_k}\right)$ $\mathrm{GE}(N_a) \leq 1 + \sum_{k \neq k^\star} p_k(N_a)$	Δ_k: mean per-trace correct-vs.-wrong score advantage σ_k^2: trace-to-trace variance of the pairwise score gap b_k: largest centered per-trace deviation Large Δ_k, small σ_k^2, b_k: stronger, more stable score separation	More attack traces average out score fluctuations. Keys with large and stable separation are eliminated quickly, whereas closely competing or highly variable keys dominate the required N_a.
Single-attack top-r rank	Bernstein, rank tail $\Pr[R_{N_a} > r] \leq q_r(N_a) \leq \delta_a$ $\rightarrow \Pr[R_{N_a} \leq r] \geq 1 - \delta_a$ End-to-end confidence: $1 - \delta_p - \delta_a$	$\Delta_k, \sigma_k^2, b_k$: pairwise score-separation statistics r: target maximum correct-key rank δ_a: attack-stage failure probability δ_p: profiling-stage failure probability	Answers the practical question: *with N_a traces, how high can the correct key rank in a single attack with the requested confidence?*
Non-profiled SCA			
GE	Cantelli, pairwise $p_k(N_a) = \frac{V_k(N_a)}{V_k(N_a) + \Delta_k^2(N_a)}$ $\mathrm{GE}(N_a) \leq 1 + \sum_{k \neq k^\star} p_k(N_a)$	$\Delta_k(N_a)$: mean correct-vs.-wrong MMD-score gap $V_k(N_a)$: variance of the complete MMD-score gap $\Delta_k^2(N_a)/V_k(N_a)$: squared separation relative to fluctuation Large ratio: stable correct-vs.-wrong separation	If the correct key produces an MMD score that is consistently larger than a wrong key, its crossing probability becomes small. GE decreases as more wrong keys become well separated from the correct key. No attack-stage δ_a is required because GE is already an expectation.
Single-attack top-r rank	Cantelli, ordered gap $q_r(N_a) = \frac{V_r(N_a)}{V_r(N_a) + \Delta_r^2(N_a)} \leq \delta_a$ $\rightarrow \Pr[R_{N_a} \leq r] \geq 1 - \delta_a$	$\Delta_r(N_a)$: mean separation at the rth rank boundary $V_r(N_a)$: variance of the rth rank-boundary gap $\Delta_r(N_a)/\sqrt{V_r(N_a)}$: rank-boundary separation relative to fluctuation r: target maximum correct-key rank δ_a: attack-stage failure probability	Answers the same practical question as the profiled bound, but directly from MMD-score separation: *how many attack traces are needed for a single non-profiled attack to place the correct key within the top r with confidence $1 - \delta_a$?*