CERTIFYING ROBUSTNESS BY BETTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Randomized smoothing provides strong robustness certificates, but often requires tens of thousands of Monte Carlo evaluations per input. Adapting this computational budget to each input requires sequentially valid inference, since fixed-n Monte Carlo confidence bounds generally do not retain their nominal guarantees under repeated inspection or data-dependent stopping. We formulate certification as a sequential hypothesis-testing problem in which the null corresponds to failure of the population-level certificate. For mean, mean–variance, and discretized-CDF certificates, we construct predictable bounded witnesses whose betting products form e-processes, thereby controlling false certification at any level α under arbitrary stopping. We also design reverse information projection (RIPr) guided e-processes that show modest additional sampling gains over standard constructions. Experiments on CIFAR-10 and ImageNet attribute most small-budget gains to direct certificate witnesses rather than monitoring alone. Finally, we introduce a validity-preserving futility rule that terminates uncertifiable inputs early, applies to any of these procedures, and abandons a certifiable input only with probability bounded by its own error level. The rule significantly reduces the mean evaluation cost for every method, including the baselines, with no observed loss of certificates in the reported experiments.

1 INTRODUCTION

Randomized smoothing (RS) turns a base model into a certifiably robust one by averaging its output over random, e.g. Gaussian, perturbations of the input. Different certificates are based on different summary statistics of the score distribution. The smoothed mean alone supports the Neyman–Pearson (NP) certificate (Cohen et al., 2019; Kumar et al., 2020a), pairing that mean with the score variance yields the variance certificate (VC) (Schuchardt et al., 2023), and discretizing the entire score cumulative distribution function (CDF) yields the CDF certificate (Kumar et al., 2020a). All three depend on population quantities that must be inferred from finitely many samples.

Conventional procedures fix a terminal sample size n and substitute a confidence bound into the population certificate. The resulting guarantee applies at the prespecified budget, but repeated inspection and stopping at the first successful certification generally invalidate that fixed-sample guarantee. One established remedy estimates the underlying quantities with a time-uniform confidence sequence (CS), valid at all sample sizes at once. Betting-based sequences provide tight bounds for bounded means (Howard et al., 2021; Waudby-Smith & Ramdas, 2024), and Voráček (2024) have already brought such sequences to smoothing certification for the hard-label mean. Plugging a CS into a certificate gives us a strong baseline, but we show that there is a more powerful approach.

We test the certification condition directly. For a given input, the goal is to certify that the model's smoothed score for a target class stays above a threshold τ under every perturbation of ℓ_2 norm at most r. We fix r, τ, and a certificate before sampling. The *bad null* consists of score laws for which the population certificate is at most τ. Rejecting the bad null certifies the input. To accumulate evidence against this null, we repeatedly evaluate the model on independently perturbed copies of the input. Each forward pass yields a score. Before seeing it, we choose a bet using the scores observed so far. A *witness* turns the score into a signed payoff with nonpositive expectation under every law in the bad null. A bounded, nonnegative *stake* scales the gain or loss relative to our current capital. If the null holds, these bets are fair or unfavorable given the past. The resulting capital process is an *e-process*, whose growth provides evidence against the null. Ville's inequality guarantees that under

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

the bad null, the probability that capital ever reaches $1/\alpha$ times its initial value is at most α (Ville, 1939; Shafer & Vovk, 2019). We can therefore stop and certify as soon as the capital reaches $1/\alpha$.

We make three contributions. (i) *Certificate-specific e-processes*: we design bounded witnesses for three certificates, yielding e-processes that outperform both fixed-n and anytime-valid plug-in baselines in terms of both certified accuracy and number of samples. We then use ideas from reverse information projection (RIPr) to guide witness and stake tuning. (ii) *Multiclass certification*: we extend these tests to more powerful multiclass certificates by comparing the target with each competitor. We show how to avoid classwise Bonferroni correction with an intersection–union test, which also applies to plug-in tests. (iii) *Protected futility stopping*: we introduce a rule that uses upper confidence bounds to stop sampling early on uncertifiable inputs. It preserves false-certification control and bounds the probability of abandoning a certifiable input (§ 4). Our contributions build on standard betting and intersection–union principles and show the value of tailored tests.

Experiments on CIFAR-10 and ImageNet show comparable or higher certified accuracy, especially at low budgets. Our tests use similar or fewer samples than adaptive baselines and far fewer than fixed-n baselines, with a median of 20 samples in some cases. Our futility rule applies to all methods, including the baselines, reducing mean sample cost by 3–15$\times$ with no observed loss of certificates. Ablations separate the effects of witness construction, working-law adaptation and error correction (§ 7).

2 Related work

Smoothing certificates. Cohen et al. (2019) certify an ℓ_2 radius from the smoothed top-class probability alone; additional information about the score distribution can yield tighter guarantees. Lee et al. (2019) sharpen the certificate, Kumar et al. (2020a) certify from a discretized score CDF, and Schuchardt et al. (2023) constrain variance as well as mean. We build on the NP, VC and CDF certificates from this line and develop their statistical layer, which infers a population guarantee from finitely many noisy evaluations (Voráček, 2024). Delattre et al. (2025) address the multiplicity correction, reducing the number of corrected comparisons below the number of classes; their method is a baseline in § 7.2. Two works bring sequential inference to smoothing. Voráček (2024) replace the fixed-n bound on the hard-label mean by a betting confidence sequence, which is anytime-valid; their Krichevsky–Trofimov sequence is our strongest plug-in baseline on hard labels (§ 7.2). Cullen et al. (2026) run an e-process on the same hard-label probability, meta-learning a per-image betting prior and stopping on wealth- and precision-based exit rules. Both are single-class and mean-only. We test the certificate itself, with a witness built for it: for VC and CDF, which a $0/1$ stream fixes through its mean alone, this splits no α between components and inverts no bound. We also carry the guarantee to the argmax, and bound the probability that early abstention discards a certifiable input (Prop. 3). Certification cost is also reduced by other means: input-specific sample sizes (Chen et al., 2022), variance-reduced base classifiers (Horváth et al., 2022), certificate reuse across model versions (Ugare et al., 2024). All retain a fixed-budget statistical analysis (§ K).

Sequential testing, e-values, and betting. Adapting the sample size goes back to Wald (1947), with the modern nonnegative-supermartingale form resting on Ville (1939) and Shafer & Vovk (2019) and surveyed by Ramdas et al. (2023). Time-uniform confidence sequences (Howard et al., 2021), betting bounds for bounded means (Waudby-Smith & Ramdas, 2024), and monotone-process variants for CDF-type functionals (Clerico et al., 2025) give sharp anytime-valid *estimates*, and are the plug-in comparators of § 7.1; reverse information projection supplies e-variable optimality theory (Lardy et al., 2024; Larsson et al., 2025), which we use only as a design device. Our fixed-budget comparators come from empirical-Bernstein bounds (Maurer & Pontil, 2009; Martinez-Taboada & Ramdas, 2026) and the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality with Massart's constant (Massart, 1990). These bounds infer parameters; we construct tests directly for smoothing certificates. Betting has also been used to stop the resampling of a Monte Carlo test early (Fischer & Ramdas, 2025), and empirical-Bernstein stopping (Mnih et al., 2008) is the classical adaptive-sample-size precedent. Two extensions we do not pursue are in § K.

3 Certificates as sequential tests

Fix an input $\boldsymbol{x} \in \mathbb{R}^d$, a target class, and a score function $f : \mathbb{R}^d \to [0, 1]$ of the base model. For example, f can be the target class's softmax score or the indicator that the model predicts

that class (the hard label). For a noise level $\sigma > 0$, the smoothed score at the original input is $g(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\epsilon}}[f(\boldsymbol{x} + \boldsymbol{\epsilon})]$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_d)$ and $\boldsymbol{I}_d$ is the $d \times d$ identity. Before sampling, we also fix a radius $r > 0$, a certificate threshold $\tau \in (0, 1)$, an initial test level $\alpha \in (0, 1)$, and, for the CDF certificate, the grid on which it is discretized. Independent draws $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_d)$ then reveal scores $X_t = f(\boldsymbol{x} + \boldsymbol{\epsilon}_t) \in [0, 1]$. Their common law is P. With $\rho = r/\sigma$, a population certificate $\mathrm{Cert}_j(P; \rho)$ lower-bounds the smoothed score at perturbed inputs: $g(\widetilde{\boldsymbol{x}}) \geq \mathrm{Cert}_j(P; \rho)$ whenever $\|\widetilde{\boldsymbol{x}} - \boldsymbol{x}\|_2 \leq r$. We consider three certificates (NP, VC, CDF) based on different summary statistics.

Set $M_P = \mathbb{E}_P[X_t] = g(\boldsymbol{x})$, $V_P = \mathrm{Var}_P(X_t)$, $T_\rho(u) = \Phi\big(\Phi^{-1}(u) + \rho\big)$, and $\kappa_\rho = \sqrt{e^{\rho^2} - 1}$, with Φ the standard normal CDF and the endpoint values of the composition at $u \in \{0, 1\}$ defined by continuity. The NP (Cohen et al., 2019) and VC certificates take the form

$$\mathrm{Cert}_{\mathrm{NP}}(P; \rho) = \Phi\big(\Phi^{-1}(M_P) - \rho\big), \tag{1}$$

$$\mathrm{Cert}_{\mathrm{VC}}(P; \rho) = M_P - \kappa_\rho \sqrt{V_P}. \tag{2}$$

where Eq. 2 specializes the certificate of Schuchardt et al. (2023) to scalar scores. The CDF certificate uses a prespecified grid $0 = b_0 < \cdots < b_K = 1$, with $F_{P,i} = P(X_t \leq b_i)$ and $\Delta_i = b_i - b_{i-1}$,

$$\mathrm{Cert}_{\mathrm{CDF}}(P; \rho) = b_{K-1} - \sum_{i=1}^{K-1} \Delta_i T_\rho(F_{P,i}). \tag{3}$$

Our goal is to certify that $\mathrm{Cert}_j(P; \rho) > \tau$, typically with $\tau = 1/2$ for classification. At that threshold the target class holds a majority of the smoothed softmax mass or hard-label votes, certifying the top-1 decision throughout the ball from one scalar per draw. We test the negation: for $j \in \{\mathrm{NP}, \mathrm{VC}, \mathrm{CDF}\}$ the composite *bad null* is $H_j(\rho, \tau) = \{Q : \mathrm{Cert}_j(Q; \rho) \leq \tau\}$, abbreviated H_j. Testing acts as a stochastic proof by contradiction, we reject $P \in H_j$ and certify only upon observing an unlikely event of probability at most α under every law in H_j. The NP null depends only on the mean M_P, the VC null on the mean and variance (M_P, V_P), and the CDF null on the CDF values $(F_{P,1}, \ldots, F_{P,K-1})$.

A strength of RS is that it applies to black-box models, including via API (Salman et al., 2020). The single-class construction above works with either a predicted label (hard) or the target class's probability (soft). If the model/API also provides probabilities for all classes, we can construct more powerful multiclass certificates by comparing the target class directly against its competitors (§ 6).

4 Anytime-valid certificate tests

Let $(\mathcal{F}_t)_{t \geq 0}$ be the filtration generated by the observed scores together with any pre-sampling side information (the multiclass designs of § 6 use one noiseless forward pass this way), and call a map or stake *predictable* if it is $\mathcal{F}_{t-1}$-measurable. Conditionally on $\mathcal{F}_{t-1}$, the mean, variance and grid-CDF values of X_t under P are again M_P, V_P and $F_{P,i}$. We state the standard betting construction to fix notation (Shafer & Vovk, 2019; Howard et al., 2021).

Proposition 1 (Standard bounded-witness construction)**.** *Let H be a composite bad null, $u_t : [0, 1] \to \mathbb{R}$ a bounded predictable witness, and $\lambda_t \geq 0$ a predictable stake. Set $U_t = u_t(X_t)$ and $e_t = 1 + \lambda_t U_t$. Suppose $\mathbb{E}_P[U_t \mid \mathcal{F}_{t-1}] \leq 0$ holds for every $P \in H$, and that $e_t \geq 0$ over the full observation support. Then $\mathcal{K}_t = \prod_{s=1}^t e_s$ is an e-process for H and $\sup_{P \in H} \mathbb{P}_P(\sup_{t \geq 0} \mathcal{K}_t \geq 1/\alpha) \leq \alpha$.*

In § A we elaborate the stronger statement that $\mathcal{K}_t$ is a nonnegative supermartingale under *every* $P \in H$ at once, which allows us in § 6 to apply Ville's inequality to one pairwise null inside a union. In § C we verify the deployed procedures' false-certification rate on synthetic boundary laws. The two conditions serve different roles: nonpositivity of $\mathbb{E}_P[U_t \mid \mathcal{F}_{t-1}]$ yields $\mathbb{E}_P[e_t \mid \mathcal{F}_{t-1}] \leq 1$, while $e_t \geq 0$ is a separate support requirement. We next construct a witness u_t for each certificate.

NP. Since Φ is monotone, the NP null constrains only the mean: $\mathrm{Cert}_{\mathrm{NP}}(P; \rho) \leq \tau \iff M_P \leq T_\rho(\tau)$. Thus $U_t^{\mathrm{NP}} = X_t - T_\rho(\tau)$ has nonpositive conditional mean on H_{NP}, and stakes $0 \leq \lambda_t \leq 1/T_\rho(\tau)$ keep its factor nonnegative. In § A.2 we provide the reduction and support verification.

VC. Pick predictable $m_t \in (0, 1)$ and $0 < v_t \leq m_t(1 - m_t)$ subject to $m_t - \kappa_\rho \sqrt{v_t} \geq \tau$, then set

$$\beta_t = \frac{\kappa_\rho}{2\sqrt{v_t}}, \quad Y_t(x) = \beta_t x^2 - (1 + 2\beta_t m_t)x, \quad D_t = \beta_t v_t - m_t - \beta_t m_t^2. \tag{4}$$

Every $P \in H_{\mathrm{VC}}$ obeys $\mathbb{E}_P[Y_t(X_t) \mid \mathcal{F}_{t-1}] \geq D_t$ (see § A.3 for details), making $U_t^{\mathrm{VC}} = D_t - Y_t(X_t)$ a witness with nonpositive conditional mean under that null. For the support hypothesis, U_t^{VC} is concave, so its minimum $u_{\min,t}$ over $[0,1]$ is at an endpoint and stakes $0 \leq \lambda_t \leq -1/u_{\min,t}$ keep the factor nonnegative.

CDF. For $\rho \geq 0$ the map T_ρ is increasing and concave, strictly so when $\rho > 0$. Fix predictable $q_{i,t} \in (0,1)$ and set $w_{i,t} = \Delta_i T'_\rho(q_{i,t})$ and $d_t = b_{K-1} - \tau - a_t$, where

$$a_t = \sum_{i=1}^{K-1} \Delta_i \{T_\rho(q_{i,t}) - T'_\rho(q_{i,t}) q_{i,t}\}, \qquad U_t^{\mathrm{CDF}} = d_t - \sum_{i=1}^{K-1} w_{i,t} \mathbf{1}\{X_t \leq b_i\}. \tag{5}$$

The coordinatewise tangent bounds imply $\mathbb{E}_P[U_t^{\mathrm{CDF}} \mid \mathcal{F}_{t-1}] \leq 0$ across H_{CDF}. Factor nonnegativity follows from $0 \leq \lambda_t \leq 1/(W_t - d_t)$ whenever $W_t = \sum_{i=1}^{K-1} w_{i,t}$ satisfies $0 < d_t < W_t$, otherwise we may use the *neutral factor* $e_t = 1$, which is always valid (see § A.4 for the derivations).

Error control. The NP witness is the standard bet on a bounded mean. The VC and CDF witnesses are, to our knowledge, new. One source of their power is avoiding error splits: each tests the certificate condition through a *single* payoff at the full level α, regardless of the number of statistics involved. A standard plug-in approach uses a union bound instead to ensure simultaneous coverage. For VC, this means a lower bound on M_P and an upper bound on V_P, e.g. each with $\alpha/2$ confidence (§ D.2). Fan et al. (2025) jointly test upper bounds on the mean and variance, but do not study certification.

Futility. An e-process can certify early, but otherwise continues to the evaluation cap. In our experiments, 39–50% of images remain uncertified within the maximum budget. The complementary decision is to stop sampling and *abstain*. Fix a budget N, let T be the first crossing time of the e-process, and let F be any stopping time adapted to $(\mathcal{F}_t)$. The procedure with futility certifies if $T \leq \min(F, N)$, abstains for futility if $F < \min(T, N)$, and abstains at the max budget otherwise.

Proposition 2 (Futility preserves validity). *If* $\sup_{P \in H} \mathbb{P}_P(T \leq N) \leq \alpha$, *then* $\sup_{P \in H} \mathbb{P}_P\{T \leq \min(F, N)\} \leq \alpha$ *for every adapted* F.

Proof. The event $\{T \leq \min(F, N)\}$ is a subset of $\{T \leq N\}$, so any futility rule, adapted or not, preserves the false-certification guarantee. □

Futility may still reduce power, by abandoning an input that would otherwise be certified. Therefore, we construct next a futility rule that bounds the loss of certificates as well, built from an *upper* confidence bound on the population certificate.

Proposition 3 (Futility protection). *Let* $\overline{C}_t \geq \mathrm{Cert}_j(P; \rho)$ *hold at every futility checkpoint with probability at least* $1 - \beta_{\mathrm{f}}$, *and let* F *be the first checkpoint at which* $\overline{C}_t \leq \tau$. *Then*

$$\sup_{P:\, \mathrm{Cert}_j(P;\rho) > \tau} \mathbb{P}_P(F < T) \leq \beta_{\mathrm{f}}. \tag{6}$$

Proof. For a certifiable input, $\{F < T\}$ can occur only if $\overline{C}_F$ fails to cover the true certificate, because $\overline{C}_F \leq \tau < \mathrm{Cert}_j(P; \rho)$. Thus $\mathbb{P}_P(F < T) \leq \beta_{\mathrm{f}}$ by simultaneous coverage. □

Population futility. We construct the upper bounds with a simple staged approach, using Hoeffding bounds for NP and VC and a one-sided DKW band for CDF. We can reuse the certification samples because the protection guarantee requires simultaneous coverage, not independence. We allocate per-checkpoint errors summing to β_{f} over a fixed grid to obtain the simultaneous coverage required by (6) (see § H.4). Alternatively, we could invert suitable e-processes to obtain the bounds. The guarantee holds for any budget and limits the probability of abandoning a certifiable input to β_{f}. We observe no lost certificates in our experiments (§ 7.1).

Slope-based futility. We also consider a cheaper alternative: we fit a slope to the recent log capital and extrapolate to the sample maximum budget, abstaining if the projected path does not reach $\log(1/\alpha)$. By Prop. 2, we retain false-certification control, but obtain no counterpart of (6), since a fitted slope does not bound future growth with controlled error (see § H.4). In our experiments, it adds trade-offs on the Pareto frontier between certified accuracy and sample cost (Figure 3).

5 PREDICTABLE RIPR-GUIDED TUNING

All three certificates use factors $e_t = 1 + \lambda_t U_t$. Each witness encodes its certificate's null, while Prop. 1 supplies a common validity proof. We guide the witness and stake choices using a *working law* estimated from past scores. Under standard conditions, reverse information projection (RIPr) identifies a likelihood-ratio e-variable that maximizes expected log-growth under that working law (Lardy et al., 2024; Larsson et al., 2025). Our common construction lets us use RIPr for tuning without deriving an exact continuous-score RIPr factor for each null. We leave this to future work.

We use finite-alphabet RIPr as a *design device* for these factors. We summarize past scores by a categorical working law $\widehat{P}_t$ on a fixed K-cell partition of $[0,1]$, using the b_i grid by default (§ B.2). The corresponding RIPr problem is $P_t^\star \in \arg\min_{Q \in H_j^K(\rho,\tau)} \mathrm{KL}(\widehat{P}_t \| Q)$, where H_j^K is the categorical design null and KL denotes Kullback–Leibler divergence. Under the conditions of § B, this auxiliary problem gives an exact likelihood-ratio e-variable. Its first-order optimality conditions guide our witness and stake choices (§ B.3).

For NP, we keep the mean witness and optimize only its stake. For VC, we maximize working-law expected log-growth over the RIPr-motivated family of boundary witnesses and admissible stakes. For CDF, we numerically project the working histogram to select a supporting tangent, then optimize its stake (§ B). We freeze each choice before observing X_t and evaluate e_t on the raw score. Prop. 1 guarantees validity even when the working law or projection is approximate.

6 MULTICLASS ARGMAX CERTIFICATION

So far we certify a scalar score against a fixed threshold. For multiclass classification, we aim to certify that the smoothed argmax remains unchanged in the perturbation ball. The standard approach bounds the target probability and that of its strongest competitor separately with a multiplicity correction for the maximum, dividing α by $C-1$. We cast the decision as an *intersection–union* test over pairwise margins (Berger, 1982), avoiding this penalty. This classical property also applies to plug-in tests of each pair. E-process components additionally permit continuous monitoring and stopping at the first crossing. The number of classes C affects bookkeeping but not the error budget.

We fix the target class A before sampling and let $\mathbf{X}_t$ be the model's output on the tth noisy draw, either softmax or hardmax one-hot, in the probability simplex on C classes. For each competitor $B \neq A$, we define the normalized pairwise margin $Z_t^{(B)} = \frac{1}{2}\{1 + (X_{t,A} - X_{t,B})\} \in [0,1]$. This is an ordinary $[0,1]$ score, sending a tie to $1/2$, so § 3–§ 4 apply to it verbatim. Any valid scalar certificate on its law lower-bounds the worst-case margin over the ball, so $\mathrm{Cert}_j(P_{A,B};\rho) > 1/2$ forces class A to outrank B at every admissible perturbation. Write H_B for the pairwise bad null, such as $\{\mathrm{Cert}_j(P_{A,B};\rho) \leq 1/2\}$. Rejecting every competitor's null certifies the argmax (15), so the bad null is a *union* over competitors. We run one e-process per pair until the smallest of their running maxima reaches $1/\alpha$ (17). This implements an intersection–union test: we reject all $C-1$ pairwise nulls at the same level. Writing $\mathcal{K}_{t,j}^{(B)}$ for the Bth e-process and $\overline{\mathcal{K}}_{t,j}^{(B)} = \max_{s \leq t} \mathcal{K}_{s,j}^{(B)}$ for its running maximum, we stop at $\tau_{\mathrm{multi},j} = \inf\{t : \min_{B \neq A} \overline{\mathcal{K}}_{t,j}^{(B)} \geq 1/\alpha\}$.

Proposition 4 (Anytime multiclass validity)**.** *If every $\mathcal{K}_{t,j}^{(B)}$ is an e-process under its own pairwise null H_B, then $\mathbb{P}_P(\tau_{\mathrm{multi},j} < \infty) \leq \alpha$ for every P in the global null $\bigcup_{B \neq A} H_B$.*

Proof. Under the global null some pairwise null, say the one indexed by $B^\star$, is true. Certifying globally requires every pair to cross, so it entails $\sup_t \mathcal{K}_{t,j}^{(B^\star)} \geq 1/\alpha$, an event of probability at most α by Ville's inequality applied to that one e-process. □

The proof needs no assumption about how the $C-1$ margins depend on one another, and in practice they are strongly dependent: one draw of $\mathbf{X}_t$ updates every pair. A round therefore costs one model evaluation, and only the bookkeeping grows with C.

Sharper hard-label test. Let μ_A and μ_B be the probabilities of predicting classes A and B under the smoothing noise. We test the null $\mu_A \leq T_{2\rho}(\mu_B)$ from the tight Gaussian certificate of Cohen et al. (2019). This is less conservative than applying the NP certificate to the pairwise margin (§ I).

We build pairwise e-processes using § 5. For hard labels, we project the three-outcome law (target, competitor, remainder) onto the exact two-mean null. For softmax margins, we test margin-certificate nulls, with VC and CDF using information beyond the mean. § I gives projection formulas, validity proofs, and the certificates supported by each output type.

7 EXPERIMENTS

Setup. From each of the CIFAR-10 test and ImageNet validation sets (Krizhevsky, 2009; Deng et al., 2009) we take 1,000 evenly spaced images and draw 10,000 independent Gaussian-noise softmax evaluations per image, from ResNet-110 and ResNet-50 base models respectively (He et al., 2016), in the smoothing setup of Cohen et al. (2019). Unless stated otherwise $\sigma = 0.25$, $r = 0.125$ (so $\rho = 0.5$; § G varies both), $\tau = 0.5$, $\alpha = 10^{-3}$, $K = 100$ CDF bins, score prefixes shared across methods, and a full-support Dirichlet working histogram. Our **E-Cert** family comprises three direct-witness designs: **E-Cert-P** is the prespecified-*portfolio* e-process (§ D), **E-Cert-R** the *RIPr-guided* design of § 5, and **E-Cert-K** the pairwise design that keeps a working-law witness and a quadratic-*Kelly* stake (§ I.2). Here R denotes RIPr guidance; E-Cert-R (no refit) freezes the working law in the same construction. Unless stated otherwise, E-Cert-R refits every round through 200, then geometrically; both scalar e-process designs update once per score. The plug-in comparators are named for the bound they invert: the **fixed-n** and **staged** plug-ins for the fixed-n bounds of § D.2 used once and at ten checkpoints, and the **CS plug-in** for the confidence-sequence versions, with the sequence named where two are compared (§ I.5). Both E-Cert and the CS plug-ins use betting; they differ in whether they test the certificate directly or first bound its components. Because $\tau = 1/2$, certifying the target score forces it to be the smoothed argmax, so certified accuracy (CA), the percentage of the 1,000 images certified, is the percentage provably predicted correctly.

Two conventions govern the tables. First, α applies to each prespecified input–certificate decision on its own, not to a table jointly. Second, the target class is fixed to the ground truth before sampling, as is standard for certified accuracy. The class-partitioning method (CPM) and Cohen CERTIFY instead pick it from an initial n_0 draws. Charging every method for that selection leaves E-Cert-R a 1.1–2.3× sampling advantage (§ I.16).

7.1 SINGLE-CERTIFICATE RESULTS: CERTIFIED ACCURACY AND STOPPING COST

Capital trajectories. On the image selected by the prespecified rule in § E, E-Cert-R reaches $1/\alpha$ at a median of 29.5 evaluations versus 41 for E-Cert-P across ten score permutations (Figure 1).

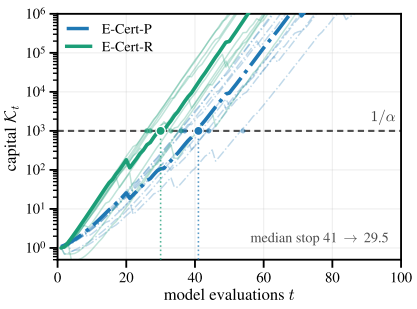


Figure 1: Capital $\mathcal{K}_t$ of the VC arms of E-Cert-P and E-Cert-R on one CIFAR-10 image, ten score permutations each, against $1/\alpha$; same witness family and scores (§ E).

The fixed-n plug-in comparators are empirical-Bernstein NP and VC bounds (VC splitting α across its two estimands) and a one-sided DKW band for CDF (Maurer & Pontil, 2009; Massart, 1990).

Certificate-specific witnesses improve small-budget accuracy. At $n = 100$ both e-process designs achieve higher certified accuracy than the prespecified fixed-n plug-in on every dataset–certificate pair, by 6–26 points; by $n = 10^4$ the methods have largely converged (Table 1). The controls below attribute most of this gain to the witness construction. A terminal-only control (our witness, rejecting only on the terminal capital $\mathcal{K}_n$) keeps 93.1–99.5% of it, leaving at most 0.7 pp to monitoring (§ F). Much of the rest is slack in the empirical-Bernstein bound: a fixed-n betting bound tuned for n trails by about 2 pp for NP and 8–10 pp for VC, and tracks the CS plug-in to within 0.6 pp (Table 7). For VC at $n = 100$, componentwise empirical Bernstein gives 28.4/29.4% (CIFAR-10/ImageNet), the same two estimands with a betting stake 44.0/42.6%, one direct boundary witness 53.2/50.6%, and RIPr on top of it 53.7/51.4%; swapping that witness's quadratic-Kelly stake for the exact one moves it by at most 0.1 pp (§ F.1). Two steps carry the gain: a tighter estimator for the same two components is worth 13–16 pp, and replacing them with one direct witness a further 8–9 pp. RIPr adds a smaller

Table 1: Certified accuracy (%) at budget n. Fixed-n tests are planned separately at each budget; sequential tests may certify earlier. Bold: best per dataset–certificate–budget, including ties. Comparator definitions: § D.2.

		Fixed-n plug-in			CS plug-in			E-Cert-P			E-Cert-R		
Dataset	Cert.	10^2	10^3	10^4	10^2	10^3	10^4	10^2	10^3	10^4	10^2	10^3	10^4
CIFAR-10	NP	36.7	53.8	56.3	49.2	**54.9**	55.9	50.7	**54.9**	**56.5**	**51.1**	**54.9**	56.4
	VC	28.4	54.6	59.7	44.0	58.9	60.6	52.0	**59.6**	**61.0**	**54.2**	**59.6**	**61.0**
	CDF	44.2	56.5	60.1	44.8	56.5	60.1	52.1	59.0	59.9	**54.4**	**59.7**	**61.1**
ImageNet	NP	36.2	48.8	50.0	45.1	**49.5**	50.1	46.5	**49.5**	**50.2**	**46.9**	**49.5**	**50.2**
	VC	29.4	51.6	54.9	42.7	53.9	55.9	50.3	**55.2**	**56.1**	**51.4**	**55.2**	**56.1**
	CDF	43.2	52.3	54.9	45.6	53.7	55.5	49.2	54.5	55.1	**51.3**	**55.2**	**55.9**

benefit at this radius: in the matched control its mean cost after futility is within 0–2.9% of the same witness, which in wall-clock leaves NP and VC tied within measurement noise and costs 7–17× on CDF, whose projection sets a break-even per-evaluation time that § D.3 quantifies. Certified accuracy orders differently at a harder threshold: repeating the same control at $\rho = 3$, where the CS plug-in and the cheap witness both certify nothing, RIPr certifies 10.8% and 18.8% (§ F.1).

Certified accuracy across radii and smoothing levels. We evaluate all four arms over a $\sigma \times r$ grid, using a separate Gaussian-augmented base classifier and score cache at each smoothing level (Figure 2, § G). Accuracy differences increase with ρ. Over 132 cells E-Cert-R leads the fixed-n plug-in in 101 and trails in one, by up to 21.3 points; it leads the CS plug-in in 85 and trails in four, each by one image, by up to 19.0 points. That plug-in shares its anytime validity, and for NP it is the same bet on $X_t - T_\rho(\tau)$ with a different stake rule, so the NP columns compare stake rules alone. In nine cells the fixed-n plug-in certifies nothing where E-Cert-R certifies a positive fraction; in eight of these the continuously monitored CS plug-in also certifies nothing. The NP design grid must resolve $T_\rho(\tau)$: if its largest cell center is no larger than this threshold, the stake is identically zero (Prop. 5). A threshold-aware grid fixed from ρ and τ avoids this failure; at $\rho = 3$, E-Cert-R certifies 0.6–5.9% across the four tested dataset–σ cells, ahead of E-Cert-P in each (§ B.2). For CDF at $\rho = 3$, E-Cert-R is the only compared arm that certifies any images.

Evaluation cost and stopping time. Figure 3 and Table 16 report evaluation cost over ten prespecified within-image permutations and all 1,000 images per dataset. An image uncertified at 10,000 is charged that cap. The *staged plug-in* monitors Table 1's fixed-n bound at ten checkpoints, each at level $\alpha/10$; the CS plug-in is inspected after every score (§ D.2).

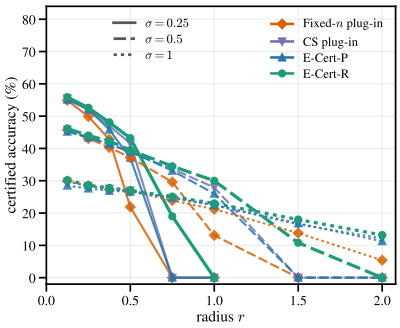


Figure 2: CDF certified accuracy versus radius on ImageNet at $n = 10^4$. Colors and markers denote methods; line styles denote σ. Full sweep: § G.

Fixing the accuracy instead of the budget, E-Cert-R reaches the fixed-n plug-in's full-budget accuracy 1.3–18.2× sooner, while the staged plug-in never reaches it: its ten checkpoints cost more accuracy than monitoring returns (§ H.2). Both e-processes beat the staged plug-in on median and mean in all 60 dataset–certificate–seed comparisons, and E-Cert-R beats the CS plug-in in all 60 too. Per image, on the inputs both arms certify, it stops earlier than the staged plug-in on *every* one, by a median factor of 4.3–9.5 (Figure 6). Part of that is structural, since a staged arm cannot stop between checkpoints: its best tested grid over $m \in \{10, 25, 50, 99, 230\}$ checkpoints at α/m improves its median by 0–21% and reverses no seed-level comparison (§ H.3). The median factors of 1.5–2.4 against the CS plug-in and 1.2–1.4 against E-Cert-P, both inspected every round, are not structural, and our arm is earlier on 89–100% of the paired images in every single-class cell (Figure 7 plots the medians). Against E-Cert-P, every seed-level median and mean drops for VC and CDF. For NP the two designs are comparable, E-Cert-R having the lower cost in 8 and E-Cert-P in 12 of the 20 dataset–seed comparisons.

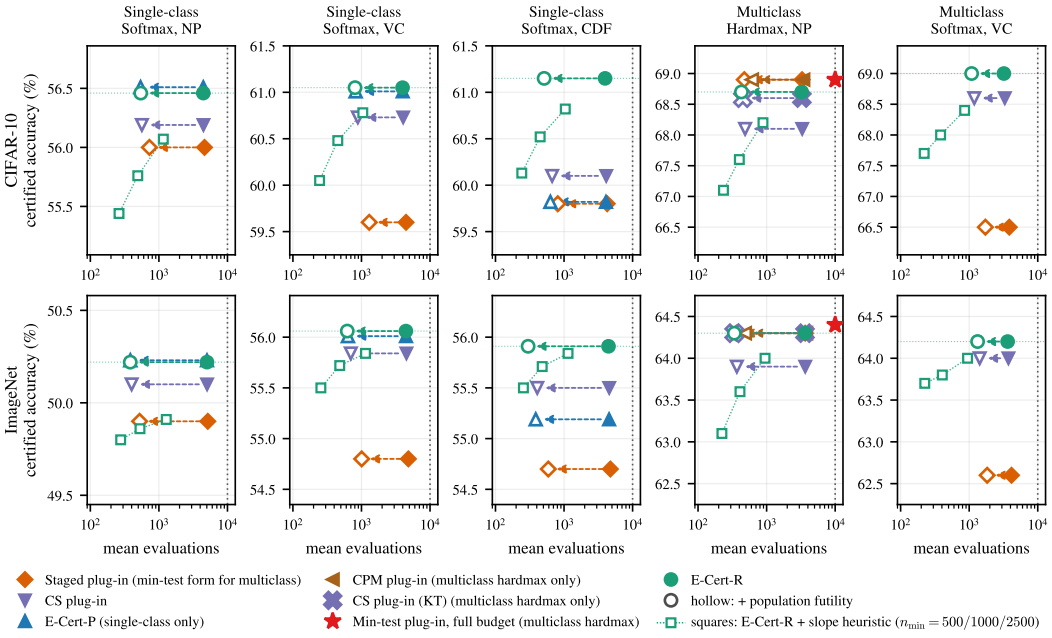


Figure 3: Certified accuracy versus mean evaluations over all 1,000 images, with unresolved inputs charged the 10^4 cap (dotted line). Arrows connect each method to its population-futility variant, reducing cost without observed certificate loss; hollow squares use slope-based futility. The red star is the fixed-budget min-test at level α. Single-class points average ten score orders (horizontal bars: ± 1 SD); multiclass points use one order. Vertical ticks are 0.5 percentage points apart; axis limits differ across certificate groups.

Futility stopping reduces budget exhaustion without observed certificate loss. The means above are dominated by the 39–50% of images that no arm certifies. The population futility rule of § 4 reads the score stream alone, so any arm can take it. With $\beta_f = \alpha$ split over six checkpoints from 100 to 5,000 it resolves 34–48% of images by abstention, leaves 1.7–6.0% at the cap, and cuts every arm's mean cost by 3.4–14.9×, to 299–1,299 evaluations. Over all 240,000 image–permutation pairs it never abandons an image its base arm would have certified within the budget, so certified accuracy is identical in every cell (Table 16). Futility accounts for most of this saving. With the rule applied to all arms, E-Cert-R retains a mean cost 1.0–1.4× below the CS plug-in's and 1.3–2.0× below the staged plug-in's. Zero certificate loss is an empirical result; Prop. 3 gives a probabilistic bound. An end-to-end boundary-law audit observes no trigger in 20,000 streams per certificate (§ H.4). The hollow squares in Figure 3 are the cheaper slope heuristic of § 4, which undercuts the population rule by 1.2–5.9× at $n_{\min} = 500$ but discards 0.4–1.6% of certificates (§ H.4).

Without futility, the e-processes spend a mean of 40–50% of the cap, and a median below 1.2% of it in five of six cells. Three checks leave this unchanged: three orders of magnitude of α, where E-Cert-R has the lowest median cost in 70 of 72 comparisons (§ H.5); a tenfold larger budget (§ H.6); and consistency-regularized checkpoints (§ H.7).

7.2 Multiclass argmax: Bonferroni, class partitioning, and e-processes

The hardmax benchmark uses the main σ, ρ and α, with 1,000 images per dataset, 10,000 class identities each and $C = 10/1{,}000$. Comparators are staged classwise Clopper–Pearson plug-ins (Bonferroni and min-test), CPM with a 100-draw partitioning prefix (Delattre et al., 2025), and continuously monitored CS plug-ins with empirical-Bernstein stakes or the KT sequence of Voráček (2024). The CS plug-ins use intersection–union; staged decisions correct for multiple checkpoints. The two E-Cert-R arms use the same exact two-mean tests and clean-softmax initialization: one freezes the working law, the other refits on a doubling schedule. We plot the stronger min-test form

of the classwise plug-in; at this operating point it differs from the Bonferroni form by under 15 mean evaluations and 0.1 pp (Table 24). Protocols and controls: § I.5–§ I.14.

Certified accuracy and evaluation cost across all images. By $n = 10^4$ the three staged plug-ins, the CS plug-in (KT) and E-Cert-R certify essentially the same images (0.688–0.689 and 0.686/0.687 on CIFAR-10, 0.643 for all five on ImageNet, at most three images apart paired; § I.6), so the anytime formulation costs no certified accuracy even at $C = 1{,}000$. On the images all arms certify, E-Cert-R's median is 19/20 against the Bonferroni plug-in's 50/50 (36/41 at its best tested grid), the min-test plug-in's 50/50 (32/32), CPM's 250/250 (50 at its best initial-phase size), the CS plug-in's 38/38 and the CS plug-in (KT)'s 27/27 (2.0/1.9× and 1.4×; Table 23, § I.13, § I.15). E-Cert-R seeds its working law from the clean forward pass and the plug-ins do not, so we also run it with a uniform prior instead: the medians become 22 and 21 and the accuracies change by at most 0.2 pp, so the gap is not that side information (§ I.8). Charging *every* image, the six arms excluding the no-refit ablation incur a mean cost of 3,248–3,772: the 31–36% they cannot certify cost 10,000 each. The argmax futility rule reverses the plug-in's two bounds, upper on the target and lower on the top observed competitor, so it upper-bounds the achievable radius (§ I.7). Added to E-Cert-R it cuts the mean from 3,248 to 429 and from 3,629 to 338, with no certificate lost, and to 337–614 for the plug-ins. After futility, E-Cert-R's mean cost is 1.1–1.4× lower than the other arms except the CS plug-in (KT), which is comparable (1.05/1.00×; Table 24).

Online refitting learns an effective working law without a separate pilot. The E-Cert-R (no refit) arm reaches only 0.539 on CIFAR-10 and 0.267 on ImageNet, both intervals well clear of the rest; the two arms differ only in whether the working law is refit. A seed-quality ablation separates the working law from the number of refits (§ I.8): a *single* design estimated from a 100-draw pilot, never refit, already reaches 0.677 and 0.636 (93.2% and 98.1% of the gap), and a full-stream oracle gives 0.689 and 0.644. Online refitting learns *while* accumulating evidence: the pilot's median stop is 119 against 19/20, largely the pilot itself.

Limits of futility under an inefficient design. Added to the E-Cert-R (no refit) arm the same rule saves far less (4,644 → 1,825 and 7,352 → 4,062) and leaves 17%/40% of images at the cap. Refitting recovers most of them (148 of 169 and 376 of 395), so they were not futile; the rest stay undecided. Refitting and futility are complements, not substitutes.

Sampling cost of the competitor-wise correction. Applying the $\alpha/(C-1)$ split to our own construction leaves accuracy within 0.3 points. It raises the median stop from 20 to 26 on CIFAR-10 and 23 to 42 on ImageNet. As the competitor count C' increases, the min-test's median is nearly constant, 24 then 23 throughout on ImageNet, while the split variant's grows approximately log-linearly. Checkpoint monitoring masks much of this cost in the staged plug-in forms (§ I.9, Table 26, § I.13).

Continuous softmax margins. The same construction applies to continuous margins $Z_t^{(B)} = \frac{1}{2}(1 + f_A - f_B)$; Figure 3 shows VC and Table 27 all three. Against the staged min-test plug-in the e-process arms certify 0–2.5 pp more images at 2–12× lower all-images median cost, 40–113 evaluations against 250–500 (5–12× for E-Cert-R of § I.2). Against the CS plug-in per pair, whose medians are 73–124, E-Cert-K is close (0.8–1.3×) and E-Cert-R needs 1.5–2.7× fewer evaluations, within 1 pp of certified accuracy. Futility again loses no certificate.

8 Conclusion

We test randomized smoothing certificates directly and stop sampling as soon as the evidence supports certification. Our betting construction controls false certification under arbitrary stopping, so we can adapt the computational budget to each input. The approach covers mean, mean–variance and CDF certificates and extends to multiclass decisions without a classwise Bonferroni correction. On CIFAR-10 and ImageNet, our tests achieve comparable or higher certified accuracy with similar or fewer samples than adaptive baselines, with the largest accuracy gains at small budgets. Direct certificate witnesses account for most of these gains; continuous monitoring contributes relatively little. RIPr guidance offers further gains at more demanding thresholds. Our futility rule bounds the risk of abandoning a certifiable input. In our experiments, it lowers mean evaluation cost for every method, including the baselines, with no observed loss of certificates. Future work could extend these tests to regression and provide anytime-valid bounds on certified radii.

AI USE STATEMENT

Following the ICLR 2027 policy we disclose how a generative-AI assistant was used in preparing this work, and how each use was checked. The research questions, the certificate constructions, the choice of baselines and comparators, and the experimental design are the authors' own, and the authors take full responsibility for the entire content of the paper.

The assistant drafted and polished exposition, including condensing appendix material into the main text and copy-editing. It helped locate related work; every citation was read and verified by the authors against the primary source, with quotations taken verbatim from it. It implemented and ran the analysis and sampling code under the authors' direction, including its numerical projection routines, self-tests and validity audits. It suggested candidate ablations and robustness checks, which the authors selected, ran and interpreted. It also helped derive and check steps in the proofs.

Two safeguards apply throughout. All theorem and proposition statements and their proofs were verified by the authors line by line, and the constructions they describe are additionally checked numerically at run time (§ C, § B). The analysis code includes consistency checks against reference experimental outputs.

ETHICS STATEMENT

This work studies statistical validity of certification procedures for randomized smoothing; it introduces no new data collection involving human subjects. § J discusses limitations and the broader-impact considerations we are aware of, including the scope of the population-level robustness guarantee itself and the risk of misreading "anytime validity" as license for uncorrected multiple testing across targets or thresholds.

REPRODUCIBILITY STATEMENT

The anonymous supplementary package provides code for computing Table 1 and Figures 1–3 from Monte Carlo score streams. Its README specifies the input format, dependencies, and commands; score generation and precomputed data or outputs are not included. § D.1 documents the base classifiers and score generation, and § D.2 specifies the comparators and aggregation. Proofs appear in the main text and § A, § B, and § I.

REFERENCES

Motasem Alfarra, Adel Bibi, Philip H. S. Torr, and Bernard Ghanem. Data dependent randomized smoothing. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 64–74. PMLR, 2022.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009. doi: 10.1016/j.tcs.2009.01.016.

Teodora Baluta, Zheng Leong Chua, Kuldeep S. Meel, and Prateek Saxena. Scalable quantitative verification for deep neural networks. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pp. 312–323. IEEE, 2021. doi: 10.1109/ICSE43902.2021.00039.

Roger L. Berger. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24(4):295–300, 1982. doi: 10.1080/00401706.1982.10487790.

Roger L. Berger and Jason C. Hsu. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11(4):283–319, 1996. doi: 10.1214/ss/1032280304.

Avrim Blum, Travis Dick, Naren Manoj, and Hongyang Zhang. Random smoothing might be unable to certify ℓ_∞ robustness for high-dimensional images. *Journal of Machine Learning Research*, 21(211):1–21, 2020.

Nicholas Carlini, Florian Tramer, Krishnamurthy Dj Dvijotham, Leslie Rice, Mingjie Sun, and J. Zico Kolter. (Certified!!) adversarial robustness for free! In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=JLg5aHHv7j.

Ruoxin Chen, Jie Li, Junchi Yan, Ping Li, and Bin Sheng. Input-specific robustness certification for randomized smoothing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6295–6303, 2022. doi: 10.1609/aaai.v36i6.20579.

Ping-yeh Chiang, Michael Curry, Ahmed Abdelkader, Aounon Kumar, John Dickerson, and Tom Goldstein. Detection as regression: Certified object detection with median smoothing. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1275–1286. Curran Associates, Inc., 2020.

Eugenio Clerico, Hamish E. Flynn, and Patrick Rebeschini. Uniform mean estimation for monotonic processes, 2025. URL https://arxiv.org/abs/2502.01244.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1310–1320. PMLR, 2019. URL https://proceedings.mlr.press/v97/cohen19c.html.

Andrew Cullen, Paul Montague, Shijie Liu, Sarah Erfani, and Benjamin I. Rubinstein. Double bubble, toil and trouble: Enhancing certified robustness through transitivity. In *Advances in Neural Information Processing Systems*, volume 35, pp. 19099–19112. Curran Associates, Inc., 2022.

Andrew C. Cullen, Paul Montague, and Benjamin I. P. Rubinstein. Halt fast! Early stopping for certified robustness. arXiv preprint arXiv:2606.27694, 2026.

D. A. Darling and Herbert Robbins. Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences*, 58(1):66–68, 1967. doi: 10.1073/pnas.58.1.66.

Blaise Delattre, Paul Caillon, Quentin Barthélemy, Erwan Fagnou, and Alexandre Allauzen. Bridging the theoretical gap in randomized smoothing. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 3997–4005. PMLR, 2025. URL https://proceedings.mlr.press/v258/delattre25a.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://doi.org/10.1109/CVPR.2009.5206848.

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

A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27(3):642–669, 1956. doi: 10.1214/aoms/1177728174.

Yixuan Fan, Zhanyi Jiao, and Ruodu Wang. Testing the mean and variance by e-processes. *Biometrika*, 112(1): asae049, 2025. doi: 10.1093/biomet/asae049.

Lasse Fischer and Aaditya Ramdas. Sequential Monte Carlo testing by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4):1200–1220, 2025. doi: 10.1093/jrsssb/qkaf014.

Axel Gandy. Sequential implementation of Monte Carlo tests with uniformly bounded resampling risk. *Journal of the American Statistical Association*, 104(488):1504–1511, 2009. doi: 10.1198/jasa.2009.tm08368.

Peter Grünwald, Rianne de Heide, and Wouter Koolen. Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5):1091–1128, 2024. doi: 10.1093/jrsssb/qkae011.

Peter D. Grünwald. Beyond Neyman–Pearson: E-values enable hypothesis testing with a data-driven alpha. *Proceedings of the National Academy of Sciences*, 121(39):e2302098121, 2024. doi: 10.1073/pnas.2302098121.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Miklós Z. Horváth, Mark Niklas Müller, Marc Fischer, and Martin Vechev. Boosting randomized smoothing with variance reduced classifiers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=mHu2vIds_-b.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. doi: 10.1214/18-PS321.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet S. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

Christopher Jennison and Bruce W. Turnbull. *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall/CRC, Boca Raton, FL, 1999. doi: 10.1201/9780367805326.

Jongheon Jeong and Jinwoo Shin. Consistency regularization for certified robustness of smoothed classifiers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 10558–10570. Curran Associates, Inc., 2020.

Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. Peeking at A/B tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1517–1525. ACM, 2017. doi: 10.1145/3097983.3097992.

Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. Always valid inference: Continuous monitoring of A/B tests. *Operations Research*, 70(3):1806–1821, 2022. doi: 10.1287/opre.2021.2135.

Wouter M. Koolen and Peter Grünwald. Log-optimal anytime-valid E-values. *International Journal of Approximate Reasoning*, 141:69–82, 2022. doi: 10.1016/j.ijar.2021.09.010.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Aounon Kumar, Alexander Levine, Soheil Feizi, and Tom Goldstein. Certifying confidence via randomized smoothing. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5165–5177, 2020a.

Aounon Kumar, Alexander Levine, Tom Goldstein, and Soheil Feizi. Curse of dimensionality on randomized smoothing for certifiable robustness. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5458–5467. PMLR, 2020b.

John M. Lachin. A review of methods for futility stopping based on conditional power. *Statistics in Medicine*, 24(18):2747–2764, 2005. doi: 10.1002/sim.2151.

K. K. Gordon Lan, Richard Simon, and Max Halperin. Stochastically curtailed tests in long-term clinical trials. *Communications in Statistics. Part C: Sequential Analysis*, 1(3):207–219, 1982. doi: 10.1080/07474948208836014.

Tyron Lardy, Peter Grünwald, and Peter Harremoës. Reverse information projections and optimal e-statistics. *IEEE Transactions on Information Theory*, 70(11):7616–7631, 2024. doi: 10.1109/TIT.2024.3444458.

Martin Larsson, Aaditya Ramdas, and Johannes Ruf. The numeraire e-variable and reverse information projection. *The Annals of Statistics*, 53(3):1015–1043, 2025. doi: 10.1214/24-AOS2487.

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

Guang-He Lee, Yang Yuan, Shiyu Chang, and Tommi Jaakkola. Tight certificates of adversarial robustness for randomly smoothed classifiers. In *Advances in Neural Information Processing Systems*, volume 32, pp. 4911–4922, 2019.

Alexander Levine and Soheil Feizi. (De)Randomized smoothing for certifiable defense against patch attacks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6465–6475. Curran Associates, Inc., 2020.

Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Linyi Li, Jiawei Zhang, Tao Xie, and Bo Li. Double sampling randomized smoothing. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 13163–13208. PMLR, 2022.

Linyi Li, Tao Xie, and Bo Li. SoK: Certified robustness for deep neural networks. In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 1289–1310. IEEE, 2023. doi: 10.1109/SP46215.2023.10179303.

Tudor Manole and Aaditya Ramdas. Martingale methods for sequential estimation of convex functionals and divergences. *IEEE Transactions on Information Theory*, 69(7):4641–4658, 2023. doi: 10.1109/TIT.2023.3250099.

Diego Martinez-Taboada and Aaditya Ramdas. Sharp empirical bernstein bounds for the variance of bounded random variables, 2026. URL https://arxiv.org/abs/2505.01987.

Pascal Massart. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, 18 (3):1269–1283, 1990. doi: 10.1214/aop/1176990746.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample-variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, pp. 115–124, 2009. URL https://www.cs.mcgill.ca/~colt2009/papers/012.pdf.

Volodymyr Mnih, Csaba Szepesvári, and Jean-Yves Audibert. Empirical Bernstein stopping. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 672–679. ACM, 2008. doi: 10.1145/1390156.1390241.

Jeet Mohapatra, Ching-Yun Ko, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. Higher-order certification for randomized smoothing. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4501–4511. Curran Associates, Inc., 2020.

Jeet Mohapatra, Ching-Yun Ko, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. Hidden cost of randomized smoothing. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 4033–4041. PMLR, 2021.

Francesco Orabona and Kwang-Sung Jun. Tight concentrations and confidence sequences from the regret of universal portfolio. *IEEE Transactions on Information Theory*, 70(1):436–455, 2024. doi: 10.1109/TIT.2023.3330187.

Aaditya Ramdas and Ruodu Wang. Hypothesis testing with E-values. *Foundations and Trends in Statistics*, 1 (1-2):1–390, 2025. doi: 10.1561/3600000002.

Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter Koolen. Admissible anytime-valid sequential inference must rely on nonnegative martingales. arXiv preprint arXiv:2009.03167, 2020.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023. doi: 10.1214/23-STS894.

Herbert Robbins. Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409, 1970. doi: 10.1214/aoms/1177696786.

Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Hadi Salman, Mingjie Sun, Greg Yang, Ashish Kapoor, and J. Zico Kolter. Denoised smoothing: A provable defense for pretrained classifiers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21945–21957. Curran Associates, Inc., 2020.

Yan Scholten, Jan Schuchardt, Aleksandar Bojchevski, and Stephan Günnemann. Hierarchical randomized smoothing. In *Advances in Neural Information Processing Systems*, volume 36, pp. 49783–49813. Curran Associates, Inc., 2023.

Jan Schuchardt, Tom Wollschläger, Aleksandar Bojchevski, and Stephan Günnemann. Localized randomized smoothing for collective robustness certification. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=-k7Lvk0GpBl.

Glenn Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(2):407–431, 2021. doi: 10.1111/rssa.12647.

Glenn Shafer and Vladimir Vovk. *Game-Theoretic Foundations for Probability and Finance*. Wiley, 2019.

Shubhanshu Shekhar and Aaditya Ramdas. On the near-optimality of betting confidence sets for bounded means. arXiv preprint arXiv:2310.01547, 2023.

Peter Súkeník, Aleksei Kuvshinov, and Stephan Günnemann. Intriguing properties of input-dependent randomized smoothing. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20697–20743. PMLR, 2022.

Chenhao Sun, Yuhao Mao, Mark Niklas Mueller, and Martin Vechev. Average certified radius is a poor metric for randomized smoothing. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 57651–57666. PMLR, 2025.

Karim Tit, Teddy Furon, and Mathias Rousset. Efficient statistical assessment of neural network corruption robustness. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9253–9263. Curran Associates, Inc., 2021.

Shubham Ugare, Tarun Suresh, Debangshu Banerjee, Gagandeep Singh, and Sasa Misailovic. Incremental randomized smoothing certification. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2305.19521.

Jean Ville. *Étude critique de la notion de collectif*. Gauthier-Villars, 1939.

Václav Voráček. Treatment of statistical estimation problems in randomized smoothing for adversarial robustness. In *Advances in Neural Information Processing Systems*, volume 37, pp. 133464–133486, 2024. doi: 10.52202/079017-4242. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f11394cdd377aab9ff5e2a4e9f27367f-Abstract-Conference.html.

Václav Voráček and Matthias Hein. Sound randomized smoothing in floating-point arithmetic. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=HaHCoGcpV9.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021. doi: 10.1214/20-AOS2020.

Abraham Wald. *Sequential Analysis*. John Wiley & Sons, 1947.

Larry Wasserman, Aaditya Ramdas, and Sivaraman Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020. doi: 10.1073/pnas.1922664117.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024. doi: 10.1093/jrsssb/qkad009.

Stefan Webb, Tom Rainforth, Yee Whye Teh, and M. Pawan Kumar. A statistical approach to assessing neural network robustness. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1xcx3C5FX.

Lily Weng, Pin-Yu Chen, Lam Nguyen, Mark Squillante, Akhilan Boopathy, Ivan Oseledets, and Luca Daniel. PROVEN: Verifying robustness of neural networks with a probabilistic approach. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6727–6736. PMLR, 2019.

Greg Yang, Tony Duan, J. Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10693–10705. PMLR, 2020.

Runtian Zhai, Chen Dan, Di He, Huan Zhang, Boqing Gong, Pradeep Ravikumar, Cho-Jui Hsieh, and Liwei Wang. MACER: Attack-free and scalable robust training via maximizing certified radius. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rJx1Na4Fwr.

A PROOFS FOR THE BOUNDED-WITNESS CONSTRUCTION

A.1 PROOF OF PROP. 1

Fix $P \in H$ and condition on $\mathcal{F}_{t-1}$. The predictable map u_t and stake are then fixed. Nonnegativity holds by assumption, and

$$\mathbb{E}_P[e_t \mid \mathcal{F}_{t-1}] = 1 + \lambda_t \mathbb{E}_P[U_t \mid \mathcal{F}_{t-1}] \le 1.$$

Thus $\mathcal{K}_t$ is a nonnegative supermartingale with $\mathcal{K}_0 = 1$. Ville's inequality gives $\mathbb{P}_P(\sup_{t\ge 0} \mathcal{K}_t \ge 1/\alpha) \le \alpha$. Taking the supremum over $P \in H$ proves the stated bound.

Two normalizations make the support condition explicit. If $Z_t \in [0,1]$, $0 < c \le 1$, and the null implies $\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] \le c$, then

$$e_t = 1 + \lambda_t(Z_t - c), \qquad 0 \le \lambda_t \le 1/c,$$

is nonnegative. If instead $0 \le c < 1$ and the null implies $\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] \ge c$, use

$$e_t = 1 + \lambda_t(c - Z_t), \qquad 0 \le \lambda_t \le 1/(1-c).$$

For an omitted endpoint, apply its structural null separately; the neutral factor is always valid. Finite convex mixtures of these e-processes remain e-processes.

A.2 NP NULL REDUCTION AND SUPPORT BOUND

Since Φ and Φ^{-1} are increasing,

$$\begin{aligned} \mathrm{Cert}_{\mathrm{NP}}(P;\rho) \le \tau &\iff \Phi\big(\Phi^{-1}(M_P) - \rho\big) \le \tau \\ &\iff \Phi^{-1}(M_P) - \rho \le \Phi^{-1}(\tau) \\ &\iff M_P \le \Phi\big(\Phi^{-1}(\tau) + \rho\big) = T_\rho(\tau). \end{aligned}$$

Thus, for every $P \in H_{\mathrm{NP}}$, $\mathbb{E}_P[U_t^{\mathrm{NP}} \mid \mathcal{F}_{t-1}] = M_P - T_\rho(\tau) \le 0$. Moreover, for every $x \in [0,1]$,

$$1 + \lambda_t\{x - T_\rho(\tau)\} \ge 1 - \lambda_t T_\rho(\tau) \ge 0$$

whenever $0 \le \lambda_t \le 1/T_\rho(\tau)$.

A.3 PROOF AND SUPPORT BOUND FOR THE VC WITNESS

Condition on the past and abbreviate (m_t, v_t, β_t) by (m, v, β). Let (M_P, V_P) be the moments of $P \in H_{\mathrm{VC}}$. The null and safe-side inequalities give

$$M_P - m \le \kappa_\rho(\sqrt{V_P} - \sqrt{v}). \tag{7}$$

Direct expansion yields

$$\begin{aligned} \mathbb{E}_P[Y_t(X_t)] - D_t &= \beta(V_P - v) - (M_P - m) + \beta(M_P - m)^2 \\ &\ge \frac{\kappa_\rho}{2\sqrt{v}}(\sqrt{V_P} - \sqrt{v})^2 + \beta(M_P - m)^2 \ge 0, \end{aligned}$$

where the inequality uses (7) and $\beta = \kappa_\rho/(2\sqrt{v})$. This proves the claimed conditional-mean bound.

For $u_t(x) = D_t - Y_t(x)$, global nonnegativity of $1 + \lambda_t u_t(x)$ follows from the fact that u_t is concave in x. Its minimum on $[0,1]$ is therefore $u_{\min,t} = \min\{u_t(0), u_t(1)\}$. If $u_{\min,t} < 0$, the exact endpoint-safe range is

$$0 \le \lambda_t \le -1/u_{\min,t};$$

if $u_{\min,t} \ge 0$, nonnegativity imposes no upper bound (although a bounded implementation may still cap the stake). The witness is a supporting normal to the VC bad-null boundary at equality (and a separating normal on the certified side), avoiding a single-observation variance estimate.

A.4 CDF CONCAVITY AND TANGENT REDUCTION

Let ϕ denote the standard normal density and $z = \Phi^{-1}(u)$. For $u \in (0,1)$,

$$T'_\rho(u) = \frac{\phi(z+\rho)}{\phi(z)} = \exp\left(-\rho z - \frac{\rho^2}{2}\right),$$
$$T''_\rho(u) = -\rho\,\frac{\phi(z+\rho)}{\phi(z)^2}.$$

Thus T_ρ is increasing and, for $\rho > 0$, strictly concave. (At $\rho = 0$ it is the identity, so the following tangent inequality still holds with equality.) For every $q \in (0,1)$ and $u \in [0,1]$,

$$T_\rho(u) \le T_\rho(q) + T'_\rho(q)(u-q). \tag{8}$$

At $u \in \{0,1\}$ this follows by continuity. Applying (8) to each $F_{P,i}$ shows that $P \in H_{\mathrm{CDF}}$ implies

$$a_t + \sum_{i=1}^{K-1} w_{i,t} F_{P,i} \ge b_{K-1} - \tau, \qquad \sum_{i=1}^{K-1} w_{i,t} F_{P,i} \ge d_t.$$

Because $\mathbb{E}_P[\sum_{i=1}^{K-1} w_{i,t}\mathbf{1}\{X_t \le b_i\} \mid \mathcal{F}_{t-1}] = \sum_{i=1}^{K-1} w_{i,t} F_{P,i}$, the increment in (5) has conditional mean at most zero. All $w_{i,t}$ are nonnegative. With $W_t = \sum_{i=1}^{K-1} w_{i,t}$, $Z_t = \sum_{i=1}^{K-1} w_{i,t}\mathbf{1}\{X_t \le b_i\}/W_t$, and $c_t = d_t/W_t$, the nondegenerate case $0 < c_t < 1$ uses

$$e_t = 1 + \widetilde{\lambda}_t(c_t - Z_t), \qquad 0 \le \widetilde{\lambda}_t \le 1/(1-c_t).$$

Because $c_t - Z_t = U_t^{\mathrm{CDF}}/W_t$, this normalized stake equals the main-text stake through $\lambda_t = \widetilde{\lambda}_t/W_t$; hence $\lambda_t \le 1/(W_t - d_t)$. Degenerate tangent choices can be replaced by the neutral factor.

B PREDICTABLE REVERSE INFORMATION PROJECTION

B.1 FINITE-ALPHABET CONSTRUCTION AND PROOF

For an L-cell alphabet, let $J_t \in \{1,\dots,L\}$ be the cell containing X_t, let $N_{t-1,k}$ be the number of previous scores in cell k, and fix a Dirichlet pseudocount $\xi > 0$. Define

$$\widehat{p}_{t,k} = \frac{N_{t-1,k} + \xi}{t - 1 + L\xi}, \qquad \boldsymbol{p}_t^\star \in \arg\min_{\boldsymbol{p} \in H_j^L(\rho,\tau)} \mathrm{KL}(\widehat{\boldsymbol{p}}_t \| \boldsymbol{p}), \qquad e_t(k) = \frac{\widehat{p}_{t,k}}{p^\star_{t,k}}. \tag{9}$$

Here $H_j^L(\rho,\tau)$ is the induced categorical null and is assumed nonempty, closed, and convex; the selected projection has positive coordinates. If $\widehat{\boldsymbol{p}}_t \in H_j^L(\rho,\tau)$, we set $\boldsymbol{p}_t^\star = \widehat{\boldsymbol{p}}_t$ and obtain the neutral factor. At $t = 1$, $\widehat{\boldsymbol{p}}_t$ is the full-support prior; the first score is still processed and updates the next working law. We use L here for the number of cells in this auxiliary finite-alphabet *design* discretization, distinct from the K CDF-certificate grid cells fixed in Eq. 3. K is part of the population CDF certificate's own definition, while L is a computational device used only to choose the next witness and stake. In the experiments of § 7 $L = K = 100$, but nothing in the construction requires them to coincide. Neither value was tuned. Sweeping K over $\{25, 50, 100, 200, 500\}$ moves the mean fixed-n CDF certificate monotonically upward ($0.5116 \to 0.5307$ on CIFAR-10, $0.4602 \to 0.4774$ on ImageNet at $n = 1{,}000$). The discretization is therefore conservative as derived, and $K = 100$ leaves 0.3–0.4 pp of certified accuracy on the table relative to $K = 500$; E-Cert-R's median stop moves by at most five evaluations over $K \in \{25,\dots,200\}$.

We fix the Dirichlet pseudocount at $\xi = 0.01$ per cell without tuning. Any $\xi > 0$ gives a full-support working law based on completed scores; validity follows from Prop. 1. We assess its effect on efficiency below. Re-running the deployed loop at $\xi \in \{0.001, 0.01, 0.1, 1\}$ on all 1,000 images per dataset leaves certified accuracy flat – no cell moves by more than 0.3 percentage points over that range – while cost is flat near the default and degrades above it: the median stop over certified images is 22–27 at $\xi \le 0.01$, 1.5–1.7× larger at $\xi = 0.1$ and 5.0–6.1× larger at $\xi = 1$, where a pseudocount comparable to the first hundred observations holds the working law near the uniform prior for hundreds of rounds. Thus performance is stable near the default, while larger pseudocounts delay certification. Per-cell results are included in the release.

Fix the past and abbreviate $\widehat{\boldsymbol{p}}_t$ by $\widehat{\boldsymbol{p}}$. For any $\boldsymbol{\pi} \in H_j^L(\rho, \tau)$, convexity makes $\boldsymbol{p}_\epsilon = (1-\epsilon)\boldsymbol{p}^\star + \epsilon\boldsymbol{\pi}$ feasible. First-order optimality of $\boldsymbol{p}^\star$ gives

$$0 \le \frac{\mathrm{d}}{\mathrm{d}\epsilon}\mathrm{KL}(\widehat{\boldsymbol{p}}\|\boldsymbol{p}_\epsilon)\Big|_{\epsilon=0+} = 1 - \sum_{k=1}^{L} \pi_k \frac{\widehat{p}_k}{p_k^\star}.$$

Consequently, $\mathbb{E}_{\boldsymbol{\pi}}[e_t(J_t) \mid \mathcal{F}_{t-1}] = \sum_k \pi_k \widehat{p}_k / p_k^\star \le 1$ for every null law $\boldsymbol{\pi}$. Repeating this argument conditionally proves that the product of the predictable factors in (9) is an e-process.

This likelihood-ratio result is exact only for a genuinely finite score alphabet, or for a conservative quantization whose categorical null contains every induced null law. It is not the factor used in our continuous-score experiments: there, a past-only histogram working law selects a supporting direction and legal stake, while Prop. 1 validates the resulting factor over every $x \in [0,1]$. Thus the practical factors are RIPr-guided and make no exact finite-alphabet log-growth claim.

B.2 Design-grid resolution at large ρ

The design grid affects power but not validity (Prop. 1). For NP, a uniform grid can force the stake to remain zero: its largest cell center must exceed the certification threshold. Prop. 5(i) establishes this condition, which depends only on the prespecified ρ, τ and grid. The rule below selects a grid that satisfies it and is implemented by `online.minimal_power`.

Write $\theta = T_\rho(\tau)$ for the NP threshold, let the design grid have cell centers $c_1 < \cdots < c_L$ and half-width $\eta = \max_k \frac{1}{2}|\mathrm{cell}_k|$, and confine stakes to $[0, \bar{\lambda}]$ with $\bar{\lambda} < 1/\theta$, so that $\kappa = 1 - \bar{\lambda}\theta > 0$ and every factor is at least κ. The deployed stake is the working-law Kelly choice $\lambda_t \in \arg\max_{0\le\lambda\le\bar{\lambda}} \mathbb{E}_{\widehat{P}_t} \log\{1 + \lambda(X-\theta)\}$, and $\widehat{P}_t$ is the pseudocount-smoothed histogram of the completed scores, a law on $\{c_1, \dots, c_L\}$.

Proposition 5 (Grid resolution and log-growth)**.**

(i) (Unresolved grid.) *If $c_L \le \theta$, then $\mathbb{E}_{\widehat{P}_t}[X] \le \theta$ for* every *law on the grid, so $\lambda_t = 0$ at every round and $\mathcal{K}_t \equiv 1$. The arm certifies for no budget, no significance level and no data.*

(ii) (Resolved grid.) *Suppose instead $M_P \ge \theta + \gamma$ for some $\gamma > 0$, and put $g_0 = \max_{0\le\lambda\le\bar{\lambda}} \mathbb{E}_P \log\{1 + \lambda(X-\theta)\} > 0$. If $g_0 > 2\bar{\lambda}\eta/\kappa$, then almost surely $\liminf_t t^{-1} \log \mathcal{K}_t \ge g_0 - 2\bar{\lambda}\eta/\kappa > 0$; consequently, writing $T_\alpha = \inf\{t : \mathcal{K}_t \ge 1/\alpha\}$ for the stopping time (distinct from the certificate threshold τ), $T_\alpha < \infty$ almost surely for every α and $\limsup_{\alpha\downarrow 0} T_\alpha / \log(1/\alpha) \le (g_0 - 2\bar{\lambda}\eta/\kappa)^{-1}$.*

Proof. (i) $\mathbb{E}_{\widehat{P}_t}[X] = \sum_k q_k c_k \le c_L \le \theta$ for any probability vector q. The Kelly objective $\widehat{G}_t(\lambda) = \mathbb{E}_{\widehat{P}_t} \log\{1 + \lambda(X-\theta)\}$ is concave with $\widehat{G}_t'(0) = \mathbb{E}_{\widehat{P}_t}[X] - \theta \le 0$, so its maximizer over $[0, \bar{\lambda}]$ is $\lambda_t = 0$ and every factor is 1.

(ii) Let $G(\lambda) = \mathbb{E}_P \log\{1 + \lambda(X-\theta)\}$. Then $G(0) = 0$ and $G'(0) = M_P - \theta \ge \gamma > 0$, so $g_0 > 0$. On $[0, \bar{\lambda}]$ the map $x \mapsto \log\{1 + \lambda(x-\theta)\}$ is Lipschitz with constant $\lambda/\{1 + \lambda(x-\theta)\} \le \bar{\lambda}/\kappa$, so $\sup_{\lambda\le\bar{\lambda}} |G(\lambda) - \widehat{G}_t(\lambda)| \le (\bar{\lambda}/\kappa) W_1(\widehat{P}_t, P) =: \varepsilon_t$, with W_1 the Wasserstein-1 distance. Quantizing to cell centers moves mass by at most η, and the pseudocount contributes $O(L\xi/t)$, so $\limsup_t W_1(\widehat{P}_t, P) \le \eta$ almost surely by the Glivenko–Cantelli theorem, whence $\limsup_t \varepsilon_t \le \bar{\lambda}\eta/\kappa$. (Only the $\limsup$ is available: W_1 need not converge, and η is an upper bound on the quantization error rather than its limit.) Let λ_0 attain g_0. Since λ_t maximizes $\widehat{G}_t$,

$$G(\lambda_t) \;\ge\; \widehat{G}_t(\lambda_t) - \varepsilon_t \;\ge\; \widehat{G}_t(\lambda_0) - \varepsilon_t \;\ge\; G(\lambda_0) - 2\varepsilon_t \;=\; g_0 - 2\varepsilon_t.$$

Finally $\log \mathcal{K}_t = \sum_{s\le t}[\log\{1 + \lambda_s(X_s - \theta)\} - G(\lambda_s)] + \sum_{s\le t} G(\lambda_s)$. Each uncentered term lies in $[\log\kappa, \log\{1 + \bar{\lambda}(1-\theta)\}]$, so each *centered* increment of the first sum lies in the difference of that interval with itself, $[\log\kappa - \log\{1 + \bar{\lambda}(1-\theta)\},\ \log\{1 + \bar{\lambda}(1-\theta)\} - \log\kappa]$; it is a martingale with bounded increments and hence $o(t)$ almost surely. The second sum is at least $t(g_0 - 2\bar{\varepsilon}_t)$ with $\bar{\varepsilon}_t = t^{-1}\sum_{s\le t}\varepsilon_s$, and since $\varepsilon_t \le \bar{\lambda}/\kappa$ is bounded with $\limsup_t \varepsilon_t \le \bar{\lambda}\eta/\kappa$, also $\limsup_t \bar{\varepsilon}_t \le \bar{\lambda}\eta/\kappa$. Dividing by t gives the growth claim; a positive growth rate makes T_α finite, and $\log \mathcal{K}_{T_\alpha} \ge \log(1/\alpha)$ with $\liminf_t t^{-1}\log\mathcal{K}_t > 0$ gives the stated rate. □

Part (i) is the $\rho \geq 3$ NP regime, and only that. With $\tau = 1/2$ the NP threshold is $\theta = \Phi(\rho)$, so $\theta = 0.99865$ at $\rho = 3$, while a uniform $L = 100$ grid has $c_L = 0.995$: no working law on it can leave the NP null, the Kelly stake is identically zero, and the arm would report nothing however many scores it draws. E-Cert-P, which fixes its stakes in advance rather than reading them off a working law, is unaffected, so a uniform grid would leave it certifying where a working-law design cannot.

VC and CDF do not have the same obstruction. The maximum certificate value over grid-supported laws is attained by a point mass on the top cell. At $\rho = 3$ that is 0.336 for NP, below τ and hence (i), but 0.995 for VC and 0.990 for CDF, both far above it. To assess their behavior, we also compute the *full-stream empirical certificate*: the fraction of images whose certificate at the empirical law of all 10^4 draws clears τ, read from the raw moments for NP and VC and from the K-cell definition (3) for CDF. At $\rho = 3$ and $\sigma = 0.25/0.5$ it is: NP $8.4/1.2\%$ on CIFAR-10 and $7.5/4.1\%$ on ImageNet, VC $9.4/1.2\%$ and $13.5/7.5\%$, CDF $26.1/7.8\%$ and $38.2/25.4\%$. Against those figures, E-Cert-P reaches $5.6/0.4\%$ and $4.5/2.8\%$ on NP and nothing elsewhere, and the fixed-n and CS plug-ins reach nothing anywhere. E-Cert-R reaches $10.8/1.7\%$ and $19.0/10.8\%$ on CDF, where it is the *only* arm that certifies at all, $5.9/0.6\%$ and $5.5/3.1\%$ on NP at the grid the design rule below selects, and nothing on VC.

At $\rho = 3$ that reference is not a bound on what a valid procedure could reach. A certificate evaluated at the empirical law carries no coverage of its own, and $\kappa_\rho = \sqrt{e^{\rho^2} - 1} = 90.0$ there, so it magnifies any finite-sample shortfall in $\sqrt{V}$: a law with $P(X = 0) = 4 \times 10^{-5}$ has population VC certificate $0.431 < \tau$, yet returns an empirical certificate of 1 on the all-ones stream it produces with probability 0.67. These results distinguish three cases: a proven grid obstruction for NP, an empirical reference gap for VC that does not establish a power deficit, and positive CDF certification only by E-Cert-R. The 151 images certified by NP with a resolved grid have full-stream means from 0.99947 to 1 (median 0.99993). These descriptive statistics agree with the decisions; the error guarantee comes from the e-process.

Increasing the number of uniform cells does not restore certification within the reported budget. A uniform grid resolves θ once $L > 1/\{2(1-\theta)\}$, i.e. $L \geq 371$ at $\rho = 3$, so $L = 500$ and $L = 1000$ clear (i)'s bar; both still certify nothing. The reason is that L controls two things at once. The working law is the Dirichlet-smoothed histogram $\widehat{\boldsymbol{p}}_t$ of (9). Its $L\xi$ units of pseudocount sit at the prior and pull its mean toward the middle of $[0, 1]$, so $\mathbb{E}_{\widehat{P}_t}[X]$ approaches the truth only at rate $L\xi/t$ and a finer grid carries proportionally more of it. Measured on these streams, the working mean first exceeds θ at $t = 7{,}126$–$8{,}041$ at $L = 500$ and $t = 5{,}868$–$6{,}634$ at $L = 1000$, and the margin it then holds over θ is at most 3.5×10^{-4}, so the stake is near zero for whatever budget remains. These delays are measured under the deployed schedule. They are not predictions of part (ii), whose sufficient condition is not met at the deployed stake cap, as quantified below.

A nonuniform grid resolves the threshold while keeping L and the total pseudocount $L\xi$ fixed. Equal pseudocounts per bin also change the prior: its mean rises from 0.500 at $p = 1$ to 0.667 at $p = 2$ and 0.750 at $p = 3$. The experiment therefore compares the joint effect of grid resolution and prior shape; it does not isolate their individual contributions. Clearing (i) is also necessary rather than sufficient: a grid with $c_L > \theta$ can still round a null-violating law back inside the null. At $\rho = 3$ and $p = 2$ the law with $P(X = 0.9983) = 13/17$ and $P(X = 1) = 4/17$ has mean $0.99870 > \theta$ but gridded mean $0.99842 < \theta$, so its working law would stop betting however long the stream ran. Part (ii) excludes this behavior under $g_0 > 2\lambda\eta/\kappa$, a sufficient condition that is not met with our deployed cap $\kappa = 10^{-8}$. Edges at $1-(1-k/L)^p$ with $L = 100$ and $p \in \{2, 3\}$ give $c_L = 0.99995$ and 0.9999995, while leaving $L\xi$ at its deployed value. The working mean then clears θ by $t = 185$–450, and the arm certifies. Table 2 reports all five grids over all $1{,}000$ images of both reported $\rho = 3$ cells. At $p = 2$ the arm certifies $5.9/0.6\%$ of CIFAR-10 and $5.5/3.1\%$ of ImageNet images at $\sigma = 0.25/0.5$, against 0% at every uniform grid and $5.6/0.4\%$ and $4.5/2.8\%$ for E-Cert-P. With this grid, E-Cert-R exceeds E-Cert-P's certified accuracy in all four cells.

The design rule. We select the grid from the certification threshold. With edges $1-(1-k/L)^p$ the largest center is $c_L = 1 - \frac{1}{2}L^{-p}$, so (i) is cleared exactly when $L^{-p} < 2(1-\theta)$, and the deployed design takes the smallest integer $p \geq 1$ that satisfies it,

$$p(\rho) \;=\; \max\Big\{1,\; \Big\lfloor \frac{\log\{2(1 - T_\rho(\tau))\}}{\log(1/L)} \Big\rfloor + 1 \Big\}.$$

Table 2: The $\rho = 3$ cells of § G, E-Cert-R's NP arm, at five design grids. Cells are numbers of images certified out of 1,000, in percent; c_L is the largest cell center and $L\xi$ the total Dirichlet pseudocount. The NP threshold is $\theta = \Phi(3) = 0.998650$, so a grid with $c_L \le \theta$ has Kelly stake identically zero (Prop. 5(i)). Edges are $1 - (1 - k/L)^p$; the design rule selects $p = 1$, the uniform grid, at every reported cell with $\rho \le 2$ and $p = 2$ here. The grid is a design device fixed before sampling, so every column certifies the same claim at the same level.

			CIFAR-10		ImageNet	
Design grid	c_L	$L\xi$	σ=0.25	σ=0.5	σ=0.25	σ=0.5
uniform L=100	0.9950000	1.0	0.0	0.0	0.0	0.0
uniform L=500	0.9990000	5.0	0.0	0.0	0.0	0.0
uniform L=1000	0.9995000	10.0	0.0	0.0	0.0	0.0
power p=2, L=100	0.9999500	1.0	5.9	0.6	5.5	3.1
power p=3, L=100	0.9999995	1.0	6.0	0.6	5.7	3.1
E-Cert-P (no grid)	–	–	5.6	0.4	4.5	2.8

The floor-plus-one enforces the strict inequality: $L^{-p} < 2(1-\theta)$ is strict, so when $\log\{2(1-\theta)\}/\log(1/L)$ lands on an integer a, the power a puts c_L exactly *at* θ and (i) still applies. That happens whenever $2(1-\theta)$ is a power of $1/L$; at $L = 100$ and $\tau = 1/2$ it is $\rho = \Phi^{-1}(0.995) = 2.5758$, which no reported cell hits but which a ceiling would resolve incorrectly. The implementation verifies the power it names against the strict inequality rather than trusting a floating-point floor at that boundary. The rule depends only on ρ, τ and L, fixed before sampling, so Prop. 1 covers it unchanged. It returns $p = 1$, the uniform grid, at every reported (ρ, τ) setting of § G with $\rho \le 2$, which is 18 of the 22 (σ, r) pairs; every number in § G outside those four cells is a uniform-grid one for that reason, not by a separate convention. The rule also binds the multiclass softmax NP arm at $\rho \ge 3$ (§ I.12), where the uniform grid leaves the Kelly stake identically zero. At the four remaining cells it returns $p = 2$ at $\rho = 3$ and $p = 3$ at $\rho = 4$, and those are the grids § G reports. The $\rho = 3$ cells are the ones above. At $\rho = 4$ the rule changes nothing, and the obstruction there is the budget rather than the grid or the images. The NP bet is linear in $X - \theta$ with stakes capped at $\bar\lambda = (1 - 10^{-8})/\theta$, so the most a single draw can do is multiply capital by $1 + \bar\lambda(1-\theta)$, and reaching $1/\alpha$ takes at least $\log(1/\alpha)/\log\{1 + \bar\lambda(1-\theta)\}$ draws even on the all-ones stream: 218,105 at $\rho = 4$ against a budget of 10^4. Thus NP cannot certify at this budget, regardless of the observed scores; this does not imply that the population certificate fails. The same floor is 5,114 at $\rho = 3$, inside the budget by only a factor of two, which is why the resolved grid certifies 5.9 and 0.6% on CIFAR-10 and 5.5 and 3.1% on ImageNet ($\sigma = 0.25$ and 0.5) rather than the double-digit accuracies the same arm reaches at $\rho \le 2$.

Part (ii) gives a qualitative dependence on grid error, but its sufficient condition does not cover the deployed constants. We cap stakes at $\bar\lambda = (1 - 10^{-8})/\theta$, so $\kappa = 10^{-8}$. At the main operating point ($\rho = 1/2$), its threshold is $2\bar\lambda\eta/\kappa \approx 1.45 \times 10^6$, whereas log-growth is bounded by $\log\{1 + \bar\lambda(1-\theta)\} \approx 0.369$. The condition therefore fails at the grid resolutions considered here. For a cap bounded away from $1/\theta$, the result describes how quantization error η affects growth. Part (i) identifies when the design must remain inactive; it gives no rate once betting becomes possible. § B.4 gives separate oracle stopping-time bounds for VC and CDF.

B.3 Certificate-specific geometry

For NP, the null is the half-space $\mathbb{E}X \le T_\rho(\tau)$. Its valid linear factors have the form

$$e_t(x) = 1 + \lambda_t\{x - T_\rho(\tau)\}, \qquad 0 \le \lambda_t \le 1/T_\rho(\tau).$$

RIPr therefore changes the stake, not the statistic. Under a working law $\widehat{P}_t$, the corresponding Kelly choice solves

$$\lambda_t \in \arg\max_{0 \le \lambda \le 1/T_\rho(\tau)} \mathbb{E}_{\widehat{P}_t} \log[1 + \lambda\{X_t - T_\rho(\tau)\}].$$

Relation to the binary hard-label case. When scores are binary (the hard label $X_t \in \{0, 1\}$, e.g. an indicator of predicting the target class) rather than continuous, H_{NP} reduces to a single Bernoulli

mean being at most $T_\rho(\tau)$. Voráček (2024) study exactly this case. Their betting confidence sequence stakes the Krichevsky–Trofimov running-mean estimate of the head probability on every score, which is the Kelly-optimal bet against that estimate. Against a *known* true mean it coincides with Wald's sequential probability ratio test (Wald, 1947). Our NP factor coincides with it when the working law $\widehat{P}_t$ is the Bernoulli law at their Krichevsky–Trofimov running mean *and* that mean exceeds γ; below γ ours takes the neutral factor one while theirs keeps staking, so the two *factors* coincide only on the rounds where the working mean is already in the alternative. The wealths do not follow: once one round differs the products part and later agreement between factors does not bring them back together. It is in that sense a working-law-relative Kelly bet, though they do not describe it as one. They prove it valid with width $\sqrt{\{\log(1/\alpha) + \log t\}/t}$, which is not asymptotically rate-optimal. They give a separate, non-betting construction (randomized Clopper–Pearson intervals at a doubling schedule with level $\alpha/\{k(k+1)\}$) that attains the law-of-the-iterated-logarithm-optimal rate instead. Empirically the betting sequence needs fewer samples in their benchmark, so it is the baseline we run in § I.5. RIPr keeps the same working-law-relative Kelly principle. It applies that principle to continuous $[0,1]$-valued scores and, through the VC and CDF geometries below, to certificates without a single sufficient scalar statistic. No comparable rate-optimal benchmark is available there, so § 5 treats RIPr as a design heuristic whose validity Prop. 1 supplies independently of any growth-optimality claim. § B.4 does bound the stopping time of a witness built at the alternative, which is a statement about the certificate's geometry rather than about RIPr's estimate of it.

Bernoulli hard-label specialization. For the hard-label score $X_t \in \{0,1\}$, write $p = \mathbb{E}[X_t]$ and $\gamma = T_\rho(\tau)$. The NP bad null is simply $p \le \gamma$. Given completed observations, a symmetric $\mathrm{Beta}(a,a)$ working law has posterior-predictive mean

$$\widehat{p}_t = \frac{a + \sum_{s<t} X_s}{2a + t - 1}. \tag{10}$$

If $\widehat{p}_t > \gamma$, the reverse projection of $\mathrm{Bern}(\widehat{p}_t)$ onto $\{p \le \gamma\}$ is $\mathrm{Bern}(\gamma)$, and its likelihood-ratio factor is

$$\begin{aligned} e^{\mathrm{hard}}_{t,\mathrm{RIPr}}(x) &= \left(\frac{\widehat{p}_t}{\gamma}\right)^x \left(\frac{1-\widehat{p}_t}{1-\gamma}\right)^{1-x} \\ &= 1 + \frac{\widehat{p}_t - \gamma}{\gamma(1-\gamma)}(x - \gamma). \end{aligned} \tag{11}$$

If $\widehat{p}_t \le \gamma$, we use the neutral factor one. Because $\widehat{p}_t$ is $\mathcal{F}_{t-1}$-measurable, this is exactly the predictable NP factor above, evaluated at the working mean instead of an arbitrary Bernoulli parameter. A symmetric prior initializes the process before score 1; the first score may have a neutral factor but is still used to form the round-2 working law.

For VC with $\rho > 0$, the population functional $P \mapsto M_P - \kappa_\rho \sqrt{V_P}$ is convex, so its sublevel null is convex. The log-growth-optimal design moments (m_t, v_t), found by maximizing working-law expected log-growth over the null-boundary curve $m - \kappa_\rho \sqrt{v} = \tau$ rather than by an explicit reverse projection, supply the supporting normal in (4). The calculation in § A.3 validates that direction for every $P \in H_{\mathrm{VC}}$, independently of numerical projection accuracy, provided the safe-side condition is enforced. At $\rho = 0$, $\kappa_\rho = 0$ and $\mathrm{Cert}_{\mathrm{VC}}(P;\rho) = M_P$, so the mean construction gives the degenerate case; all reported radii have $\rho > 0$.

For CDF, the null is convex because it is the superlevel set of the concave map $Q \mapsto \sum_{i=1}^{K-1} \Delta_i T_\rho(F_{Q,i})$. At an active projected boundary, the gradient gives the tangent weights in (5). The tangent argument in § A.4, rather than numerical equality at the solution, supplies the null expectation bound.

For the online threshold e-processes in Table 1 and Table 16, a Dirichlet working law on $L = 100$ cells with pseudocount 0.01 per cell is refit on a predictable schedule using only completed scores; each design is frozen until the next refit. Refit *frequency* is a pure efficiency choice, never a validity one: Prop. 1 supplies validity for any predictable design regardless of how often it is refit. Refitting only before rounds $1, 2, 4, \ldots, 8192$ (a geometric doubling schedule) incurs the largest observed efficiency loss in the first few hundred rounds, when most successful certifications occur. Refitting every round reduces this gap and empirically lowers E-Cert-R's stopping time (never raises it) on the paired comparisons we checked, but incurs substantial computational cost for streams that remain near the certification boundary, since each CDF refit is a constrained numerical projection. The

default schedule therefore refits every round through round 200, then at ≈ 100 geometrically spaced rounds out to the 10,000-evaluation horizon. A working law in the null, a nonpositive-growth design, or a numerical solver failure invokes the neutral factor. For each Table 16 permutation, this schedule restarts at the first permuted score.

B.4 Local efficiency of a direct witness

We derive a witness direction and a stopping-time bound for VC and CDF from the geometry of their nulls. The design is constructed at the true alternative, so this is an oracle analysis. It does not bound the cost of estimating the design from past scores. The fitted-design growth result in Prop. 5(ii) concerns NP and requires a condition not met by the deployed stake cap.

Write $\boldsymbol{\theta}_P \in \mathbb{R}^d$ for the moment vector the certificate reads, and $\boldsymbol{\phi}$ for the map that produces it, so $\boldsymbol{\theta}_P = \mathbb{E}_P[\boldsymbol{\phi}(X)]$:

$$\boldsymbol{\phi}(x) = (x, x^2), \quad d = 2 \quad \text{(VC)}; \qquad \boldsymbol{\phi}(x) = \big(\mathbf{1}\{x \le b_i\}\big)_{i<K}, \quad d = K-1 \quad \text{(CDF)}.$$

Both nulls are convex in these coordinates: for VC because $\boldsymbol{\theta} \mapsto \theta_1 - \kappa_\rho\sqrt{\theta_2 - \theta_1^2}$ is convex, for CDF because $\boldsymbol{\theta} \mapsto \sum_i \Delta_i T_\rho(\theta_i)$ is concave. Write $H \subset \mathbb{R}^d$ for the image of the null under $P \mapsto \boldsymbol{\theta}_P$, a closed convex set. Call a witness *moment-affine* when $U(x) = c - \langle \boldsymbol{u}, \boldsymbol{\phi}(x)\rangle$, so that $\mathbb{E}_Q U = c - \langle \boldsymbol{u}, \boldsymbol{\theta}_Q\rangle$ is affine in $\boldsymbol{\theta}_Q$; both (4) and (5) are, with $\boldsymbol{u} = (-1 - 2\beta_t m_t,\ \beta_t)$ and $c = D_t$ for VC and $\boldsymbol{u} = (w_{i,t})_i$, $c = d_t$ for CDF. Let $\Sigma_P = \mathrm{Cov}_P\{\boldsymbol{\phi}(X)\}$, assumed nonsingular, and let $\|\boldsymbol{z}\|_{\Sigma_P} = (\boldsymbol{z}^\top \Sigma_P \boldsymbol{z})^{1/2}$ with dual norm $\|\cdot\|_{\Sigma_P^{-1}}$.

Proposition 6 (The supporting normal maximizes the standardized margin)**.** *Fix an alternative $P \notin H$ and let*

$$\xi_P = \mathrm{dist}_{\Sigma_P^{-1}}(\boldsymbol{\theta}_P, H) = \inf_{\boldsymbol{\theta}\in H} \|\boldsymbol{\theta}_P - \boldsymbol{\theta}\|_{\Sigma_P^{-1}}$$

be the Mahalanobis margin of P to the null, in the metric the moment map's own covariance induces. Then, over moment-affine witnesses U that are valid ($\mathbb{E}_Q U \le 0$ for every Q in the null) and not identically zero,

$$\sup_U \frac{\mathbb{E}_P U}{\sqrt{\mathrm{Var}_P(U)}} = \xi_P,$$

and the supremum is attained by the outward supporting normal of H at the $\|\cdot\|_{\Sigma_P^{-1}}$-projection $\boldsymbol{\theta}^\star$ of $\boldsymbol{\theta}_P$ onto H, with $c = \langle \boldsymbol{u}, \boldsymbol{\theta}^\star\rangle$. This is the direction (4) and (5) take.

Proof. Validity of $U(x) = c - \langle \boldsymbol{u}, \boldsymbol{\phi}(x)\rangle$ over the null reads $c \le \inf_{\boldsymbol{\theta}\in H}\langle \boldsymbol{u}, \boldsymbol{\theta}\rangle$, so the largest admissible c is that infimum, and then $\mathbb{E}_P U = \inf_{\boldsymbol{\theta}\in H}\langle \boldsymbol{u}, \boldsymbol{\theta}\rangle - \langle \boldsymbol{u}, \boldsymbol{\theta}_P\rangle = \inf_{\boldsymbol{\theta}\in H}\langle \boldsymbol{u}, \boldsymbol{\theta} - \boldsymbol{\theta}_P\rangle$. Also $\mathrm{Var}_P(U) = \boldsymbol{u}^\top \Sigma_P \boldsymbol{u} = \|\boldsymbol{u}\|_{\Sigma_P}^2$, which does not depend on c. The ratio is invariant to positive rescaling of $\boldsymbol{u}$, so

$$\sup_U \frac{\mathbb{E}_P U}{\sqrt{\mathrm{Var}_P(U)}} = \sup_{\|\boldsymbol{u}\|_{\Sigma_P}\le 1} \inf_{\boldsymbol{\theta}\in H} \langle \boldsymbol{u}, \boldsymbol{\theta} - \boldsymbol{\theta}_P\rangle = \mathrm{dist}_{\Sigma_P^{-1}}(\boldsymbol{\theta}_P, H),$$

the last equality being the support-function formula for the distance from a point to a closed convex set, in the norm dual to $\|\cdot\|_{\Sigma_P}$. A maximizing $\boldsymbol{u}$ separates $\boldsymbol{\theta}_P$ from H and is attained at the projection, where it is an outward supporting normal. □

Proposition 7 (Growth rate and stopping time at the oracle design)**.** *Let U be a bounded witness with $\Delta = \mathbb{E}_P U > 0$, $S = \mathbb{E}_P U^2$ and $U \ge u_{\min}$ with $u_{\min} < 0$, bet at a fixed stake $\lambda \in [0, \min\{\bar\lambda,\ 1/(2|u_{\min}|)\}]$ on i.i.d. draws from P. Then $\mathbb{E}_P \log(1 + \lambda U) \ge \lambda\Delta - \lambda^2 S$. If $\Delta/(2S)$ is admissible, the stake $\lambda = \Delta/(2S)$ gives log-growth $g \ge \Delta^2/(4S)$, and $T_\alpha = \inf\{t : \mathcal{K}_t \ge 1/\alpha\}$ obeys $T_\alpha < \infty$ almost surely with*

$$\limsup_{\alpha\downarrow 0} \frac{T_\alpha}{\log(1/\alpha)} \le \frac{1}{g} \le \frac{4S}{\Delta^2}.$$

Writing $\xi = \Delta/\sqrt{\mathrm{Var}_P(U)}$, the bound is $4(1+\xi^2)/\xi^2$, and at the direction of Prop. 6 $\xi = \xi_P$.

Proof. For $y \geq -1/2$, $h(y) = \log(1+y) - y + y^2$ has $h(0) = 0$ and $h'(y) = y(1+2y)/(1+y)$, which is nonnegative on $[0, \infty)$ and nonpositive on $[-1/2, 0)$; hence $h \geq 0$ there, i.e. $\log(1+y) \geq y - y^2$. The stake range makes $\lambda U \geq -1/2$, so taking expectations gives the first claim. Maximizing $\lambda\Delta - \lambda^2 S$ over $\lambda \geq 0$ gives $\lambda = \Delta/(2S)$ and value $\Delta^2/(4S)$. Then $\log \mathcal{K}_t$ is a random walk with drift $g > 0$ and bounded increments, so $t^{-1} \log \mathcal{K}_t \to g$ almost surely and $T_\alpha < \infty$. The growth inequality is non-asymptotic; the stopping-time statement is a $\limsup$ as $\alpha \to 0$, since it reads the rate off $\log \mathcal{K}_{T_\alpha} \geq \log(1/\alpha)$. Finally $S = \mathrm{Var}_P(U) + \Delta^2$, so $\Delta^2/S = \xi^2/(1+\xi^2)$. □

The bound is non-asymptotic, and we do not sharpen its constant by expanding the exact Kelly growth as $\xi \downarrow 0$, because boundedness and a positive mean do not license that expansion. Take $P(U = 1, -1, 0) = (3\epsilon, \epsilon, 1 - 4\epsilon)$, a bounded witness legal under the null it makes $\mathbb{E}U \leq 0$ at. Its optimal stake is $\frac{1}{2}$ for *every* ϵ, so it does not vanish with the margin, and the exact growth $\epsilon\{3\log 1.5 + \log 0.5\} = 0.5232\,\epsilon$ exceeds $\xi^2/\{2(1+\xi^2)\} = 0.5\,\epsilon$ by a term of the same order as the leading one, not a smaller one. A sharper constant would need a local family in which the optimal stake is interior and vanishing, with uniform control of the moments; we state only what the displayed inequality gives.

Three readings follow. First, the bound scales as ξ_P^{-2}: what it costs to certify an input is governed by how far its moment vector sits from the null, measured in the moment map's own covariance. Second, the direction attaining that margin is the supporting normal at the projection, which is what (4) and (5) bet on. That is a statement about the standardized margin and about the bound the margin controls, not about exact log-growth: $\max_\lambda \mathbb{E}_P \log(1 + \lambda U)$ is not a function of ξ alone, and the two objectives can disagree. At $\rho = \tau = 1/2$ on $P(X = 0, 0.6, 0.95) = (0.04, 0.30, 0.66)$, whose VC certificate is 0.685, the largest-ξ boundary point ($m = 0.6703$, $\xi = 0.6158$) has log-growth 0.08596, while $m = 0.6839$, with the *smaller* $\xi = 0.6132$, reaches 0.08653. Whether the deployed design attains exact growth optimality is a question these propositions do not settle. Third, for NP the two collapse to $d = 1$ and $\xi_P = \{M_P - T_\rho(\tau)\}/\mathrm{sd}_P(X)$, the standardized margin, which is the familiar form.

Association between margin and stopping time. We assess whether the standardized margin ξ_P predicts the stopping time of the fitted procedure. For every image at the main operating point we build the deployed design from the *full-stream* working law, read U off the resulting factor, and compute ξ on the raw scores. Over the 503–611 images certified in each of the six dataset–certificate cells, the Spearman correlation between ξ and the stopping time the published E-Cert-R took on the same image is -0.947 to -0.969. This supports ξ as an index of certification difficulty. Rank correlation cannot validate the constants in Prop. 7, since it is unchanged by a strictly monotone transformation. On levels, the bound $\log(1/\alpha)\,4(1+\xi^2)/\xi^2$ read at the oracle design sits above the observed stopping time at the median of every cell, by a factor of 1.06–1.28, and above it on 62–94% of individual images. It is not a finite-α guarantee for the deployed arm and we do not read it as one. The median ξ is 3.8–4.6 on CIFAR-10 and 9.7–11.2 on ImageNet, far outside any small-margin regime, and the gap between the two sides collects the slack in $\log(1+y) \geq y - y^2$, the overshoot past $\log(1/\alpha)$, the stake cap, and the cost of fitting the design from the past, none of which this comparison separates.

The analysis leaves two questions open. It does not bound the cost of replacing the oracle direction at $\boldsymbol{\theta}_P$ with a past-only estimate; Prop. 5(ii) addresses a fitted design only for NP. Componentwise confidence-bound plug-ins also lie outside the moment-affine class considered here. Their comparison with direct witnesses is empirical (§ F.1).

C Empirical false-certification rate under continuous monitoring

The validity results are accompanied by roundwise numerical checks of the witness inequality (the hardmax benchmark's refits are checked against their supporting inequality on a 1,001-point grid, at the eight competitors whose factors deviate most, worst excess 6.4×10^{-15}). A procedure that refits from data and solves numerical projections also needs an end-to-end check. This appendix supplies one. The deployed code, monitored after every draw and stopped at its first crossing, is run on synthetic streams whose law sits *exactly* on the null boundary. Its realized false-certification rate

Table 3: Realized false-certification rate against nominal level, on synthetic streams placed exactly on the relevant null boundary and stopped at the first crossing; all rows use E-Cert-R, the single-class rows on the deployed refit schedule. "One-boundary" puts the binding competitor on the boundary with the rest deep in the alternative; "all-boundary" puts every competitor on it. Single-class rows use each certificate family's own boundary. Counts and 95% Clopper–Pearson bounds are in the result CSV; no lower bound exceeds α.

			Nominal level α				
Test	Null configuration	Streams	0.2	0.1	0.05	0.01	0.001
Multiclass	$C = 10$, one-boundary	20,000	0.1424	0.0703	0.0330	0.0067	0.0007
	$C = 10$, all-boundary	20,000	0.0003	0.0001	0.0001	0.0000	0.0000
	$C = 1{,}000$, one-boundary	20,000	0.1424	0.0703	0.0330	0.0067	0.0007
NP	Two-point (wide)	20,000	0.1454	0.0712	0.0320	0.0076	0.0009
	Two-point (narrow)	20,000	0.1341	0.0646	0.0302	0.0069	0.0006
	Beta (concentrated)	20,000	0.1238	0.0601	0.0299	0.0052	0.0006
VC	Two-point (wide)	20,000	0.1449	0.0766	0.0398	0.0068	0.0007
	Two-point (narrow)	20,000	0.1346	0.0633	0.0312	0.0064	0.0008
	Beta (concentrated)	20,000	0.1286	0.0630	0.0317	0.0057	0.0006
CDF	Two-point (wide)	5,000	0.1372	0.0656	0.0322	0.0088	0.0002
	Two-point (narrow)	5,000	0.1096	0.0498	0.0214	0.0046	0.0006
	Beta (concentrated)	5,000	0.1184	0.0616	0.0296	0.0056	0.0006

is compared with the nominal level α, swept by overriding the single stopping threshold the code reads. Two facts about the boundary matter. The multiclass arms project onto the exact two-mean null $\mu_A \le T_{2\rho}(\mu_B)$, not onto the margin-NP null $\mathbb{E}[Z] \le \Phi(\rho)$; a law on the latter boundary lies strictly inside the former's alternative, so placing the audit there would report correct certifications as false ones. Likewise the three scalar nulls are not nested, so each certificate is audited on its *own* boundary, constructed by bisecting a parametric family (two-point laws at two spreads and a concentrated Beta) against the deployed certificate definitions.

Table 3 reports the result. For the multiclass min-test, the least favorable configuration puts the target-versus-binding-competitor pair exactly on the boundary ($p_A = T_{2\rho}(p_B)$, residual $\le 2 \times 10^{-16}$) and every other competitor deep in the alternative, at 20,000 streams per cell. It gives 0.1424, 0.0703, 0.0330, 0.0067 and 0.0007 against $\alpha = 0.2, 0.1, 0.05, 0.01, 0.001$. The rate at $C = 1{,}000$ equals the rate at $C = 10$ in every cell, on identical counts: the level is governed by the single binding pair and does not degrade as competitors are added. This is consistent with Prop. 4 and complements the competitor-count cost experiment in Table 26. Putting every competitor on the boundary is *more* conservative (0.00035 at $\alpha = 0.2$), since a false certification then requires every pair to cross at once.

For the single-class e-processes each certificate is audited on its own boundary, at 20,000 streams per cell for NP and VC and 5,000 for CDF, whose per-prefix search costs more than two orders of magnitude more per stream than the other two. The worst of the nine cells are 0.145, 0.077, 0.040, 0.0088 and 0.0009. At the deployed level $\alpha = 10^{-3}$ the worst cell is 18 false certifications in 20,000 streams, a rate of 9.0×10^{-4} with a 95% Clopper–Pearson upper bound of 1.3×10^{-3}; every other cell is lower. No cell's lower confidence bound exceeds its level, at any level, in either part of the audit. Across the 45 scalar cells the realized rate runs at 0.20–0.90 of nominal, so these tests are conservative on these boundaries without being vacuous.

These runs use the deployed dense-to-round-200 refit schedule of Table 1 and Table 16, so the audit exercises the code path the reported results use, with its few hundred data-dependent design choices per stream. Repeating it on the doubling schedule, which refits fourteen times, gives 0.130, 0.060, 0.033, 0.0083 and 0.0017 at 600 streams per cell; the last is one stream in 600 and is not significantly above its level. Both schedules are consistent with level control on the tested boundary laws. The audit does not establish validity beyond those laws.

D E-CERT-P CONSTRUCTIONS AND EXPERIMENTAL PROTOCOL

We specify E-Cert-P, the plug-in comparators, and the sampling and stopping protocols. E-Cert-P fixes its stake or tangent grid before score collection at the reported threshold τ. Each construction is a non-adaptive instance of Prop. 1; predictable online adaptation is described in § B.

E-Cert-P's CDF construction. It chooses its complete finite bank of tangents before score collection. For a predictable tangent $\boldsymbol{q}$, retain $a(\boldsymbol{q})$, $w_i(\boldsymbol{q})$, and $W(\boldsymbol{q}) = \sum_{i=1}^{K-1} w_i(\boldsymbol{q})$ from (5), and define

$$d(\boldsymbol{q}, \tau) = b_{K-1} - \tau - a(\boldsymbol{q}), \qquad c_{\boldsymbol{q}}(\tau) = d(\boldsymbol{q}, \tau)/W(\boldsymbol{q}).$$

When $0 < c_{\boldsymbol{q}}(\tau) < 1$ and $Z_t(\boldsymbol{q}) = \sum_{i=1}^{K-1} w_i(\boldsymbol{q})\mathbf{1}\{X_t \le b_i\}/W(\boldsymbol{q})$, the factor $1 + \lambda_t\{c_{\boldsymbol{q}}(\tau) - Z_t(\boldsymbol{q})\}$ is valid under the CDF bad null H_{CDF} whenever $0 \le \lambda_t \le 1/(1 - c_{\boldsymbol{q}}(\tau))$; inadmissible or degenerate tangents use factor one instead. For a prespecified (ρ, τ), it uses $q_i = F_{\mathrm{Beta}(a_{\mathrm{B}}, b_{\mathrm{B}})}(b_i)$, clipped to $[10^{-6}, 1 - 10^{-6}]$, for the 28 shapes

$$\begin{aligned}\mathcal{B} = {} & \{(1/2, 1/2)\} \cup \{(2^k, 2^\ell) : k = 0, \dots, 3,\ \ell = -1, \dots, k-1\} \\ & \cup \{(16, 1), (16, 2), (16, 4), (32, 1), (32, 2)\} \cup \big(\{1, 2, 4, 8, 16, 32\} \times \{1/8, 1/4\}\big).\end{aligned}$$

Each shape therefore defines the tangent $\boldsymbol{q} = (q_1, \dots, q_{K-1})$; a template is retained exactly when $0 < c_{\boldsymbol{q}}(\tau) < 1$, and each retained template includes the three legal factors

$$1 + \frac{h}{1 - c_{\boldsymbol{q}}(\tau)}\{c_{\boldsymbol{q}}(\tau) - Z_t(\boldsymbol{q})\}, \qquad h \in \{0.95, 0.99, 0.999\}.$$

The reported capital is the equal-weight mixture over all retained template–stake pairs; mixture weights are never renormalized post hoc, and different predictable choices may change growth but not coverage. Thus bank selection, stake selection, and mixture normalization depend only on (ρ, τ) and the fixed grid, never on the observed score stream.

E-Cert-P's NP and VC mixtures. Let $\mathcal{H}_{13} = \{0.05 + 0.075\ell : \ell = 0, \dots, 12\}$. The fixed NP e-process mixes equally over the 13 factors with $\lambda = h/T_\rho(\tau)$ for $h \in \mathcal{H}_{13}$.

The VC mixture uses the single-sample boundary witness of (4), the same witness family as E-Cert-R, at prespecified boundary points. A boundary point is a moment pair (m, v) with $m - \kappa_\rho\sqrt{v} = \tau$, parametrized by m: $v = ((m-\tau)/\kappa_\rho)^2$, and it is the moment pair of some $[0,1]$-law iff $v \le m(1-m)$, which holds for $m \in (\tau, m_{\max}]$ with $m_{\max}$ the larger root of $((m-\tau)/\kappa_\rho)^2 = m(1-m)$ ($m_{\max} = 0.735$ at $\rho = 0.5$). The mixture takes the 13 points $m_k = \tau + k(m_{\max} - \tau)/14$, $k = 1, \dots, 13$, and at each the 13 stakes $\lambda = h\,\lambda_{\max}(m_k)$, $h \in \mathcal{H}_{13}$, where $\lambda_{\max}(m) = -1/\min\{U(0), U(1)\}$ is the support-safe cap of § 4 (the witness is concave, so its minimum over $[0,1]$ is at an endpoint), and mixes equally over the $13 \times 13 = 169$ components. Each component is valid under H_{VC} by § A.3, so the mixture is; the grid depends only on (ρ, τ) and is fixed before score collection. One factor is evaluated per model evaluation, exactly as for E-Cert-R, so the two VC arms differ only in whether the boundary point and stake are chosen before the stream or refit from it.

D.1 SCORE CACHES AND BASE CLASSIFIERS

The experiments in § 7 use cached Monte Carlo score streams: for each fixed image, 10,000 independent $\mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I})$ perturbations are drawn (10^5 for § H.6), passed through a base classifier, and reduced either to a scalar or to the full softmax vector. The scalar is the target class's softmax entry, or the hard label, the indicator that the hardmax prediction is the target class. The real-data tables and figures are computed from these cached streams (§ D.2); none of that analysis depends on which base classifier or dataset produced them.

The smoothing noise is isotropic Gaussian, drawn in the original input coordinates, and the base classifier's per-dataset input normalization is applied *after* it, so a certified radius is a radius in those coordinates.

Model checkpoints. Every score cache in this paper comes from pretrained Gaussian-augmented base classifiers that their authors released publicly, downloaded from the official repositories and used without modification or retraining: those of Cohen et al. (2019) throughout, and additionally those of Jeong & Shin (2020) for § H.7. We take the released checkpoints trained at the same noise level used for sampling. The default experiments use the $\sigma = 0.25$ ResNet-110 (CIFAR-10) and ResNet-50 (ImageNet) checkpoints. The smoothing-level sweep in § G adds the corresponding $\sigma = 0.5$ and $\sigma = 1$ checkpoints for each dataset, with fresh score caches sampled at those levels. The two default $\sigma = 0.25$ checkpoints record 90 completed training epochs, with final base-classifier accuracy of 75.4% (CIFAR-10) and 68.7% (ImageNet) on noise-augmented evaluation data. Within each dataset and smoothing level, all certification methods read the same released base classifier.

D.2 Comparator and stopping-protocol details

Plug-in comparator definitions. For a candidate mean $m \in (0, 1]$, define

$$\mathcal{K}_t^M(m;\delta) = \prod_{s=1}^{t}\{1 + \lambda_s^M(m;\delta)(X_s - m)\}, \qquad \lambda_s^M(m;\delta) = \min\{\eta_s(\delta), (2m)^{-1}\},$$
$$L_{M,t}^{\mathrm{B}}(\delta) = \inf\{m : \max_{0\le\ell\le t} \mathcal{K}_\ell^M(m;\delta) < 1/\delta\}.$$

Following Waudby-Smith & Ramdas (2024), we use the predictable empirical-Bernstein base stake

$$\eta_s(\delta) = \sqrt{\frac{2\log(1/\delta)}{s\log(1+s)\widehat{v}_{s-1}}}.$$

The resulting inversion gives an anytime-valid lower mean bound. Here δ is the componentwise miscoverage level: NP uses $\delta = \alpha$, whereas VC combines level-$\alpha/2$ mean and variance confidence sequences. For CDF, the CS plug-in comparator instead uses the DKW-type confidence sequence described next: what has no analogue is inverting a scalar mean sequence, since the CDF certificate is a functional of the whole distribution, not of its mean. Specifically, $\widehat{\mu}_t = (0.5 + \sum_{u=1}^{t} X_u)/(t+1)$ and $\widehat{v}_t = \{0.25 + \sum_{u=1}^{t}(X_u - \widehat{\mu}_u)^2\}/(t+1)$, so $\widehat{v}_0 = 0.25$. We update the running intersection after every score from $s = 1$. Table 1 tests its fixed threshold directly, without numerical endpoint inversion. For the VC comparator, $\overline{V}_t^{\mathrm{B}}$ is the variance upper confidence sequence of Martinez-Taboada & Ramdas (2026, Corollary 4.3), using $(c_1, c_2, c_3, c_4) = (1/2, 1/16, 1/4, 1/2)$, and

$$\underline{\mathrm{Cert}}_{\mathrm{NP},t}^{\mathrm{B}} = \Phi\big(\Phi^{-1}(L_{M,t}^{\mathrm{B}}(\alpha)) - \rho\big), \qquad \underline{\mathrm{Cert}}_{\mathrm{VC},t}^{\mathrm{B}} = \left[L_{M,t}^{\mathrm{B}}(\alpha/2) - \kappa_\rho\sqrt{\overline{V}_t^{\mathrm{B}}(\alpha/2)}\right]_+.$$

Every fixed-n plug-in table entry is a separately planned terminal analysis at its displayed budget; it is not repeatedly monitored as a fixed-time bound.

The CDF confidence sequence. For CDF, the CS plug-in comparator is the DKW-type confidence sequence of Clerico et al. (2025). Write $F_P(y) = P(X_t \le y)$ for the population CDF at a threshold y, so $F_{P,i} = F_P(b_i)$. At every threshold y they bet on the coin $c_t(y) = \mathbf{1}\{X_t \le y\} - F_P(y)$ with a Krichevsky–Trofimov mixture forecaster, then merge the resulting per-y test martingales through a continuous Donsker–Varadhan union bound that stays non-vacuous because F_P is monotone in y. For Bernoulli observations, exactly our case since $\mathbf{1}\{X_t \le y\}$ is a 0/1 event at every threshold, their bound (their Eq. 7) coincides with their exact band (their Proposition 1) and reduces to

$$F_P(y) \le \min_{k_n(y)\le t\le n} \mathrm{kl}_+^{-1}\left(\frac{t}{n}, \frac{1}{n}\log\frac{2\sqrt{n}}{\delta\,(X_{(t+1)} - y)}\right) \qquad \text{simultaneously over all } n \ge 1 \text{ and } y,$$

where $k_n(y)$ counts, among the first n scores, those at or below y, $X_{(1)} \le \cdots \le X_{(n)}$ are the order statistics of the first n scores (with $X_{(n+1)} := 1$), and $\mathrm{kl}_+^{-1}(p, c) = \sup\{q \ge p : \mathrm{kl}(p, q) \le c\}$ inverts the binary relative entropy $\mathrm{kl}(p, q) = p\log(p/q) + (1-p)\log\{(1-p)/(1-q)\}$. We set $\delta = \alpha$, evaluate this bound at each grid point b_i, and plug the resulting upper bounds on $F_{P,i}$ into (3), in place of the fixed-n DKW band, to obtain an anytime-valid CDF certificate lower bound. The kl_+^{-1} root is located by 30 bisection steps, rounded outward so that finite-precision truncation can only widen, never tighten, the reported bound.

For Table 16, the staged plug-in is the fixed-n plug-in bound of Table 1 recomputed on the prefix at each of $\{10, 25, 50, 100, 250, 500, 1000, 2500, 5000, 10000\}$ evaluations. Each checkpoint uses level $\alpha/10$, divided equally between the VC mean and variance components, so the ten checkpoints are valid together by a union bound. The CS plug-in comparator uses level α for NP, level $\alpha/2$ for each VC component, and $\delta = \alpha$ for CDF. It is effectively monitored after *every* score for all three certificates. Checking it only at the ten-point staged grid would be exact for the staged monitor, but for a comparator meant to represent continuous monitoring it would overestimate the stopping time by up to the gap to the next checkpoint. Because the sequence is simultaneously valid over every sample size, this requires no error-budget split either way, unlike the staged monitor. Evaluating the CDF sequence at all $10{,}000$ prefixes is expensive: each evaluation rescans the whole score prefix, so an image the sequence never certifies costs about 46 minutes on its own on the reference 72-core machine used for this project. A coarse-to-fine search that checks a geometrically spaced grid and refines only the gap in which it *sees* the certificate above τ misses a crossing that reverts before that gap's right endpoint. For example, on the stream whose first 204 values are 0.5899 and whose rest are 0, the certificate is 0.500010 at $t = 204$ and 0.481418 at $t = 205$, so at the horizon $n = 10^4$ such a search returns nothing where per-sample monitoring returns 204. Missed crossings would overestimate this comparator's stopping cost or underestimate its certified accuracy.

We retain per-prefix monitoring and use an upper-bound screen to avoid unnecessary evaluations. Replace the gap $X_{(i+1)} - y$ in the band's width by the largest value it can take, $1 - y$, and the order-statistic index by the smallest feasible one. Both substitutions can only shrink the band, and so can only raise the certificate. The resulting upper bound costs $O(K)$ per prefix instead of $O(nK)$, and is computed for a whole block of prefixes in one vectorized pass, with blocks growing geometrically until a crossing is found or the horizon is reached. A prefix at which it fails to clear τ is therefore ruled out without being evaluated. Every prefix the screen admits is then evaluated exactly, in increasing order, and the first crossing is reported. The screen preserves the exact first crossing and introduces no tuning parameter. Over the $20{,}000$ streams of the stopping-cost run it settles $8{,}379$ of the $8{,}440$ that never certify (99.3%) on the screen alone, with no exact evaluation at all. On a stream that does certify it performs a median of four exact evaluations, and locates a crossing at a median prefix of 37, about a tenth of the prefixes below the crossing. Elapsed time is reported in § D.3.

For each of the ten prespecified seeds s, dataset code d (0 for CIFAR-10 and 1 for ImageNet), and image index i, a seed sequence initialized by (s, d, i) generates one permutation of the cached $10{,}000$ scores. The staged plug-in, CS plug-in, E-Cert-P, and E-Cert-R receive that same permuted stream. Every procedure is recomputed from its first permuted score, including the online refit schedule.

For every certificate $j \in \{\mathrm{NP}, \mathrm{VC}, \mathrm{CDF}\}$, let $\mathcal{M}_j$ contain the staged plug-in, the CS plug-in comparator, E-Cert-P, and E-Cert-R. Let $T^k_{s,d,j,i}$ be the exact first-crossing evaluation count of $k \in \mathcal{M}_j$ for certificate j, and ∞ if it does not certify by the sample-size cap $n_{\max} = 10{,}000$. The reported cost is the censored $\min(T^k_{s,d,j,i}, n_{\max})$, so an uncertified image is charged the cap rather than discarded. For each procedure, we compute the median and mean of that censored cost over *every* sampled image $i \in \{1, \ldots, 1{,}000\}$ for that seed, dataset, and certificate – Table 16 reports the arithmetic mean and sample standard deviation of those per-seed statistics across the ten seeds. One alternative is to restrict each median and mean to the subset $\mathcal{I}_{s,d,j} = \{i : \max_{k \in \mathcal{M}_j} T^k_{s,d,j,i} < \infty\}$ of images every procedure certifies by $n_{\max}$. That is a stricter selection than any single procedure's own success event, so it can favor whichever procedure is least often the bottleneck for membership in $\mathcal{I}_{s,d,j}$. Comparing on the same fixed population of $1{,}000$ images removes that asymmetry, at the cost of no longer measuring conditional stopping efficiency on a jointly-easy subset. Figure 6 measures that instead. The standard deviations quantify score-order sensitivity conditional on the cached multisets, not uncertainty across fresh Monte Carlo samples; the latter is measured in § H.9. Under this protocol, which combines the screened per-prefix CDF search above with the adaptive RIPr refit schedule of § B, the full ten-seed run on the real cached scores takes a few hours with 60 worker processes on the 72-core reference machine, the CDF comparator dominating.

Analysis time versus sampling time. We measure sampling and analysis time separately to identify computational bottlenecks, then report combined online time below. Drawing one image's 10^4 noisy scores costs 0.23 s (CIFAR-10) and 1.45 s (ImageNet) on one GH200 in fp32 at the batch sizes of the cache-generating runs, that is 23 and 145 microseconds per model evaluation. Against that, the median single-core cost of analysing the full 10^4-score cache is at most 0.03 s for the staged plug-in and for each of E-Cert-P's three certificates, 0.22/0.05 s for the CDF confidence sequence monitored

at every prefix through the screen above, and 3.5/3.2 s for E-Cert-R on the dense-to-200 schedule, of which essentially all is the CDF projection (NP alone 0.02 s, VC alone 0.03 s, the doubling schedule of § G 0.75/0.71 s). These analysis times are strongly right-skewed, so a mean and a median answer different questions: the mean is what a run over every image costs, the median what a typical image costs. The CDF confidence sequence is the extreme case, since the few streams that first cross near the cap admit thousands of prefixes – its worst image among the 50 timed took 462/13 s, which on its own carries the mean to 9.7/0.7 s – and evaluating every prefix without the screen would be minutes per image throughout. Restricting E-Cert-R's dense schedule to its first 50 refits likewise leaves the median almost unchanged (3.7%/1.3%) while moving the mean substantially (7.6 → 3.8 s and 5.0 → 2.9 s), a truncation rather than a deployable procedure. End-to-end cost is therefore measured separately in the online experiments below. Per-arm means, medians and maxima are in the released results.

Full-cache timings use different batch sizes from online certification and cannot establish a speedup over another early-stopping procedure. The online comparisons in § D.3 and § D.4 include the cheaper direct-witness baseline and measure total elapsed time on soft scores and hard labels.

D.3 Actual online certification time

To measure deployment cost rather than infer it from cached analyses, we run sampling and certification together on one GH200, using the same released checkpoints and preprocessing as the main experiments. We select 200 evenly spaced images per dataset and predefine chunk sizes 64 and 256. Each method receives the same Gaussian-noise prefix within a chunk setting; three timing repeats reuse that noise, so they measure runtime variability, not independent statistical replicates. All methods use the same population futility checks; the E-Cert-R arms use doubling refits and support cap 0.99 as in § F.1. Model loading and warm-up are excluded, whereas monitor initialization, noise generation, forward passes, score transfers and analysis are included. We charge every generated draw, including the remainder of the batch containing the first crossing, and split chunks at futility checkpoints. E-process capital can cross inside a chunk; the CS plug-in is inspected at chunk endpoints, with no additional per-checkpoint error split. Its VC/CDF implementation recomputes prefix bounds, so the comparison concerns these implementations and not all possible confidence-sequence implementations. Chunk sizes can yield different noise streams or check times and are therefore reported separately.

Table 4 reports all five methods, both chunks and all three certificates, with no restriction to certified images. At chunk 256, CDF RIPr takes 0.623 s on CIFAR-10 and 0.505 s on ImageNet, compared with 0.036/0.045 s for the same exact-stake optimizer using a working-law witness, at identical certified accuracy in each dataset. The ratios of mean time (working-law witness divided by RIPr) are 0.058 [0.040, 0.077] and 0.088 [0.054, 0.135], with paired image-bootstrap 95% intervals. The projection dominates RIPr's time. This ordering persists at chunk 64; by contrast, NP has essentially identical time for the two exact-stake arms, and VC is close (0.048 versus 0.050 s and 0.123 versus 0.123 s at chunk 256). Larger chunks reduce overhead but can generate more unused draws. Means first average the three timing repeats for each image and then average over images; bootstrap intervals resample images jointly across methods, never treating the timing repeats as independent images.

When an evaluation saved is a second saved. Write one image's end-to-end cost as $T = n\,t_s + C_a$: n model evaluations at t_s each, plus analysis C_a. The two arms compared here share the stake, the futility rule and the refit schedule and differ only in the witness, so $T_{\mathrm{RIPr}} - T_{\mathrm{plug}} = \Delta C_a - \Delta n\, t_s$ and the projection pays for itself exactly when

$$t_s \;>\; t_s^\star \;:=\; \Delta C_a / \Delta n. \tag{12}$$

This is an accounting identity under three conditions, which these runs meet to within their own noise: both arms pay the same t_s (the two differ by at most 5×10^{-4} ms in every cell), $\Delta n > 0$, and the stopping and analysis behaviour is otherwise held fixed. What it buys is that both sides are measurable in the same run, and that they separate sharply by certificate. Reading ΔC_a, Δn and t_s off the runs behind Table 4: NP is degenerate, since the two arms carry the same witness and agree image by image (§ I.11), leaving $\Delta n = 0$ and a ΔC_a of 4–10 μs that is the harness's own noise. VC has $\Delta C_a = 1.0$–1.4 ms against $\Delta n = 5.3$–9.9 evaluations, so $t_s^\star = 0.14$–0.20 ms, the same order as the $t_s = 0.060$–0.21 ms these runs pay, which is itself set by the batch the sampler fills. (12) therefore predicts a tie on VC, and the clock agrees without resolving the sign: those ratios run

0.997–1.025, inside the jitter floor reported below. CDF has $\Delta C_a = 0.45$–0.59 s against Δn of at most 10.3, putting $t_s^\star$ at 44 ms or more, two to three orders of magnitude over the available t_s; the clock shows 7–17×. One CDF cell has no finite break-even at all: on CIFAR-10 at chunk 256 the two arms generate the same 508.0 draws on average, so $\Delta n = 0$ and the 0.59 s of projection buys nothing that a larger t_s could repay.

The CDF gap is therefore not marginal at this operating point, and closing it would need t_s larger by orders of magnitude rather than by a constant. Two things raise t_s, neither of them measured here: the batch the sampler can fill, since the chunks of 64 and 256 used here already amortise the launch overhead that a latency-bound deployment would pay per draw, and the cost of the base classifier, since denoised smoothing (Salman et al., 2020) charges a denoiser and a classifier per draw and its diffusion variant (Carlini et al., 2023) a denoising pass and a large ViT. We claim no crossing for either; (12) says what would have to hold, and on the base classifier and hardware used here it does not, which is why § 7.1 recommends the direct witness and treats RIPr as a sampling-design option. We also cannot resolve how $t_s^\star$ moves with instance difficulty from these runs: Δn exceeds 500 on exactly one of the 200 ImageNet images, and C_a on the images that draw past 10^3 is measured on two to seven images per band, too few to separate a per-refit from a per-draw cost.

The strongest hard-label comparator, Voráček's Krichevsky–Trofimov sequence, cannot appear in this table: it is a coin-betting estimator for a $0/1$ stream, while the observable timed here is the target's continuous softmax score. § D.4 repeats the measurement on the hard label, where it does apply.

Table 4: Actual online cost over all 200 images per dataset, population futility enabled. CA is certified accuracy (%); draws and seconds are means, including batch overshoot and all uncertified images. CPU is the mean analysis time at chunk 256; totals also include sampling, transfer and monitor initialization. Three repeats reuse the same noise.

			Chunk 64			Chunk 256			CPU (256)
Dataset	Cert.	Method	CA	Draws	Seconds	CA	Draws	Seconds	Seconds
CIFAR-10	NP	Working-law witness / quadratic stake	55.0	393.9	0.082	55.0	424.4	0.032	0.001
		Working-law witness / exact stake	55.0	374.6	0.079	55.0	423.1	0.032	0.001
		RIPr witness / exact stake	55.0	374.6	0.079	55.0	423.1	0.032	0.001
		E-Cert-P	55.0	370.3	0.077	55.0	421.9	0.031	0.000
		CS plug-in	55.0	383.6	0.081	54.5	443.4	0.032	0.000
	VC	Working-law witness / quadratic stake	57.5	754.0	0.152	57.0	762.3	0.048	0.003
		Working-law witness / exact stake	57.5	751.2	0.152	57.0	758.8	0.048	0.003
		RIPr witness / exact stake	57.5	741.4	0.151	57.0	751.1	0.050	0.004
		E-Cert-P	57.0	750.6	0.153	57.0	752.4	0.048	0.002
		CS plug-in	57.0	775.3	0.470	57.0	778.6	0.124	0.077
	CDF	Working-law witness / quadratic stake	57.5	465.2	0.096	57.0	513.1	0.036	0.004
		Working-law witness / exact stake	57.5	462.7	0.096	57.0	508.0	0.036	0.004
		RIPr witness / exact stake	57.5	457.9	0.687	57.0	508.0	0.623	0.591
		E-Cert-P	56.5	505.3	0.111	56.5	549.3	0.043	0.008
		CS plug-in	56.5	534.8	1.464	56.5	581.1	0.433	0.397
ImageNet	NP	Working-law witness / quadratic stake	43.0	335.3	0.060	43.0	346.5	0.054	0.001
		Working-law witness / exact stake	43.0	335.9	0.061	43.0	346.5	0.054	0.001
		RIPr witness / exact stake	43.0	335.9	0.061	43.0	346.5	0.054	0.001
		E-Cert-P	43.0	335.2	0.060	43.0	346.5	0.053	0.000
		CS plug-in	43.0	345.6	0.062	43.0	353.1	0.054	0.000
	VC	Working-law witness / quadratic stake	48.5	766.9	0.136	48.5	805.4	0.122	0.003
		Working-law witness / exact stake	48.5	768.6	0.137	48.5	806.1	0.123	0.003
		RIPr witness / exact stake	48.5	761.4	0.137	48.5	800.9	0.123	0.004
		E-Cert-P	48.5	772.0	0.139	48.5	810.7	0.126	0.003
		CS plug-in	48.5	834.5	0.464	48.5	866.2	0.213	0.085
	CDF	Working-law witness / quadratic stake	48.5	258.8	0.047	48.5	274.7	0.044	0.002
		Working-law witness / exact stake	48.5	265.9	0.049	48.5	280.4	0.045	0.002
		RIPr witness / exact stake	48.5	255.6	0.498	48.5	270.7	0.505	0.464
		E-Cert-P	48.0	351.4	0.071	48.0	369.7	0.064	0.008
		CS plug-in	48.0	377.8	0.856	48.0	384.5	0.257	0.199

D.4 Actual online certification time on the hard label

The protocol above transfers unchanged except for the observable. Every draw now contributes the indicator $\mathbf{1}\{\hat{f}(x+\varepsilon) = y\}$ rather than the target's softmax score, so the certificate is NP at $\rho = 1/2$

and both hard-label comparators apply. We time four arms: the staged plug-in (Cohen's rule made staged, a one-sided Clopper–Pearson bound at level $\alpha/10$ at ten predefined checkpoints), the CS plug-in (KT) inspected at every draw, and the two exact-stake e-processes. Both comparator monitors reproduce their published counterparts stream for stream before any timing is recorded. Table 5 reports the result.

Each arm is implemented in the form its own decision needs, so that the timing charges it for the arithmetic that decision requires and no more. For KT the certification event is one-sided at the fixed threshold θ: since $\log W_t(p)$ is convex with its minimum at the running mean, every $p < \theta$ is rejected exactly when $H_t/t > \theta$ and the wealth at θ clears $\log(1/\alpha)$, and the KT wealth depends on the stream only through (H_t, t). The arm therefore carries the running counts and tests the threshold in $O(1)$ per draw, which the self-checks confirm reproduces the published confidence sequence's decision prefix by prefix and stream by stream.

Certified accuracy separates the four arms by at most 0.5 percentage points (61.0–61.5% on CIFAR-10, 59.5–60.0% on ImageNet). On the images an arm certifies, its median crossing is 22 draws for both exact-stake e-processes against 25 for both comparators, in all four dataset–chunk cells. These 200 images and the cached benchmark behind Figure 7 condition on different sets, so the two sets of medians are not interchangeable.

That margin does not convert into wall-clock: the ratios of mean time, comparator divided by ours, run 0.940–1.001 and our arm is between 0 and 6% slower in every cell. Two things absorb it. A three-draw margin at a median crossing of 22–25 is invisible beside a chunk of 64 or 256 draws and beside the futility rule all four arms share, and mean generated draws are 329–525 for every arm. Per-draw arithmetic then decides what is left, and on a $0/1$ stream every arm here is computing a one-dimensional statistic, so there is little to separate them: analysis costs 0.0003–0.0018 s per image across the four.

On a hard label, VC and CDF are determined by the mean (§ I.1), and all arms test the same scalar probability, so (12) degenerates rather than binds: every arm computes a one-dimensional statistic, which puts ΔC_a at the jitter floor, and the three-draw margin is worth 0.2–0.5 ms of sampling against a chunk that costs 4–39 ms. The reduction in evaluations therefore does not yield a runtime advantage in this experiment, and could not at this granularity. The softmax experiment in § D.3 assesses the additional cost of exploiting the score distribution.

The two exact-stake NP arms have the same witness and agree image by image in crossing time and outcome, as required by § I.11. The exporter checks this equality. Their time ratios for identical work, 1.000 [0.989, 1.013] and 0.965 [0.880, 1.000], measure this experiment's jitter floor; the staged-rule ratios should be read against it.

Table 5: Actual online cost on the hard label, all 200 images per dataset, NP certificate, population futility enabled. CA is certified accuracy (%); draws and seconds are means, including batch overshoot and all uncertified images. CPU is the mean analysis time at chunk 256. The last column is the baseline's mean time divided by ours at chunk 256, with paired image-bootstrap 95% intervals; above one favours ours. Three repeats reuse the same noise.

		Chunk 64			Chunk 256			CPU (256)	Time ratio (256)
Dataset	Method	CA	Draws	Seconds	CA	Draws	Seconds	Seconds	baseline/ours [95% CI]
CIFAR-10	Staged plug-in	61.5	445.3	0.0935	61.5	524.8	0.0364	0.0006	0.959 [0.940, 0.982]
	CS plug-in (KT)	61.5	455.6	0.0953	61.0	517.6	0.0360	0.0003	0.948 [0.933, 0.963]
	Working-law witness / exact stake	61.0	464.3	0.0983	61.0	524.7	0.0380	0.0018	1.001 [0.990, 1.014]
	RIPr witness / exact stake	61.0	464.3	0.0985	61.0	524.7	0.0380	0.0018	—
ImageNet	Staged plug-in	59.5	334.3	0.0600	60.0	349.8	0.0544	0.0005	0.940 [0.827, 1.005]
	CS plug-in (KT)	59.5	329.4	0.0588	59.5	356.1	0.0547	0.0003	0.946 [0.854, 1.004]
	Working-law witness / exact stake	59.5	329.2	0.0595	59.5	360.1	0.0558	0.0013	0.966 [0.882, 1.001]
	RIPr witness / exact stake	59.5	329.2	0.0600	59.5	360.1	0.0578	0.0013	—

E CAPITAL TRAJECTORIES AND CERTIFICATION TIMES

Figure 4 plots the accumulated capital $\mathcal{K}_t = \prod_{s \le t} e_s$ against the threshold $1/\alpha$ it must cross to certify.

E-Cert-P and E-Cert-R use the same image, certificate and score order. E-Cert-P fixes its mixture before sampling; E-Cert-R adapts its design to completed observations (§ 5). Ten curves per arm give the ten prespecified permutations behind Table 16, with seed 2026 solid and the other nine faded. Those permutations reorder one cached score multiset, so their spread is sensitivity to score *order* conditional on the draws, not fresh-draw Monte Carlo variability, which is measured separately in § H.9.

The image is selected by a prespecified rule. Among the 563 of 1,000 CIFAR-10 images that both arms certify on all three certificates under all ten seeds, we keep the upper quartile by across-seed spread. This excludes nearly constant streams whose curves obscure score-order variation. Of those we take the one closest to the eligible-set median stopping time, which is unique. Both criteria are symmetric in the two arms. The selected image is also checked for representativeness: its fixed-to-RIPr median ratios are $1.15/1.39/1.39$ for NP/VC/CDF against $1.17/1.38/1.43$ for the whole eligible set, a drift of at most 0.046, and the median-spread threshold would have selected a different image with ratios equal to the population's.

Certificate-specific effects of adaptive design. The three certificates exhibit different responses to adaptation. On NP the two designs are nearly on top of each other (31.5 against 27.5 evaluations at the median), because the NP witness is identical between the arms and RIPr can only retune the stake. On VC (41 against 29.5) and CDF (37.5 against 27) the gap is wider, because there RIPr chooses the witness, the boundary point or the tangent, and not only the stake. These are medians on one image and are smaller than Table 16's, which average over all 1,000 images with an uncertified image charged the 10,000 cap.

The curves continue past their first crossing, where the deployed procedure stops; the generating script asserts that every plotted path's first crossing equals the stopping time the deployed code reports, against both the published per-image records and a fresh call to the deployed routine.

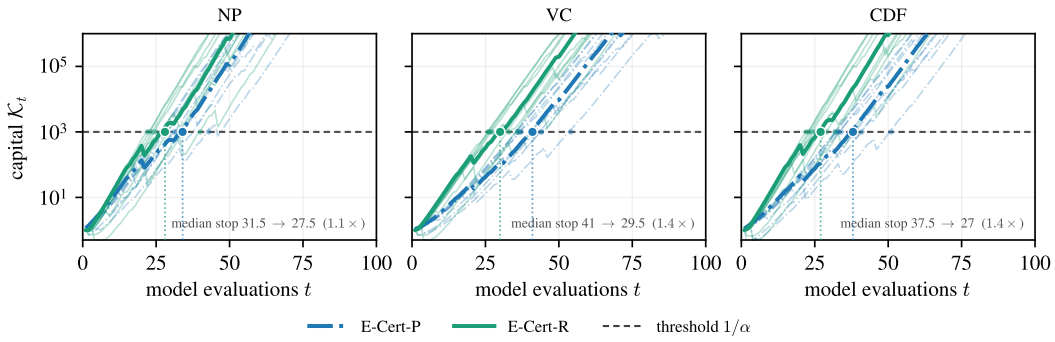


Figure 4: Capital $\mathcal{K}_t$ against model evaluations on one prespecified CIFAR-10 image, at $\rho = 0.5$ and $\alpha = 10^{-3}$, for E-Cert-P and E-Cert-R on each certificate. Ten curves per arm are the ten permutation seeds behind Table 16; the solid curve is seed 2026, the drop-lines mark first crossings and the small dots the other seeds'. Curves continue past their crossing to show the trajectory; the deployed procedure stops there.

F MONITORING VERSUS WITNESS SHARPNESS

§ 7.1 reports that at $n = 100$ both e-process designs achieve higher certified accuracy than the prespecified fixed-n plug-in on every dataset–certificate pair. This comparison combines changes in witness construction and monitoring; this appendix separates them on the single-certificate main experiment's default refit schedule (every round through 200, then 100 geometrically spaced rounds

Table 6: Decomposing the small-budget accuracy advantage into witness sharpness and monitoring, at $n = 100$. Certified accuracy (%) for the three arms, and the two differences that sum to the total gain. The witness column is "RIPr-guided fixed-n minus Maurer–Bernstein fixed-n", the monitoring column "RIPr-guided anytime minus RIPr-guided fixed-n"; both arms in the second difference share the same witness and the same design.

		Certified accuracy			Decomposition	
Dataset	Cert.	Fixed-n plug-in	RIPr terminal	RIPr anytime	witness	monitoring
CIFAR-10	NP	36.7	50.7	51.1	+14.0	+0.4
	VC	28.4	53.7	54.2	+25.3	+0.5
	CDF	44.2	53.7	54.4	+9.5	+0.7
ImageNet	NP	36.2	46.8	46.9	+10.6	+0.1
	VC	29.4	51.3	51.4	+21.9	+0.1
	CDF	43.2	51.1	51.3	+7.9	+0.2

up to 10,000). To separate their effects, note that the fixed-n NP bound is the one-sided Maurer–Pontil empirical-Bernstein bound, whose slack term $7\log(2/\delta)/(3(n-1))$ alone is 0.179 at $n = 100$, $\delta = 10^{-3}$. This correction reduces the lower mean bound independently of the monitoring schedule.

We introduce a third procedure to separate these effects:

Arm	Witness	Decision rule
Maurer–Bernstein fixed-n	loose	terminal only
RIPr-guided fixed-n	sharp	terminal only
RIPr-guided anytime	sharp	$\sup_{t \le n}$

The middle arm uses our own predictable RIPr design and factors but may reject only on the terminal capital $\mathcal{K}_n$, never on $\sup_{t \le n} \mathcal{K}_t$. It is therefore a legitimate fixed-n test of the same null with the same witness, and the two differences decompose the reported gap into witness sharpness (middle minus top) and monitoring (bottom minus middle). Both decisions are read off one capital path per stream using the main experiment's default refit schedule. The replay is checked against the deployed per-image loop on real streams, and every first crossing is compared with the saved image-level results underlying Table 1 before the table is exported.

Table 6 gives the decomposition at $n = 100$. The terminal-only direct test retains 93.1–99.5% of the advantage; continuous monitoring adds at most 0.7 pp. At $n = 1{,}000$ and 10,000 the two e-process arms agree to within 0.1 pp in every cell, monitoring adding $+0.0$ or $+0.1$, while the witness still accounts for 0.1–5.0 pp over the fixed-n plug-in. The table therefore reports the budget at which the components differ most. This pattern is consistent with certification using a single $\log(1/\alpha)$ log-capital threshold rather than a union bound over monitoring times. Thus witness construction accounts for most of the measured accuracy improvement, while anytime validity justifies the adaptive stopping used in Table 16.

Against a sharper fixed-n bound. The decomposition above measures sharpness relative to the empirical-Bernstein bound the fixed-n plug-in uses, whose slack at $n = 100$ is large. The strongest fixed-n comparator we know is the betting bound of Waudby-Smith & Ramdas (2024) evaluated at the terminal time only, with its predictable empirical-Bernstein stake tuned for n (so it pays no $\log t$ for time-uniformity); for VC its variance term is the upper confidence sequence of Martinez-Taboada & Ramdas (2026) read at n. Table 7 puts it beside the arms of Table 1. At $n = 100$ it recovers most of the gap. It is within about 2 pp of the e-processes for NP on both datasets, and for VC within 8–10 pp (44.0 against 52.0/54.2 on CIFAR-10, 42.6 against 50.3/51.4 on ImageNet). It tracks the CS plug-in to within 0.6 pp at every budget, as it should: the two are the same bet with the stake tuned for n or for all t. The CS plug-in in Table 1 therefore approximates this sharper fixed-n comparator within the stated tolerance. The 6–26 point gain over empirical Bernstein substantially exceeds the gain over this comparator. The residual is consistent with the benefit of testing a certificate directly: NP uses the same centred-score witness as the mean bound, while VC's single boundary-normal witness replaces two separately bounded components. However, these arms also differ in stake design; the

following control isolates that choice. At $n = 1{,}000$ the empirical-Bernstein VC bound still trails the best arm by 5.0 pp on CIFAR-10 and 3.6 pp on ImageNet; at $n = 10^4$ the largest remaining gap is 1.3 pp.

Table 7: Certified accuracy (%) at three budgets for the fixed-n plug-in of Table 1 (empirical Bernstein), a fixed-n betting bound tuned for n, and the anytime-valid arms, on the same streams; NP and VC only (the CDF plug-in's DKW band is unchanged).

Dataset	Cert.	Method	$n = 10^2$	$n = 10^3$	$n = 10^4$
CIFAR-10	NP	Fixed-n plug-in (empirical Bernstein)	36.7	53.8	56.3
		Fixed-n plug-in (betting, tuned for n)	49.1	55.1	56.5
		CS plug-in	49.2	54.9	55.9
		E-Cert-P	50.7	54.9	56.5
		E-Cert-R	51.1	54.9	56.4
	VC	Fixed-n plug-in (empirical Bernstein)	28.4	54.6	59.7
		Fixed-n plug-in (betting, tuned for n)	44.0	59.2	61.0
		CS plug-in	44.0	58.9	60.6
		E-Cert-P	52.0	59.6	61.0
		E-Cert-R	54.2	59.6	61.0
ImageNet	NP	Fixed-n plug-in (empirical Bernstein)	36.2	48.8	50.0
		Fixed-n plug-in (betting, tuned for n)	44.8	49.5	50.3
		CS plug-in	45.1	49.5	50.1
		E-Cert-P	46.5	49.5	50.2
		E-Cert-R	46.9	49.5	50.2
	VC	Fixed-n plug-in (empirical Bernstein)	29.4	51.6	54.9
		Fixed-n plug-in (betting, tuned for n)	42.6	54.0	56.0
		CS plug-in	42.7	53.9	55.9
		E-Cert-P	50.3	55.2	56.1
		E-Cert-R	51.4	55.2	56.1

F.1 Separating witness choice from stake optimization

The comparisons above change several design choices simultaneously. We therefore add a single-class control on the same 1,000 images per dataset, at $\sigma = 0.25$, $\rho = 0.5$, $\tau = 0.5$ and $\alpha = 10^{-3}$, with a 10,000-draw cap. All three arms share the same past-only 100-bin histogram (pseudocount 0.01 per bin), its bin-center moments, raw-score capital updates, and refit times. The first uses a *working-law witness* – a direct certificate witness read off the past-only working law, with no RIPr projection – and a quadratic-Kelly stake; the second the same witness with an exact expected-log stake; the third the RIPr-guided witness with the same exact stake optimizer. These are direct-witness controls, not confidence-bound plug-in monitors: we use them to isolate the incremental benefit of RIPr-guided design. The distinction is certificate-specific. The NP null is a half-space whose witness is fixed, so all three arms share it and the two exact-stake arms are identical by construction, which is verified on every image; only VC and CDF have a witness to choose. For VC and CDF, comparing them isolates witness selection from stake optimization within the shared working model.

We predefine both the dense-to-200 and doubling refit schedules and both support-cap fractions 0.5 and 0.99. The cap bounds the stake by that fraction of $-1/u_{\min}$ and is imposed inside the VC design search, rather than clipping a witness optimized at another cap. We report all four settings in Table 9; doubling and 0.99 are the predefined setting shared with the online timing experiment, not a choice of the best test-set result. Every arm is also evaluated with the same population futility rule at its six checkpoints, using $\beta_f/6$ per checkpoint with $\beta_f = 10^{-3}$. This budget is separate from certification: e-process certification still uses the capital threshold $1/\alpha$, while the staged comparator elsewhere retains $\alpha/10$ at each certification checkpoint. An image not certified or abandoned earlier pays the full cap.

Table 8 shows that the added RIPr benefit over a cheap adaptive witness is small in all-images mean cost. Relative to the exact-stake working-law witness, with futility, the VC reductions are 1.4% and 0.3% on CIFAR-10 and ImageNet, and the CDF reductions 2.2% and 2.9%; NP is identical. Without

futility all four reductions are below 0.3%, because uncertified images dominate the mean. Terminal accuracy differs by at most one image. At $n = 100$, the three designs separate more clearly in certified accuracy. For VC the working-law witness certifies $53.2/50.6\%$ (CIFAR-10/ImageNet) with the quadratic-Kelly stake and $53.2/50.5\%$ with the exact one, and RIPr raises it to $53.7/51.4\%$; for CDF the three are $52.7/50.1$, $52.4/50.0$ and $53.9/51.3\%$. Against the $28.4/29.4\%$ that the componentwise empirical-Bernstein VC bound reaches at the same budget, and the $44.0/42.6\%$ of the tuned fixed-n betting bound above, the direct witness adds 8–9 points over the tuned betting bound, followed by 0.5–0.9 points from RIPr. Exact stake optimization provides no consistent finite-sample improvement over the quadratic rule in these settings.

The same control at a large radius. Those numbers are all at $\rho = 1/2$, where the certificate is easy and every arm certifies most images; the cells where the paper's strongest result lives are at $\rho \geq 2$. We therefore rerun the control there, on the cached $\sigma = 0.25$ streams in their stored order, so that the CS plug-in of § G — which is run on those same streams at the same α — serves as a fourth column testing the same discretized certificate. All four keep the uniform $K = 100$ grid: the threshold-aware rule of § B.2 is an NP device, and warping the grid for CDF would change the functional itself rather than the design. The three design arms share the doubling refits and the support cap 0.99 of the table above. Certified accuracy is read with futility off; the rule abandons no certificate any of these arms reaches, so it moves no entry here.

		CS plug-in	cheap, quadratic	cheap, exact	RIPr, exact
$\rho = 2$	CIFAR-10	34.2	34.5	34.5	36.2
	ImageNet	41.6	41.5	41.5	43.2
$\rho = 3$	CIFAR-10	0.0	0.0	0.0	10.8
	ImageNet	0.0	0.0	0.0	18.8

Certified accuracy (%) at $n = 10^4$ on the CDF certificate. At $\rho = 2$ the picture matches $\rho = 1/2$: the cheap adaptive witness is level with the CS plug-in and RIPr adds 1.7 pp (paired bootstrap $[0.9, 2.5]$ pp on both datasets). At $\rho = 3$ it does not. There neither cheap witness certifies anything under either stake rule, while RIPr certifies 10.8 and 18.8%. Against these three comparators, at this radius and budget, the projection is not a refinement worth a point at the margin; it is the difference between an arm that certifies and one that does not. That is a statement about the three arms tested, not about every CDF construction without a projection. On VC the control is flat throughout — 27.0 and 32.3 for all three designs at $\rho = 2$, zero for all of them at $\rho = 3$ — so it is also specific to the CDF witness.

What separates the RIPr column from § G is the support cap, not the schedule: both runs use doubling refits, and rerunning these arms at the cap that appendix uses ($1 - 10^{-8}$ rather than 0.99) leaves every RIPr entry unchanged except ImageNet at $\rho = 3$, which reads 19.0 and reproduces it exactly. Two cells move in all: that one, and the cheap quadratic witness on CIFAR-10 at $\rho = 2$, by 0.1 pp. We keep 0.99 here to match the cap and refit schedule of the table above; the draw order differs from it, as noted.

At $\rho = 3$ the RIPr solver fails its success-or-boundary check at least once on 194 of $1{,}000$ CIFAR-10 images and 247 of $1{,}000$ on ImageNet (266 and 293 refits in total). Each failure uses the neutral factor 1 until the next valid refit, which preserves validity; all affected runs remain included in the reported results. Skipping a bet is not automatically a loss of power, since a round the design would have staked on can also move capital down. These are failures of the projection solver, not of the stake optimizer.

Independent-noise replication and numerical failures. We repeat the matched comparison using three independently generated noise caches (seeds 2027, 2028, 2029), keeping the same 200 images per dataset, checkpoint and labels. Table 10 reports mean and sample standard deviation over the three seeds at doubling/0.99; the saved results also include cap 0.5. NP's exact-stake arms again coincide, and CDF retains a small mean-cost improvement. VC is not uniformly improved: on ImageNet its seed-averaged cost increases from 768.9 to 770.9. All main and noise-replicate settings have zero observed certificate loss from population futility; this observation does not imply zero population abandonment probability.

Table 8: Single-class matched designs, doubling refits and support cap 0.99. CA is certified accuracy (%) at $10{,}000$; both mean costs charge all $1{,}000$ images per dataset. Lost counts certificates present without futility but absent with it. All three arms use the same working law.

Dataset	Cert.	Method	CA (%)	Mean	Mean + futility	Lost
CIFAR-10	NP	Working-law witness / quadratic stake	56.3	4465.5	561.4	0
		Working-law witness / exact stake	56.6	4455.4	551.3	0
		RIPr witness / exact stake	56.6	4455.4	551.3	0
	VC	Working-law witness / quadratic stake	61.0	3997.1	827.3	0
		Working-law witness / exact stake	61.0	3990.0	820.2	0
		RIPr witness / exact stake	61.1	3978.5	808.6	0
	CDF	Working-law witness / quadratic stake	61.2	3983.3	523.3	0
		Working-law witness / exact stake	61.2	3980.2	520.3	0
		RIPr witness / exact stake	61.2	3968.7	508.7	0
ImageNet	NP	Working-law witness / quadratic stake	50.2	5030.1	378.7	0
		Working-law witness / exact stake	50.2	5030.4	379.0	0
		RIPr witness / exact stake	50.2	5030.4	379.0	0
	VC	Working-law witness / quadratic stake	56.0	4464.3	640.6	0
		Working-law witness / exact stake	56.0	4459.4	635.7	0
		RIPr witness / exact stake	56.0	4457.7	634.0	0
	CDF	Working-law witness / quadratic stake	55.9	4464.5	312.6	0
		Working-law witness / exact stake	55.9	4465.5	313.5	0
		RIPr witness / exact stake	55.9	4456.4	304.4	0

Table 9: All predefined matched-control settings. Ratio of all-images mean evaluations: exact-stake working-law witness divided by RIPr, both with futility; larger than one favors RIPr. Brackets are paired image-bootstrap 95% intervals. NP has ratio 1.000 in every setting. These are pointwise descriptive intervals, conditional on the cached streams.

Dataset	Schedule	Cap	VC ratio [95% CI]	CDF ratio [95% CI]
CIFAR-10	dense	0.5	1.014 [1.007, 1.025]	1.040 [1.032, 1.052]
		0.99	1.017 [1.008, 1.033]	1.022 [1.017, 1.028]
	doubling	0.5	1.011 [1.007, 1.017]	1.037 [1.029, 1.047]
		0.99	1.014 [1.008, 1.023]	1.023 [1.017, 1.031]
ImageNet	dense	0.5	1.007 [1.005, 1.010]	1.074 [1.057, 1.103]
		0.99	1.008 [1.005, 1.011]	1.033 [1.025, 1.047]
	doubling	0.5	1.007 [0.998, 1.019]	1.064 [1.049, 1.083]
		0.99	1.003 [0.997, 1.008]	1.030 [1.021, 1.042]

Across the main experiment's $8{,}000$ CDF RIPr image–schedule–cap runs, 420 of $667{,}225$ refits fail the numerical solver's success/boundary check, affecting 167 runs. Counts use each base path once, rather than duplicating it for the with/without-futility summaries. Every failure uses the neutral factor 1 until a subsequent valid refit; failed runs remain in the tables. This preserves the null-side guarantee but can reduce power and increase cost. In the noise repetitions there is one failed refit (seed 2029), again with neutral fallback. These failures are retained in the reported costs and accuracies.

G Certified accuracy across radii and smoothing levels

Changing the radius at fixed σ leaves the sampled score distribution untouched and recomputes only the ρ-dependent tests. Varying σ is a different experiment: each level has its own Gaussian-augmented base classifier and therefore its own score streams, so it needs its own cache. We sampled $1{,}000$ images $\times$ $10{,}000$ draws per dataset at $\sigma \in \{0.25, 0.5, 1.0\}$, using the released checkpoint trained at each level, and evaluated an 8×3 grid of radii $r \in \{0.125, 0.25, 0.375, 0.5, 0.75, 1, 1.5, 2\}$ and smoothing levels. This gives the certified-accuracy-versus-radius curves that are conventional in

Table 10: Independent-noise matched controls: mean ± sample standard deviation over three seeds, 200 fixed images per dataset, doubling refits, cap 0.99, population futility enabled. Repeats vary Gaussian noise, not training seeds or architectures.

Dataset	Cert.	Method	CA (%)	Mean evaluations + futility
CIFAR-10	NP	Working-law witness / quadratic stake	54.67 ± 0.29	358.8 ± 27.2
		Working-law witness / exact stake	54.67 ± 0.29	358.3 ± 26.9
		RIPr witness / exact stake	54.67 ± 0.29	358.3 ± 26.9
	VC	Working-law witness / quadratic stake	56.83 ± 0.29	786.4 ± 34.8
		Working-law witness / exact stake	56.83 ± 0.29	783.2 ± 38.3
		RIPr witness / exact stake	56.83 ± 0.29	779.5 ± 39.6
	CDF	Working-law witness / quadratic stake	56.83 ± 0.29	479.3 ± 47.9
		Working-law witness / exact stake	57.00 ± 0.00	476.1 ± 46.1
		RIPr witness / exact stake	57.00 ± 0.00	467.7 ± 45.3
ImageNet	NP	Working-law witness / quadratic stake	43.00 ± 0.00	337.5 ± 24.0
		Working-law witness / exact stake	43.00 ± 0.00	337.9 ± 24.0
		RIPr witness / exact stake	43.00 ± 0.00	337.9 ± 24.0
	VC	Working-law witness / quadratic stake	48.50 ± 0.00	770.5 ± 11.1
		Working-law witness / exact stake	48.50 ± 0.00	768.9 ± 11.7
		RIPr witness / exact stake	48.50 ± 0.00	770.9 ± 10.0
	CDF	Working-law witness / quadratic stake	48.50 ± 0.00	256.3 ± 6.5
		Working-law witness / exact stake	48.50 ± 0.00	256.6 ± 6.3
		RIPr witness / exact stake	48.50 ± 0.00	247.9 ± 5.6

this literature, for all four arms of Table 1 on identical streams: the fixed-n plug-in, the CS plug-in, E-Cert-P, and E-Cert-R (Figure 5; Table 11–Table 13).

To reduce computation across the 44 dataset–σ–radius cells, E-Cert-R uses doubling refits instead of the main tables' dense-to-round-200 schedule (§ D.2). Both schedules are predictable and valid. We check their agreement at the shared operating point $(\sigma, r) = (0.25, 0.125)$. There the sweep reproduces Table 1's fixed-n, CS plug-in and E-Cert-P accuracies *exactly* in all eighteen dataset–certificate cells, and its RIPr accuracies to within one image (0.565 against 0.564 for CIFAR-10 NP, 0.501 against 0.502 for ImageNet NP, exact elsewhere). The generating script asserts this. The schedule change therefore alters certified accuracy by at most 0.1 percentage points at that setting. Every arm is a function of $\rho = r/\sigma$: E-Cert-P's tangent bank is rebuilt at each ρ, its NP stake grid uses $T_\rho(\tau)$ and its VC boundary grid κ_ρ, and the plug-in bounds take ρ directly. The CS plug-in provides the strongest soft-score plug-in comparison, while E-Cert-P controls for anytime validity without RIPr guidance.

The grid is also restricted by numerical precision. Cells with $\rho = r/\sigma > 4$ are left unevaluated, and the binding constraint is the VC certificate. Its feasible working mean is bounded above by the root $m_{\max}$ of $((m-\tau)/\kappa)^2 = m(1-m)$ with $\kappa = \sqrt{e^{\rho^2} - 1}$, and since $1 - m_{\max} \approx \frac{1}{4}/(e^{\rho^2} - 1)$ that root collapses toward 1 super-exponentially: $1 - m_{\max}$ is 4.6×10^{-3} at $\rho = 2$, 2.8×10^{-8} at $\rho = 4$, 3.5×10^{-12} at $\rho = 5$, and below double precision by $\rho = 6$. The root does not cease to exist, since $f(1) > 0$ always, and $m_{\max}$ is an upper bound rather than a target: any $m \in (\tau, m_{\max}]$ is feasible, and a constant score certifies at every ρ because its variance is zero. What collapses is the variance the working law may carry, which is at most $\{(m-\tau)/\kappa\}^2$ and so falls like $e^{-\rho^2}$; at $\rho = 4$ it is 7×10^{-9} at $m = 0.75$ and 2.8×10^{-8} at the largest feasible m, so the whole admissible range sits below 10^{-7}. Estimating a variance that small from 10^4 draws, and bracketing the root near 1 in double precision, is what fails – our solver's bracketing procedure gives up at $\rho > 4.40$. We therefore report $\rho \leq 4$ and flag this as a real limitation of the VC form at extreme radius-to-σ ratios under the current numerical implementation.

The preferred smoothing level varies with the target radius. In particular, $\sigma = 1.0$ attains the highest certified accuracy at some radii above $r = 1$. For the CDF certificate the sequence of preferred smoothing levels is identical on both datasets: $\sigma = 0.25$ optimal for $r \leq 0.5$, $\sigma = 0.5$ for $r \in \{0.75, 1.0\}$, and $\sigma = 1.0$ for $r \geq 1.5$. NP and VC, whose accuracy falls faster with ρ, favor larger

smoothing levels at smaller radii; $\sigma = 0.5$ is already best at $r = 0.5$ for VC on both datasets, and from $r = 0.375$ for NP on ImageNet. Each level runs out at a predictable ceiling: $\sigma = 0.25$ reaches zero by $r = 1$ and $\sigma = 0.5$ by $r = 2$, where $\sigma = 1.0$ still certifies 3–13%. The best smoothing level thus depends on the target radius.

At the main operating point, the four arms have similar certified accuracy, with NP differences of about a tenth of a point. Differences increase at larger radii. The sweep has 144 cells in all; the tables above report the 132 with $r \leq 2$, and the counts that follow use the same 132. Over them E-Cert-R is ahead of the fixed-n plug-in in 101, level in 30, and behind in exactly one (CIFAR-10 NP at $\sigma = 0.5$, $r = 0.25$, by a single image: 0.368 against 0.369). Against the *CS plug-in*, the comparison that isolates what RIPr guidance adds beyond anytime validity, it is ahead in 85, level in 43 and behind in 4, each of the four by a single image. Restoring $r = 3$ and $r = 4$ changes none of this: over all 144 cells the two tallies become $103/40/1$ and $87/53/4$, since ten of the twelve added cells are ties at zero and two are further leads (ImageNet at $\sigma = 1$, $r = 3$: CDF 3.3% and NP 1.5%, each against 0%). The CS plug-in itself leads the fixed-n one in 83 of the 132 cells by up to 19.7 points. A large part of the fixed-n gap is therefore the tightness of the betting bound relative to the empirical-Bernstein and DKW fixed-n bounds, which the e-processes share with it. § F attributes the small-budget gap to the witness rather than to monitoring. These cell counts reuse the same images across σ, radius and certificate, with one cache per σ, so they measure consistency across conditions rather than independent confirmations; replication across fresh noise draws is § H.9. The margin grows with radius: over the fixed-n plug-in E-Cert-R gains up to 21.3 points, over the CS plug-in up to 19.0. The largest differences occur near the sampling budget's certification limit. At $\sigma = 0.25$, $r = 0.5$ the CDF e-process certifies 43.2% of ImageNet images against 21.9% for the fixed-n plug-in and 41.6% for the anytime-valid one. There are *nine* cells, at every smoothing level, in which the fixed-n plug-in certifies *nothing at all* while the e-process certifies a positive fraction. Five are CDF: 10.8% (CIFAR-10) and 19.0% (ImageNet) at $\sigma = 0.25$, $r = 0.75$; 1.7% and 10.8% at $\sigma = 0.5$, $r = 1.5$; and 3.2% at $\sigma = 1$, $r = 2$ on CIFAR-10. Four are NP at $\rho = 3$: 5.9% and 5.5% at $\sigma = 0.25$, $r = 0.75$, and 0.6% and 3.1% at $\sigma = 0.5$, $r = 1.5$. In eight of the nine the CS plug-in *also* certifies nothing, all but CIFAR-10 at $\sigma = 1$, $r = 2$, where it reaches 2.5%. At the margin the time-uniform DKW band is as far below τ as the fixed-n one, so the difference is attributable to witness construction rather than monitoring frequency. Positive certification by E-Cert-R when the comparators reach zero occurs at every smoothing level. A similar pattern appears at smaller budgets. At $n = 1{,}000$ on the $\sigma = 0.25$ streams of Table 1, the fixed-n CDF bound and E-Cert-P's CDF arm certify nothing at $r = 0.5$, while E-Cert-R's CDF arm certifies 30.8%/35.4% of CIFAR-10/ImageNet images and the CS plug-in 22.6%/27.5%. (E-Cert-P's CDF bank has 28, 27 and 20 admissible templates at $\rho = 0.5, 1, 2$.) No VC method certifies there at that budget, and every NP method does.

Comparison with E-Cert-P. In 16 of the 132 cells E-Cert-P is ahead of E-Cert-R, by one to three images out of 1,000 and never by more than 0.3 points (Table 11); ten are NP and six VC. At large ρ, design-grid resolution can cause larger differences (§ B.2). A working law represents each of its $L = 100$ cells by the cell center, so a uniform grid caps its mean at $1 - 1/(2L) = 0.995$ while the NP threshold $\Phi(\rho)$ passes that value at $\rho = 2.58$; beyond it the NP design would be neutral by construction, and E-Cert-P, which bets from the first round, would not. The rule resolves the threshold instead, and at $\rho = 3$ the NP arm certifies 5.9% of CIFAR-10 and 5.5% of ImageNet images ($\sigma = 0.25$, $r = 0.75$), and 0.6% and 3.1% at $\sigma = 0.5$, $r = 1.5$, against 5.6/4.5% and 0.4/2.8% for E-Cert-P. The grid enters no guarantee (§ 5); it is a design device, and the rule fixes its placement from ρ and τ before any score is drawn.

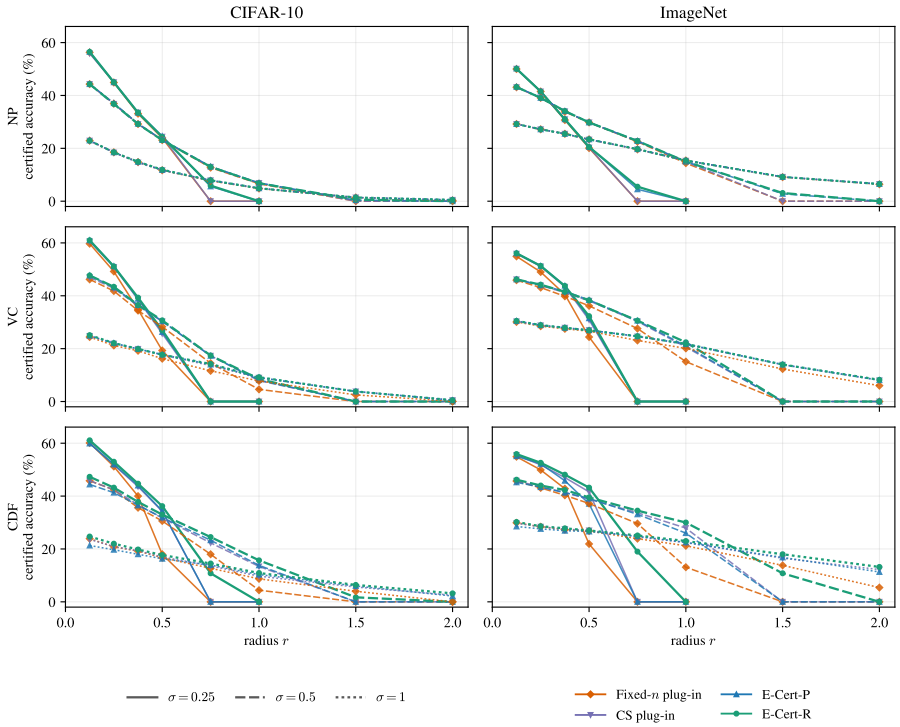


Figure 5: Certified accuracy against radius at three smoothing levels, 1,000 images per dataset and $n = 10^4$ evaluations, all four arms of Table 1; rows are certificate families, columns are datasets, color and marker are the procedure and line style the smoothing level. The four arms are nearly coincident for NP and separate with radius for VC and CDF, the two plug-ins falling to zero at the marginal radius where E-Cert-R still certifies.

Table 11: Certified accuracy (%) for the NP certificate across smoothing levels and radii, 1,000 images per dataset at $n = 10^4$, four arms. The $(\sigma, r) = (0.25, 0.125)$ column is the main paper's setting and matches Table 1. Within each σ block the column maximum is bold, columns where all four arms agree left plain; "–": $\rho = r/\sigma > 4$, outside the sweep. The $\rho \geq 3$ columns use the threshold-aware design grid of § B.2, which the rule selects from ρ and τ alone; every other column is the uniform grid it selects everywhere else.

			Perturbation radius r							
Dataset	σ	Method	0.125	0.25	0.375	0.5	0.75	1	1.5	2
CIFAR-10	0.25	Fixed-n plug-in	56.3	44.9	33.1	23.9	0.0	0.0	–	–
		CS plug-in	55.9	**45.0**	**33.5**	24.4	0.0	0.0	–	–
		E-Cert-P	**56.5**	**45.0**	**33.5**	**24.6**	5.6	0.0	–	–
		E-Cert-R	**56.5**	**45.0**	**33.5**	24.3	**5.9**	0.0	–	–
	0.5	Fixed-n plug-in	44.3	36.9	29.2	23.1	12.7	6.6	0.0	0.0
		CS plug-in	44.3	36.8	29.2	**23.2**	13.0	**6.9**	0.0	0.0
		E-Cert-P	44.3	**37.1**	**29.3**	**23.2**	**13.1**	**6.9**	0.4	0.0
		E-Cert-R	44.3	36.8	29.2	23.1	13.0	6.8	**0.6**	0.0
	1	Fixed-n plug-in	22.9	**18.5**	14.8	11.7	7.7	4.8	1.4	0.3
		CS plug-in	22.9	18.4	14.6	11.7	**7.9**	**4.9**	1.4	**0.5**
		E-Cert-P	**23.0**	**18.5**	**14.9**	11.8	**7.9**	**4.9**	1.4	**0.5**
		E-Cert-R	22.9	**18.5**	**14.9**	**11.9**	**7.9**	**4.9**	1.4	**0.5**
ImageNet	0.25	Fixed-n plug-in	50.0	41.5	30.7	20.1	0.0	0.0	–	–
		CS plug-in	50.1	41.5	**31.0**	20.3	0.0	0.0	–	–
		E-Cert-P	**50.2**	41.5	**31.0**	**20.6**	4.5	0.0	–	–
		E-Cert-R	50.1	41.5	**31.0**	20.5	**5.5**	0.0	–	–
	0.5	Fixed-n plug-in	43.1	39.1	33.8	29.8	22.5	14.4	0.0	0.0
		CS plug-in	43.1	39.1	34.0	29.8	22.7	14.9	0.0	0.0
		E-Cert-P	**43.3**	39.1	**34.1**	29.8	**22.8**	**15.0**	2.8	0.0
		E-Cert-R	43.2	39.1	**34.1**	29.8	**22.8**	**15.0**	**3.1**	0.0
	1	Fixed-n plug-in	29.2	27.2	25.5	**23.4**	19.6	**15.4**	9.1	6.4
		CS plug-in	29.2	27.2	25.5	23.3	19.6	15.3	**9.2**	**6.5**
		E-Cert-P	29.2	27.2	25.5	**23.4**	**19.7**	**15.4**	**9.2**	**6.5**
		E-Cert-R	29.2	27.2	25.5	**23.4**	**19.7**	**15.4**	**9.2**	**6.5**

Table 12: As Table 11, for the VC certificate.

Dataset	σ	Method	Perturbation radius r							
			0.125	0.25	0.375	0.5	0.75	1	1.5	2
CIFAR-10	0.25	Fixed-n plug-in	59.7	49.2	34.8	19.3	0.0	0.0	–	–
		CS plug-in	60.6	50.9	38.6	25.7	0.0	0.0	–	–
		E-Cert-P	**61.0**	**51.2**	38.9	26.3	0.0	0.0	–	–
		E-Cert-R	**61.0**	51.1	**39.3**	**27.0**	0.0	0.0	–	–
	0.5	Fixed-n plug-in	46.2	41.8	34.4	28.3	14.5	4.6	0.0	0.0
		CS plug-in	47.0	42.8	36.2	30.3	17.3	7.9	0.0	0.0
		E-Cert-P	**47.7**	43.2	**36.5**	30.4	17.3	8.1	0.0	0.0
		E-Cert-R	**47.7**	**43.4**	**36.5**	**30.7**	**17.4**	**8.5**	0.0	0.0
	1	Fixed-n plug-in	24.4	21.2	19.2	16.2	11.6	7.8	2.5	0.1
		CS plug-in	24.8	21.9	**19.9**	17.6	13.6	9.0	**3.8**	**0.5**
		E-Cert-P	24.9	**22.1**	**19.9**	**17.8**	14.0	9.1	**3.8**	**0.5**
		E-Cert-R	**25.1**	**22.1**	**19.9**	17.7	**14.2**	**9.2**	**3.8**	**0.5**
ImageNet	0.25	Fixed-n plug-in	54.9	49.0	40.5	24.4	0.0	0.0	–	–
		CS plug-in	55.9	51.1	43.4	30.9	0.0	0.0	–	–
		E-Cert-P	**56.1**	**51.3**	**43.8**	31.5	0.0	0.0	–	–
		E-Cert-R	**56.1**	**51.3**	43.7	**32.3**	0.0	0.0	–	–
	0.5	Fixed-n plug-in	45.9	43.1	39.8	36.2	27.6	15.1	0.0	0.0
		CS plug-in	**46.2**	43.9	41.2	38.1	30.3	20.6	0.0	0.0
		E-Cert-P	**46.2**	44.0	**41.7**	**38.4**	**30.6**	21.0	0.0	0.0
		E-Cert-R	**46.2**	**44.2**	41.5	38.3	**30.6**	**22.4**	0.0	0.0
	1	Fixed-n plug-in	30.1	28.5	27.5	26.6	23.1	20.1	12.3	6.0
		CS plug-in	30.3	28.8	**27.9**	**27.0**	24.6	21.6	13.9	8.0
		E-Cert-P	**30.5**	**28.9**	**27.9**	**27.0**	**24.7**	**21.8**	**14.0**	8.1
		E-Cert-R	**30.5**	**28.9**	**27.9**	26.9	**24.7**	**21.8**	**14.0**	**8.2**

Table 13: As Table 11, for the CDF certificate; Figure 2 in the main text plots its ImageNet part.

Dataset	σ	Method	Perturbation radius r 0.125	0.25	0.375	0.5	0.75	1	1.5	2
CIFAR-10	0.25	Fixed-n plug-in	60.1	51.2	40.0	18.0	0.0	0.0	–	–
		CS plug-in	60.1	51.8	43.6	34.2	0.0	0.0	–	–
		E-Cert-P	59.9	52.0	44.0	34.4	0.0	0.0	–	–
		E-Cert-R	**61.1**	**53.0**	**44.7**	**36.2**	**10.8**	0.0	–	–
	0.5	Fixed-n plug-in	45.8	42.4	35.6	30.6	18.0	4.4	0.0	0.0
		CS plug-in	45.8	42.4	36.7	31.7	22.3	13.4	0.0	0.0
		E-Cert-P	44.4	41.3	36.5	31.9	23.3	13.8	0.0	0.0
		E-Cert-R	**47.3**	**43.2**	**37.8**	**33.0**	**24.5**	**15.7**	**1.7**	0.0
	1	Fixed-n plug-in	23.9	20.8	19.2	16.7	12.7	8.7	4.0	0.0
		CS plug-in	23.6	20.9	19.3	17.0	13.6	9.7	5.6	2.5
		E-Cert-P	21.3	19.7	18.0	16.3	13.9	10.3	6.0	2.2
		E-Cert-R	**24.7**	**22.0**	**19.8**	**17.7**	**14.5**	**10.9**	**6.4**	**3.2**
ImageNet	0.25	Fixed-n plug-in	54.9	49.9	42.8	21.9	0.0	0.0	–	–
		CS plug-in	55.5	51.8	47.0	41.6	0.0	0.0	–	–
		E-Cert-P	55.1	52.2	45.8	37.1	0.0	0.0	–	–
		E-Cert-R	**55.9**	**52.6**	**48.1**	**43.2**	**19.0**	0.0	–	–
	0.5	Fixed-n plug-in	45.7	43.1	40.3	37.2	29.6	13.1	0.0	0.0
		CS plug-in	45.8	43.4	41.4	38.8	33.7	27.9	0.0	0.0
		E-Cert-P	45.2	43.4	41.3	39.1	33.1	26.0	0.0	0.0
		E-Cert-R	**46.2**	**44.0**	**42.3**	**39.5**	**34.5**	**30.0**	**10.8**	0.0
	1	Fixed-n plug-in	29.9	28.4	27.3	26.8	23.9	21.2	13.8	5.4
		CS plug-in	29.7	28.4	27.5	**27.0**	24.6	22.4	16.5	12.2
		E-Cert-P	28.5	27.6	26.9	26.6	24.8	22.5	16.8	11.3
		E-Cert-R	**30.2**	**28.6**	**27.8**	**27.0**	**25.1**	**23.0**	**18.0**	**13.2**

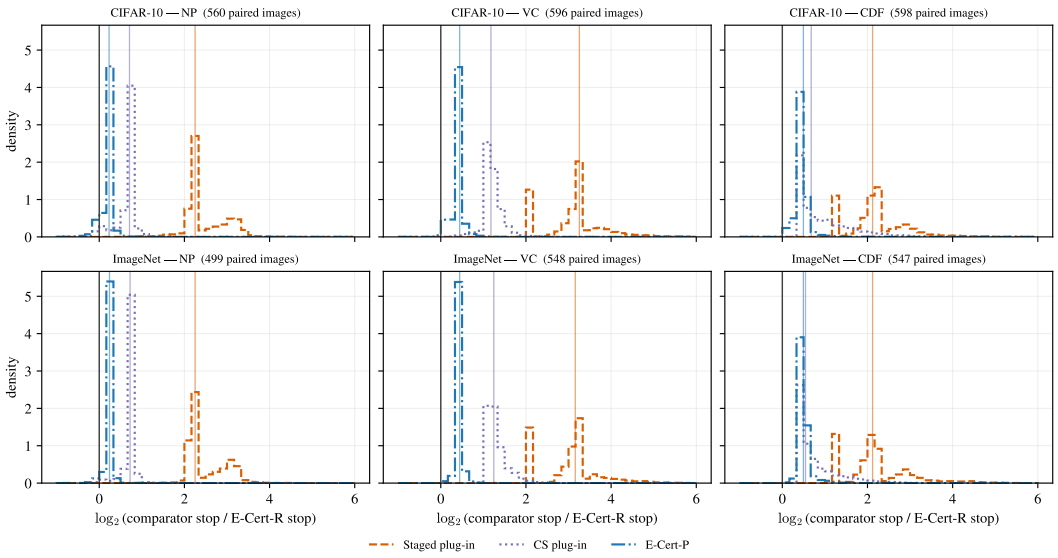


Figure 6: Per-image stopping-time ratio against E-Cert-R, on the images both arms certify; $\log_2$ scale, so 1 means twice as many evaluations and 0 means a tie. Vertical lines mark medians. Panel titles give the number of distinct images; each contributes ten permutation rows.

H STOPPING COST: DISTRIBUTIONS, FAIRNESS CONTROLS, FUTILITY, AND ROBUSTNESS CHECKS

We report stopping-cost distributions, monitoring-grid sensitivity and futility outcomes for the single-class experiments. Further controls vary the error level, evaluation budget, base-classifier training method and Monte Carlo draws.

H.1 STOPPING-COST DISTRIBUTIONS

Table 16 summarizes each arm by a median and a mean over all 1,000 images with an uncertified image charged the 10,000 cap. At this operating point 39–50% of images are uncertified. The mean is therefore dominated by that mass at the evaluation cap, and the median sits well above the typical certified cost. For CIFAR-10 NP the staged plug-in's all-images median is 500 in every permutation against 100 over the images it certifies. Certified accuracy and conditional stopping efficiency require separate summaries. Certified accuracy at a budget n is Table 1's columns at $n = 100$, 1,000 and 10,000 and Table 16's CA at the cap; the staged plug-in reaches it in steps at its ten checkpoints, which is the quantization the own-best-grid control of § H.3 addresses. Speed is the paired view of Figure 6, drawn from the same per-image records that produce the table (the generating script asserts that it reproduces all 24 published median and mean cells before drawing).

Figure 6 reports per-image stopping-time ratios on the images both compared arms certify, complementing the main text's aggregate comparisons. Nearly all of the mass lies above 1; the exception is NP, where up to 10.1% of paired rows favor E-Cert-P. The paired comparison shows that the aggregate advantage also holds for most individual images. The median factors are 4.3–9.5 against the staged plug-in, 1.5–2.4 against the CS plug-in and 1.2–1.4 against E-Cert-P. Only the first of those has a structural component: a staged arm cannot reject between checkpoints, so it can never stop before an exact crossing, though § H.3 shows that quantization accounts for at most 21% of its median. The other two compare arms that are both inspected every round. Being restricted to jointly certified images the figure says nothing about certified accuracy; each image contributes ten permutation rows, so the histogram is smoother than the count of distinct images suggests.

Table 14: Cost to match the fixed-n plug-in's full-budget accuracy. "CA at the cap" is each arm's certified accuracy at $n = 10^4$; "evaluations to match" is the smallest budget at which the arm's certified accuracy reaches the fixed-n plug-in's, mean $\pm$ SD over the ten permutations of Table 16; "saving" is 10^4 divided by it. "–": the arm's terminal accuracy is below the target, so it never matches.

Dataset	Cert.	Method	CA at the cap	Evaluations to match	Saving
CIFAR-10	NP	Fixed-n plug-in	56.3	10,000	1.0×
		Staged plug-in	56.0	–	–
		CS plug-in	56.2	$9{,}095 \pm 748$	1.1×
		E-Cert-P	56.5	$7{,}704 \pm 601$	1.3×
		E-Cert-R	56.5	$7{,}776 \pm 818$	1.3×
	VC	Fixed-n plug-in	59.7	10,000	1.0×
		Staged plug-in	59.6	–	–
		CS plug-in	60.7	$1{,}718 \pm 141$	5.9×
		E-Cert-P	61.0	971 ± 99	10.4×
		E-Cert-R	61.1	895 ± 110	11.3×
	CDF	Fixed-n plug-in	60.1	10,000	1.0×
		Staged plug-in	59.8	–	–
		CS plug-in	60.1	$9{,}297 \pm 402$	1.1×
		E-Cert-P	59.8	–	–
		E-Cert-R	61.2	$1{,}520 \pm 289$	6.8×
ImageNet	NP	Fixed-n plug-in	50.0	10,000	1.0×
		Staged plug-in	49.9	–	–
		CS plug-in	50.1	$6{,}334 \pm 2{,}197$	1.8×
		E-Cert-P	50.2	$5{,}015 \pm 1{,}637$	2.2×
		E-Cert-R	50.2	$5{,}098 \pm 1{,}270$	2.1×
	VC	Fixed-n plug-in	54.9	10,000	1.0×
		Staged plug-in	54.8	–	–
		CS plug-in	55.8	$2{,}081 \pm 121$	4.8×
		E-Cert-P	56.0	734 ± 60	13.7×
		E-Cert-R	56.1	560 ± 77	18.2×
	CDF	Fixed-n plug-in	54.9	10,000	1.0×
		Staged plug-in	54.7	–	–
		CS plug-in	55.5	$3{,}763 \pm 259$	2.7×
		E-Cert-P	55.2	$2{,}733 \pm 1{,}196$	4.2×
		E-Cert-R	55.9	623 ± 85	16.3×

H.2 Cost to match the fixed-n accuracy

Table 1 fixes the budget and compares accuracy; Table 16 fixes the accuracy target implicitly and compares cost. Table 14 states the combination directly: for each arm, the smallest budget at which its certified accuracy reaches what the fixed-n plug-in achieves with the whole 10^4, computed per permutation and reported as mean $\pm$ SD over the ten. E-Cert-R reaches that target with 1.3–18.2× fewer evaluations; the split across certificates is the one seen throughout, 1.3–2.1× for NP, where the four arms are nearly coincident, against 6.8–18.2× for VC and CDF, which use distributional information beyond the mean.

Entries marked "–" do not reach the target accuracy within the budget. The staged plug-in misses in every cell: spending $\alpha/10$ at each of ten checkpoints puts its terminal accuracy 0.1–0.3 points *below* the fixed-n bound it monitors, so the staged procedure does not reach the reference accuracy within this budget. E-Cert-P misses one cell (CIFAR-10 CDF, 59.8% against 60.1%), the same cell in which its prespecified tangent bank trails the refit design elsewhere.

H.3 Monitoring-grid sensitivity in the single-certificate benchmark

Table 16 monitors the staged plug-in on the same ten-point grid it reports at, so that arm's stopping cost can only take ten values while applying a Bonferroni factor of ten. We vary that grid here; § I.13 gives the corresponding multiclass control. We sweep the monitoring grid size m, each variant testing

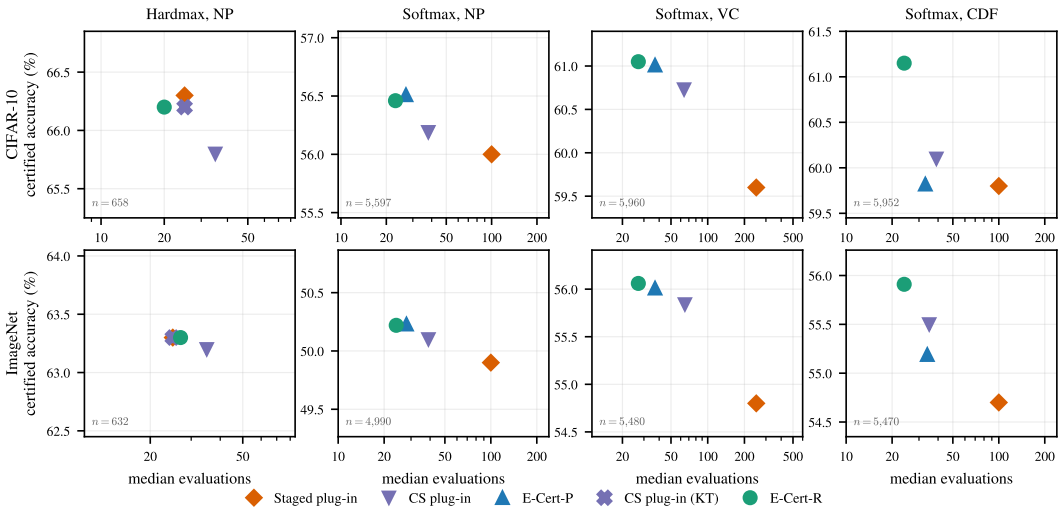


Figure 7: **Single-class** certified accuracy at $n = 10^4$ over all 1,000 images, versus median stop on the images every displayed arm certifies (n per panel). The 31–50% of uncertified inputs compress the all-images cost ratios in Figure 3 to 1.0–1.2; this conditional measure separates stopping efficiency. Columns denote observable–certificate pairs. Hardmax has no VC/CDF column because both functionals depend only on its mean. Fixed-budget procedures have no stopping time; futility leaves certified-image stops unchanged. E-Cert-R has the lowest median except on ImageNet hardmax, where staged Clopper–Pearson achieves equal accuracy at 25 evaluations versus 27.

at α/m, and recompute only the staged arm: the other three arms retain their results from the main run. Permutations are reproduced by importing the main script's own permutation routine, so the streams are identical seed-for-seed and image-for-image, and the $m = 10$ rows of Table 15 reproduce the published staged numbers exactly, a cross-check the generating script asserts at run time.

The control matters much less here than it does in the multiclass setting. The best grid lowers the staged median by 2.5% (NP) and 21.3% (VC) on CIFAR-10, and by 10.1% (CDF) and 4.9% (VC) on ImageNet. CDF on CIFAR-10 is already best at the paper's own $m = 10$. ImageNet NP sits at the 10,000-evaluation censoring cap under every grid, so no grid can move it. Mean costs move by at most 1.6% throughout, being dominated by censored non-certifying images. The contrast with § 7.2 is that these crossings happen late, at medians of several hundred to several thousand evaluations. Checkpoint spacing is then a small fraction of the distance to the e-processes. The multiclass crossings at 19–20 evaluations sit right where a ten-point grid is coarsest.

Against the staged arm at its best tested grid, both e-processes retain lower median and mean costs in all 60 dataset–certificate–seed comparisons each, 120 of 120 in total. Those come from six dataset–certificate cells times ten within-image permutations of one multiset, so they show consistency across conditions rather than 120 independent confirmations. The ordering is unchanged across the tested grids.

H.4 Futility stopping: protocol, full table, and audit

§ 4 shows that any adapted futility rule preserves the false-certification bound (Prop. 2) and that a rule stopping on an upper confidence bound for the population certificate also bounds the probability of abandoning a certifiable input, (6). This subsection gives the bounds, the protocol, the full eight-arm stopping table that Figure 3 draws, and an end-to-end audit of (6).

Upper bounds on the population certificate. Write ϵ for the per-checkpoint error and n for the prefix length at a checkpoint. For NP, the one-sided Hoeffding bound $U_n^M = \min\{1, \bar{X}_n + \sqrt{\log(1/\epsilon)/(2n)}\}$ satisfies $\mathbb{P}(U_n^M < M_P) \le \epsilon$, and since $\mathrm{Cert}_{\mathrm{NP}}$ is increasing in the mean, $\Phi(\Phi^{-1}(U_n^M) - \rho)$ upper-bounds the certificate; the rule triggers when $U_n^M \le T_\rho(\tau)$. For VC the certificate is increasing in the mean and decreasing in the variance, so an upper bound needs an upper

Table 15: Staged plug-in under a swept monitoring grid, ten paired permutations, all 1,000 images per dataset with non-certifying images censored at 10,000 evaluations. m is the realized number of checkpoints (requested 100 and 250 de-duplicate to 99 and 230), each test at α/m. Entries are across-seed mean $\pm$ SD. $\star$ marks each certificate's own best grid, chosen on the median and tie-broken on the mean. The $m = 10$ rows are the grid used in Table 16.

Dataset	Cert.	Looks m	Median	Mean
CIFAR-10	NP	10	500.0 ± 0.0	$4{,}625.6 \pm 8.0$
		25	506.0 ± 64.3	$4{,}594.3 \pm 4.1$
		50*	487.4 ± 25.1	$4{,}591.4 \pm 2.7$
		99	500.0 ± 15.5	$4{,}595.0 \pm 3.0$
		230	524.9 ± 17.6	$4{,}602.8 \pm 3.3$
	VC	10	750.0 ± 235.7	$4{,}468.4 \pm 6.5$
		25*	590.2 ± 63.4	$4{,}403.9 \pm 5.4$
		50	605.1 ± 28.8	$4{,}395.8 \pm 3.9$
		99	638.4 ± 22.2	$4{,}401.6 \pm 3.5$
		230	667.5 ± 19.8	$4{,}417.0 \pm 3.3$
	CDF	10*	275.0 ± 79.1	$4{,}265.3 \pm 7.8$
		25	316.0 ± 0.0	$4{,}227.0 \pm 2.4$
		50	299.4 ± 13.9	$4{,}226.2 \pm 3.0$
		99	307.4 ± 9.8	$4{,}231.8 \pm 2.4$
		230	319.6 ± 9.3	$4{,}241.6 \pm 2.7$
ImageNet	NP	10	$10{,}000.0 \pm 0.0$	$5{,}166.6 \pm 1.7$
		25	$10{,}000.0 \pm 0.0$	$5{,}144.2 \pm 1.3$
		50*	$10{,}000.0 \pm 0.0$	$5{,}142.6 \pm 1.5$
		99	$10{,}000.0 \pm 0.0$	$5{,}145.5 \pm 1.6$
		230	$10{,}000.0 \pm 0.0$	$5{,}150.7 \pm 1.5$
	VC	10	$1{,}000.0 \pm 0.0$	$4{,}832.6 \pm 3.8$
		25	$1{,}000.0 \pm 0.0$	$4{,}788.6 \pm 1.2$
		50*	951.4 ± 58.2	$4{,}784.0 \pm 1.7$
		99	979.9 ± 23.4	$4{,}787.4 \pm 1.3$
		230	$1{,}014.4 \pm 25.5$	$4{,}797.5 \pm 1.1$
	CDF	10	625.0 ± 212.5	$4{,}742.2 \pm 3.6$
		25*	562.0 ± 0.0	$4{,}711.6 \pm 2.9$
		50	605.1 ± 19.2	$4{,}710.0 \pm 1.8$
		99	620.8 ± 23.0	$4{,}713.2 \pm 1.8$
		230	654.6 ± 14.2	$4{,}720.7 \pm 1.8$

mean and a *lower* variance. With $m = \lfloor n/2 \rfloor$ and $W_j = (X_{2j-1} - X_{2j})^2/2$, which has $\mathbb{E}W_j = V_P$ by independence and range $[0, 1/2]$, Hoeffding gives $L_m^V = \max\{0, \overline{W}_m - \frac{1}{2}\sqrt{\log(2/\epsilon)/(2m)}\}$; with U_n^M at $\epsilon/2$ as well, $U_n^M - \kappa_\rho\sqrt{L_m^V}$ is an upper bound on $\mathrm{Cert}_{\mathrm{VC}}$ with error at most ϵ. For CDF the one-sided DKW band $F_{n,i}^L = \max\{0, \widehat{F}_{n,i} - \sqrt{\log(1/\epsilon)/(2n)}\}$ lower-bounds every $F_{P,i}$ at once with error at most ϵ (Massart, 1990), the same constant the fixed-n CDF comparator uses; because T_ρ is increasing, $b_{K-1} - \sum_i \Delta_i T_\rho(F_{n,i}^L)$ upper-bounds $\mathrm{Cert}_{\mathrm{CDF}}$. In every case the futility statistic is a bound on the same population functional the certificate reads, so a small certificate lower bound alone does not justify abstention, since another witness may still allow certification.

Protocol. Futility is checked at the staged monitor's own checkpoints from 100 to 5,000 $(100, 250, 500, 1{,}000, 2{,}500, 5{,}000)$, never at the cap, where the procedure abstains anyway. The population error $\beta_{\mathrm{f}} = \alpha$ is split equally, so each checkpoint spends $\epsilon = \alpha/6$; a union bound over the six checkpoints gives the simultaneous coverage (6) requires. The futility time F depends on the score stream alone, not on the e-process watching it. The two futility arms are therefore the published per-image crossing times T of Table 16's e-processes, combined with F on the identical permuted streams: stop at $\min(T, F, N)$, certify iff $T \le \min(F, N)$, ties certifying. Nothing about the e-processes is recomputed. The generating script asserts that disabling futility reproduces every published cell; that each upper bound dominates the corresponding fixed-n lower certificate at every

checkpoint on real prefixes; and that an all-zero stream triggers at the first checkpoint while an all-one stream never does.

Futility outcomes on the cached streams. Table 16 reports all eight arms: the four procedures, and each with the futility rule. The rule depends on the stream alone, so it applies to the plug-ins exactly as to the e-processes. Three columns describe the futility arms: the fraction of images resolved by futility, the fraction still charged the cap, and the number of certificates *lost*. Lost counts images the base arm certifies within the budget but the rule abandons first, summed over the ten permutations. The base arms leave 39–50% of images at the cap; with futility 1.7–6.0% remain there, the rest having been abstained on between 100 and 5,000 evaluations. The lost count is zero in all twenty-four cells, over 240,000 image–permutation pairs. The mean cost therefore falls without any accuracy movement. For E-Cert-R it falls by 8.2×, 4.9× and 7.7× on CIFAR-10 NP/VC/CDF, and by 13.1×, 7.2× and 14.9× on ImageNet. The other arms gain comparable factors (3.4–9.9× for the staged plug-in, 4.6–12.6× for the CS plug-in). After futility E-Cert-R's mean is 1.04–1.36× below the CS plug-in's and 1.34–1.97× below the staged plug-in's, so the ordering of the base arms is preserved but compressed: the rule removes most of the cap mass for every procedure alike. ImageNet's uncertifiable images have target probabilities far from the boundary, so almost all of them trigger at the first checkpoint ($n = 100$), which is why the ImageNet NP median sits on the boundary between the certified and the capped mass – the arms straddle half at 49.9–50.2%, so which side the median falls on changes with the arm – and moves from 5,548 to exactly 100. CIFAR-10's uncertifiable images sit closer to the boundary and 3–5% never trigger. A median moves only if abstained images, assigned their abstention time instead of the cap, fall below it. E-Cert-R's medians are all under 100 except ImageNet NP, so only that one moves. For E-Cert-P, ImageNet NP (5,465) and the ImageNet VC and CDF medians of 103 and 113 drop to 100.

Audit of the population-protection claim. We measure (6) end to end. For each certificate a $\mathrm{Beta}(8, b)$ score law is placed exactly on the boundary $\mathrm{Cert}_j(P; \rho) = \tau$ by solving for b ($b = 3.57$, 6.09 and 6.05 for NP, VC and CDF). The audit evaluates these boundary laws, without claiming they are least favorable. We then run 20,000 independent streams of 5,000 draws through the deployed futility code. No stream triggers for any certificate (0 of 20,000 for NP, VC and CDF alike); the exact 95% upper confidence limit on the trigger rate is 1.5×10^{-4}, well inside $\beta_{\mathrm{f}} = 10^{-3}$. The bounds are Hoeffding and DKW, so the rule is conservative rather than tight.

Scope. The zero observed loss should be read alongside the probabilistic guarantee (6): an input whose population certificate exceeds τ is abandoned before it certifies with probability at most β_{f}, whether certification would have come within N draws or not. The lost column is the realized count of within-budget certificates abandoned on these streams, a sample statistic, which the guarantee bounds in expectation at β_{f} per certifiable input but does not fix. An input with $\mathrm{Cert}_j(P; \rho) \leq \tau$ that the base arm would nonetheless have certified is not covered by this protection: such a certification is false and already has probability at most α.

Slope-based futility. The population rule needs a confidence bound on a population functional. A cheaper alternative reads the capital path alone: fit the recent trend of $\log \mathcal{K}_t$ at a checkpoint and abstain once the extrapolation cannot reach $\log(1/\alpha)$ by the cap, subject to a minimum count $n_{\min}$ before the rule may fire. This preserves false-certification control by Prop. 2. A fitted slope does not bound future growth, however, so the rule lacks (6)'s abandonment guarantee. At $n_{\min} = 500$ it is cheaper than the population rule in all six single-class cells, by 1.2–3.3×, and *loses* 0.4–1.0 points of certified accuracy doing so; at $n_{\min} = 2{,}500$ it is dominated in 8 of 10 cells, with both higher cost and certificate loss. Figure 3 draws it on E-Cert-R at three minimum counts, where a hollow square below the dotted accuracy line is exactly a discarded certificate. Full per-arm numbers are in the released results; the remaining experiments use population futility for its abandonment guarantee.

H.5 Sensitivity to the error level

The default efficiency experiments use $\alpha = 10^{-3}$. We assess whether the relative efficiency depends on this choice. Table 17 sweeps $\alpha \in \{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$ for all four arms of Table 16, on all 1,000 images of each dataset at the primary permutation seed.

Table 16: Stopping evaluations over ten paired permutations, over *all* 1,000 images per dataset, an image not certified by 10,000 evaluations charged 10,000. Median and mean are across-permutation mean ± SD; CA is certified accuracy (%) at the cap; Futility and Cap are the shares of images resolved by abstention and by the budget; Lost counts certificates the base arm reaches within budget that the futility rule abandons first, summed over permutations; "–": the arm has no futility rule. Smallest median and mean per block in bold. Protocol: dense CDF checkpointing, an every-round RIPr refit through round 200 then a geometric schedule, the ten-point staged grid, and futility checkpoints at 100–5,000 with $\beta_f = \alpha$ (§ D.2, § H.4).

Dataset	Cert.	Method	Median	Mean	CA	Futility	Cap	Lost
CIFAR-10	NP	Staged plug-in	500.0 ± 0.0	4,625.6 ± 8.0	56.0	–	44.0%	–
		CS plug-in	108.8 ± 3.1	4,462.9 ± 6.9	56.2	–	43.8%	–
		E-Cert-P	88.5 ± 4.4	4,438.5 ± 4.6	56.5	–	43.5%	–
		E-Cert-R	**87.9 ± 5.5**	4,442.0 ± 3.9	56.5	–	43.5%	–
		Staged plug-in + futility	100.0 ± 0.0	727.3 ± 8.8	56.0	40.4%	3.6%	0
		CS plug-in + futility	100.0 ± 0.0	564.5 ± 7.8	56.2	40.4%	3.4%	0
		E-Cert-P + futility	88.5 ± 4.4	**540.2 ± 4.9**	56.5	40.4%	3.0%	0
		E-Cert-R + futility	**87.9 ± 5.5**	543.6 ± 5.2	56.5	40.4%	3.1%	0
	VC	Staged plug-in	750.0 ± 235.7	4,468.4 ± 6.5	59.6	–	40.4%	–
		CS plug-in	143.2 ± 3.2	4,055.5 ± 7.3	60.7	–	39.3%	–
		E-Cert-P	74.2 ± 1.8	3,992.1 ± 7.0	61.0	–	39.0%	–
		E-Cert-R	**57.3 ± 2.1**	3,975.5 ± 6.4	61.0	–	39.0%	–
		Staged plug-in + futility	250.0 ± 0.0	1,298.7 ± 9.0	59.6	34.4%	6.0%	0
		CS plug-in + futility	100.0 ± 0.0	885.7 ± 8.4	60.7	34.4%	4.9%	0
		E-Cert-P + futility	74.2 ± 1.8	822.3 ± 7.4	61.0	34.4%	4.6%	0
		E-Cert-R + futility	**57.3 ± 2.1**	**805.7 ± 7.2**	61.0	34.4%	4.6%	0
	CDF	Staged plug-in	275.0 ± 79.1	4,265.3 ± 7.8	59.8	–	40.2%	–
		CS plug-in	168.8 ± 7.4	4,125.0 ± 2.1	60.1	–	39.9%	–
		E-Cert-P	72.4 ± 2.4	4,087.1 ± 10.2	59.8	–	40.2%	–
		E-Cert-R	**55.3 ± 1.5**	3,967.2 ± 6.0	61.2	–	38.9%	–
		Staged plug-in + futility	100.0 ± 0.0	810.5 ± 9.9	59.8	35.9%	4.3%	0
		CS plug-in + futility	100.0 ± 0.0	670.1 ± 8.6	60.1	35.9%	4.0%	0
		E-Cert-P + futility	72.4 ± 2.4	632.3 ± 13.2	59.8	35.9%	4.2%	0
		E-Cert-R + futility	**55.3 ± 1.5**	**512.4 ± 9.4**	61.2	35.9%	2.9%	0
ImageNet	NP	Staged plug-in	10,000.0 ± 0.0	5,166.6 ± 1.7	49.9	–	50.1%	–
		CS plug-in	7,428.6 ± 1,457.7	5,043.5 ± 4.4	50.1	–	49.9%	–
		E-Cert-P	5,465.1 ± 1,412.8	5,026.5 ± 4.5	50.2	–	49.8%	–
		E-Cert-R	5,548.4 ± 1,194.5	5,025.8 ± 3.6	50.2	–	49.8%	–
		Staged plug-in + futility	**100.0 ± 0.0**	523.1 ± 6.1	49.9	47.6%	2.5%	0
		CS plug-in + futility	**100.0 ± 0.0**	400.0 ± 6.3	50.1	47.6%	2.2%	0
		E-Cert-P + futility	**100.0 ± 0.0**	383.0 ± 6.2	50.2	47.6%	2.1%	0
		E-Cert-R + futility	**100.0 ± 0.0**	**382.3 ± 5.8**	50.2	47.6%	2.1%	0
	VC	Staged plug-in	1,000.0 ± 0.0	4,832.6 ± 3.8	54.8	–	45.2%	–
		CS plug-in	234.9 ± 6.7	4,525.2 ± 3.3	55.8	–	44.2%	–
		E-Cert-P	103.4 ± 3.9	4,460.7 ± 5.1	56.0	–	44.0%	–
		E-Cert-R	**75.7 ± 2.9**	4,446.6 ± 5.4	56.1	–	43.9%	–
		Staged plug-in + futility	250.0 ± 0.0	1,006.7 ± 5.9	54.8	40.1%	5.1%	0
		CS plug-in + futility	100.0 ± 0.0	699.2 ± 5.4	55.8	40.1%	4.0%	0
		E-Cert-P + futility	99.8 ± 0.4	634.8 ± 6.5	56.0	40.1%	3.9%	0
		E-Cert-R + futility	**75.7 ± 2.9**	**620.6 ± 5.9**	56.1	40.1%	3.8%	0
	CDF	Staged plug-in	625.0 ± 212.5	4,742.2 ± 3.6	54.7	–	45.3%	–
		CS plug-in	219.2 ± 8.5	4,559.1 ± 2.8	55.5	–	44.5%	–
		E-Cert-P	113.3 ± 4.3	4,538.9 ± 4.4	55.2	–	44.8%	–
		E-Cert-R	**74.6 ± 3.2**	4,451.7 ± 5.1	55.9	–	44.1%	–
		Staged plug-in + futility	100.0 ± 0.0	589.7 ± 5.6	54.7	42.3%	2.9%	0
		CS plug-in + futility	100.0 ± 0.0	406.6 ± 5.0	55.5	42.3%	2.1%	0
		E-Cert-P + futility	100.0 ± 0.0	386.4 ± 5.5	55.2	42.3%	2.5%	0
		E-Cert-R + futility	**74.6 ± 3.2**	**299.2 ± 3.4**	55.9	42.3%	1.7%	0

The log-capital threshold suggests an approximate scaling relation. A betting process crosses near $\log(1/\alpha)/\mathbb{E}[\log e_t]$ rounds, so every sequential arm's stopping time should grow like $\log(1/\alpha)$ while

the *ratio* between two arms, the reciprocal ratio of their per-round log-growths, remains approximately constant. The e-process results can be obtained from the same capital paths because neither e-process design depends on α. E-Cert-P's mixture is prespecified, and E-Cert-R's working law, witness and stake are functions of completed scores alone, so α enters only through the threshold the capital is compared against. One capital path per stream is therefore read at all four levels, and those columns come from literally the same run; only the two plug-in arms, which put α inside the bound, are recomputed per level. The generating script asserts that at $\alpha = 10^{-3}$ all four arms reproduce the published per-image stopping times.

Stopping-cost scaling and relative efficiency. E-Cert-R's median stopping time tracks $\log(1/\alpha)$ closely. For CIFAR-10 CDF it runs $37/53.5/71/87$ against the $37/55.5/74/92.5$ that pure logarithmic growth from the loosest level predicts. In 70 of the 72 comparator cells no comparator has a lower median stopping cost than E-Cert-R, the ratio reaching 20.41. Both exceptions are NP cells in which E-Cert-P is level with it. One is CIFAR-10 at $\alpha = 10^{-5}$ (132.5 against 133). The other is ImageNet at $\alpha = 10^{-4}$, where the two arms certify the same 50.2% of images, so their all-images medians ($7{,}836$ against $9{,}114$) straddle the point at which the censored mass begins while their means differ by less than two evaluations. The ratios themselves move little. Over three orders of magnitude the largest and smallest value of a given ratio against the two every-round comparators differ by a median factor of 1.12, and at worst 1.40, for CIFAR-10 CDF against the CS plug-in, where 3.73 falls to 2.67. Including the staged arm, whose checkpoint quantization makes its ratio jump, the median is 1.22 and the worst 2.04. Where they do move they mostly shrink, so the advantage is largest at the *loosest* level: $\alpha = 10^{-2}$ maximizes it in 11 of the 18 dataset–certificate–comparator triples and the reported 10^{-3} in only 3. Thus the default level does not maximize the observed relative advantage.

Certified accuracy and checkpoint effects. Certified accuracy is almost α-free at this budget: tightening three orders of magnitude costs at most 0.7 percentage points on any of the 24 arm–cell combinations, and at most 0.6 on E-Cert-R. At $n = 10^4$, stricter error control primarily increases sample cost, with little change in certified accuracy. And the staged arm's ratio is the most strongly non-monotone ($8.93/6.17/4.81/7.52$ on CIFAR-10 NP): it may stop only at one of ten checkpoints, so its median jumps between grid points as α moves and the ratio bounces with it. This is the checkpoint effect evaluated in § H.3.

H.6 The benchmark at a 10^5 evaluation budget

Every table in this paper caps the stream at $n = 10^4$, while much of the randomized-smoothing literature certifies with 10^5 draws per input (Cohen et al., 2019). An order of magnitude more evaluations works in the comparators' favor. A terminal plug-in bound keeps tightening at rate $n^{-1/2}$, whereas an e-process that stopped after a few dozen rounds gains nothing. It also raises the price of an image no procedure can certify, since the cap is what such an image costs. We therefore drew fresh 10^5-draw caches with the same checkpoints and the same $(\sigma, r, \alpha) = (0.25, 0.125, 10^{-3})$. The four monitored arms and the fixed-n plug-in were re-run with the cap, the reported budgets and the staged monitoring grid all extended. The staged plug-in now looks at thirteen checkpoints, the paper's $1/2.5/5$-per-decade grid continued to 10^5, each at level $\alpha/13$. Streams are used in the order they were drawn, so no permutation seed enters. The generating script asserts that the same code path reproduces the published per-image stopping times on the 10^4 cache before any of this is run.

Certified accuracies converge as the budget increases: the spread of certified accuracy across the five procedures falls from 12–47 percentage points at $n = 100$ to 1.6–6.2 at 10^3, 0.4–1.4 at 10^4 and 0.0–1.4 at 10^5, where on ImageNet NP four of the five read 50.6% and the staged plug-in 50.5%. The main benefit at large budgets is earlier certification (§ 7.1). Stopping costs on certified images remain similar across the two caps. E-Cert-R's median over the images it certifies is $23/27/24$ evaluations for NP/VC/CDF on CIFAR-10 and $24/27/24$ on ImageNet at the 10^5 cap, against $23/27/25$ and $24/27/24$ on the 10^4 caches run through the same code. The two caches use independent draws. The comparators also have similar certified-image stopping costs at both caps. On ImageNet the certified-image medians are $100/250/100$ (staged plug-in), $39/66/35$ (CS plug-in) and $28/37/34$ (E-Cert-P) at *both* caps. Those are ratios of 4.2–$9.3\times$, 1.5–$2.4\times$ and 1.2–$1.4\times$ to E-Cert-R, and CIFAR-10 agrees to within one evaluation in every cell.

Table 17: Median stopping evaluations against the error level, on all 1,000 images of each dataset at permutation seed 2026, with an image uncertified at 10,000 charged that cap as in Table 16. Both e-process columns are one run per stream read at four thresholds, since neither design depends on α. Ratios are to E-Cert-R ("–" for that arm itself); larger is worse for the comparator.

			Median stop				Ratio to E-Cert-R			
Dataset	Cert.	Method	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
CIFAR-10	NP	Staged plug-in	500	500	500	1,000	8.93	6.17	4.81	7.52
		CS plug-in	71.5	104	139.5	180.5	1.28	1.28	1.34	1.36
		E-Cert-P	56	84	110.5	132.5	1.00	1.04	1.06	1.00
		E-Cert-R	56	81	104	133	–	–	–	–
	VC	Staged plug-in	500	1,000	1,000	1,000	13.51	18.02	13.42	10.99
		CS plug-in	100	141.5	187.5	227	2.70	2.55	2.52	2.49
		E-Cert-P	52.5	73	92	112	1.42	1.32	1.23	1.23
		E-Cert-R	37	55.5	74.5	91	–	–	–	–
	CDF	Staged plug-in	250	250	500	500	6.76	4.67	7.04	5.75
		CS plug-in	138	177	200	229	3.73	3.31	2.82	2.63
		E-Cert-P	52	70.5	89.5	109.5	1.41	1.32	1.26	1.26
		E-Cert-R	37	53.5	71	87	–	–	–	–
ImageNet	NP	Staged plug-in	10,000	10,000	10,000	10,000	2.10	1.72	1.10	1.04
		CS plug-in	5,445	7,497.5	10,000	10,000	1.14	1.29	1.10	1.04
		E-Cert-P	4,951.5	5,965.5	7,836	9,957.5	1.04	1.03	0.86	1.04
		E-Cert-R	4,760	5,799	9,113.5	9,572.5	–	–	–	–
	VC	Staged plug-in	1,000	1,000	1,000	2,500	18.52	12.82	10.00	20.41
		CS plug-in	171.5	240	306.5	373	3.18	3.08	3.06	3.04
		E-Cert-P	76	104	128.5	159.5	1.41	1.33	1.28	1.30
		E-Cert-R	54	78	100	122.5	–	–	–	–
	CDF	Staged plug-in	500	1,000	1,000	1,000	9.35	13.07	10.31	8.37
		CS plug-in	180.5	219.5	260.5	303.5	3.37	2.87	2.69	2.54
		E-Cert-P	88	117.5	146	175.5	1.64	1.54	1.51	1.47
		E-Cert-R	53.5	76.5	97	119.5	–	–	–	–

All-images mean costs grow approximately in proportion to the cap: 38,776–43,761 evaluations on CIFAR-10 and 43,757–49,903 on ImageNet against 3,971–4,621 and 4,440–5,167 at 10^4, and for E-Cert-R essentially all of that is the 38.5–49.4% of images it does not certify, each charged 10^5. Population futility reduces the cost of these inputs, with larger absolute savings at the increased budget. It is added to every arm at the same $\beta_f = \alpha$, now split over the nine checkpoints the $1/2.5/5$-per-decade grid reaches before the new cap. That cuts the mean cost from 38,776–49,903 to 994–3,788 evaluations (10.6–49.9×, against 3.4–14.9× at 10^4), resolves 37–49% of images by abstention, and loses no certificate in any of the 24 arm–certificate–dataset cells (Table 18). The absolute saving is what scales with the cap: about 41,000 evaluations per image against about 3,500 at 10^4, roughly twelvefold. Only the futility times are new here; each arm's crossing times are the ones already computed for Table 18. All-images *medians* inherit the same censoring: on ImageNet NP, where 49.4% of images are uncertified, that median sits on the boundary between certified and censored mass and moves from 5,558 to 3,949 between the two caches, which is why § H.1 reports the two views separately. E-Cert-P's CDF accuracy trails E-Cert-R by 1.4 points on CIFAR-10 and 0.8 on ImageNet, much as at 10^4 (1.2 and 0.8): a prespecified tangent bank does not sharpen with the stream, and a tenfold longer stream does not change that.

The CDF comparator has an additional computational limitation in this run. The screened exact search admits every prefix a crossing could occupy, and each admitted prefix costs $O(nK)$; at the 10^5 horizon a stream that crosses near the cap admits tens of thousands of them. Twenty-three of the 2,000 streams therefore did not finish inside the twenty-minute guard we run every component under. Twenty-one of them are known to cross later, and we report the bracket's lower end, "no crossing at or below this prefix". That is a cost *bound*, not a realized stopping time: replacing it by the known upper end would move the mean by 107 (CIFAR-10) and 75 (ImageNet) evaluations, over

Table 18: The single-class benchmark at a 10^5 evaluation budget, $1{,}000$ freshly sampled images per dataset at $\sigma = 0.25$ and $r = 0.125$. Certified accuracy is read at four budgets on the same streams. "Median (certified)" is over the images that arm certifies; the two mean columns charge every image, an uncertified one at the 10^5 cap, so the two use different populations. The futility column adds the population rule, which spends $\beta_{\mathrm{f}} = \alpha$ over nine checkpoints (100 to $50{,}000$) and reuses each arm's crossing times, so only the futility time is new; it changes no arm's certified accuracy and loses no certificate in any cell, and its saving is the ratio of the two mean columns. The staged plug-in looks at thirteen checkpoints at level $\alpha/13$; the fixed-n plug-in is a separately planned terminal analysis at each budget and so has no stopping time.

Dataset	Cert.	Method	Certified accuracy (%)				Median	Mean evaluations	
			10^2	10^3	10^4	10^5	(certified)	no futility	+ futility
CIFAR-10	NP	Fixed-n plug-in	36.0	54.0	56.1	56.9	–	–	–
		Staged plug-in	29.5	53.4	55.9	56.9	100	43,761	2,056
		CS plug-in	49.4	55.1	56.1	56.9	38	43,402	1,697
		E-Cert-P	50.9	55.3	56.3	56.9	27	43,303	1,598
		E-Cert-R	50.7	55.3	56.2	56.9	23	43,307	1,602
	VC	Fixed-n plug-in	29.0	54.7	60.0	60.9	–	–	–
		Staged plug-in	7.0	53.6	59.7	60.9	250	40,055	3,788
		CS plug-in	44.7	58.8	60.5	61.3	65	39,063	2,796
		E-Cert-P	53.1	59.7	60.8	61.5	37	38,843	2,576
		E-Cert-R	54.2	59.8	60.9	61.5	27	38,786	2,519
	CDF	Fixed-n plug-in	44.6	56.4	60.3	61.2	–	–	–
		Staged plug-in	40.6	55.9	60.0	61.1	100	39,662	2,687
		CS plug-in	45.5	56.8	60.3	61.2	40	39,167	2,192
		E-Cert-P	53.0	58.6	59.8	60.0	33	40,088	3,113
		E-Cert-R	54.1	59.7	60.9	61.4	24	38,776	1,801
ImageNet	NP	Fixed-n plug-in	36.2	48.6	50.2	50.6	–	–	–
		Staged plug-in	28.4	48.2	49.8	50.5	100	49,903	1,320
		CS plug-in	44.9	49.6	50.3	50.6	39	49,638	1,055
		E-Cert-P	46.6	49.8	50.4	50.6	28	49,577	994
		E-Cert-R	46.9	49.8	50.4	50.6	24	49,578	995
	VC	Fixed-n plug-in	29.3	51.8	54.9	56.1	–	–	–
		Staged plug-in	7.5	51.0	54.8	56.0	250	44,955	3,342
		CS plug-in	42.9	54.1	55.8	56.3	66	43,930	2,317
		E-Cert-P	49.5	55.2	56.2	56.5	37	43,777	2,164
		E-Cert-R	51.5	55.4	56.2	56.4	27	43,757	2,144
	CDF	Fixed-n plug-in	43.5	52.7	55.0	56.0	–	–	–
		Staged plug-in	39.4	51.9	54.7	55.7	100	44,999	2,256
		CS plug-in	45.3	54.0	55.5	56.1	35	44,151	1,409
		E-Cert-P	48.9	54.5	55.2	55.4	34	44,691	1,948
		E-Cert-R	51.6	55.4	56.1	56.2	24	43,875	1,133

the twenty-one streams whose crossing is known to exist; the two whose existence is still open add an uncertainty this width does not cover. For the two whose existence is still open the stream is reported as uncertified, since a bound that rules out a crossing is not evidence that one exists.

Four of the twenty-one brackets straddled $n = 10^4$, so the budget columns would have credited a certificate whose existence at that budget was undecided. Rerunning the exact search on those four truncated at 10^4 settles them: none crosses at or before the budget, and their first possible crossing is $10^4 + 1$. The CDF CS entry at $n = 10^4$ is therefore 60.3 on CIFAR-10 and 55.5 on ImageNet, two images below what the bracket's lower end had implied. Our own arms cost at most 78 seconds per image on the same streams.

H.7 Other base classifiers

To assess sensitivity to the base classifier, we repeat the benchmark on consistency-regularized checkpoints (Jeong & Shin, 2020), alongside the default checkpoints of Cohen et al. (2019). Both datasets use the same architectures and $\sigma = 0.25$ as before. Everything else is held fixed at $r = 0.125$, $\alpha = 10^{-3}$, 1,000 images, the 10,000 cap and the ten-checkpoint staged grid at $\alpha/10$. The Cohen block of Table 19 is re-run through the identical code path, so the two training-method blocks differ only in the checkpoint. These runs use the streams in the order they were drawn rather than the ten permutations of Table 16, so the Cohen block is directly comparable with Table 1. It reproduces that table: certified accuracies match in 71 of its 72 cells, and differ by a single image in the last (CIFAR-10 CDF, CS plug-in at $n = 100$, 44.9% against 44.8%). There a budget's accuracy is read off the first-crossing time rather than recomputed at that budget.

Consistency training raises certified accuracy by 1.2–4.9 points over Cohen's checkpoints when comparing the same procedure, dataset and certificate. Across both training methods and datasets, the four procedures' certified accuracies at $n = 10^4$ range from 49.9% to 65.2%; within any one checkpoint–certificate setting their spread is at most 1.4 points. The cost ratios to E-Cert-R remain similar: the staged plug-in needs 4.0–9.3× as many evaluations on Cohen's checkpoints and 4.3–9.6× on the consistency checkpoints; the CS plug-in needs 1.5–2.4× and 1.4–2.3×, respectively, and E-Cert-P 1.2–1.4× and 1.2–1.5×. E-Cert-R's median among certified images is 22–27 evaluations across these settings. These medians describe each method's own certified subset; they are not a paired speed comparison on a common-success set.

The small-budget accuracy gap of Table 1 persists under consistency training. At $n = 100$ E-Cert-R certifies $51.1/54.2/54.4\%$ of CIFAR-10 images for NP/VC/CDF against the fixed-n plug-in's $36.7/28.4/44.2\%$ on Cohen's checkpoint, and $53.4/59.7/59.6\%$ against $42.6/36.4/48.1\%$ on the consistency checkpoint. The staged plug-in has a terminal accuracy shortfall of 0.2–1.4 points relative to the best procedure in each checkpoint–certificate setting, consistent with its coarser grid and $\alpha/10$ level per checkpoint. What remains unassessed is variation across architectures and across training seeds of one method (§ J); these runs vary the training procedure, not the architecture.

H.8 Sensitivity to the certification threshold

Everything reported so far fixes $\tau = 1/2$, where the certificate is a statement about the predicted label. The threshold is a free parameter of the null, and raising it turns the same machinery into a statement about *confidence*: at $\tau = 0.7$, certifying means the smoothed score cannot fall to 0.7 anywhere in the ball. This is the regime Kumar et al. (2020a) report in, and requires only updating the null: we recompute the NP threshold, CDF constant and VC feasible range while retaining the witness families, stake rules and refit schedule. Table 20 runs the five single-class arms at $\tau \in \{0.5, 0.6, 0.7, 0.8\}$ on the same 1,000 images per dataset.

As the threshold rises, E-Cert-R's accuracy gap over the fixed-n plug-in increases: for CDF it grows from 1.0 points at $\tau = 1/2$ to 2.7 (CIFAR-10) and 3.2 (ImageNet) at $\tau = 0.8$. The staged plug-in falls further behind both, by 0.2–0.3 points behind the fixed-n bound at $\tau = 1/2$ and 0.5–0.6 at $\tau = 0.8$; behind E-Cert-R it is 1.3/1.2 points at $\tau = 1/2$ and 3.3/3.7 at $\tau = 0.8$ (CIFAR-10/ImageNet). The cost gap widens with it. The staged monitor's median stop over the images it certifies is 4.0–4.3× ours at $\tau = 1/2$, and 9.9–10.2× at $\tau = 0.8$ for CDF, because a harder certification problem pushes crossings past more of its checkpoints. For NP, all five arms remain within 0.6 points at every threshold. Absolute certified accuracy decreases for every procedure, from 61.1% to 40.5% on CIFAR-10 CDF: asking for a higher confidence certifies fewer images.

A threshold above one half has different meanings for the two observables. On the hard label $\tau = 0.8$ certifies that the *vote share* p_A, the fraction of noisy copies the base model assigns to the target, stays above 0.8; on the soft score it certifies that the model's own average confidence stays above 0.8, which is Kumar et al. (2020a)'s notion. Run on the same CIFAR-10 images, the hard-label arm certifies $66.4/62.1/57.2/51.4\%$ at the four thresholds against the best soft arm's $61.1/54.9/49.1/40.7\%$, so the ordering of Table 22 persists at every threshold and widens. At each tested threshold, certifying average softmax confidence is harder than certifying the corresponding vote share.

Table 19: The same benchmark on Cohen and consistency-regularized checkpoints at $\sigma = 0.25$, $r = 0.125$ and $n = 10^4$: certified accuracy and the median number of evaluations over the images each arm certifies. 1,000 images per dataset, streams in draw order. The fixed-n plug-in, a separately planned terminal analysis with no stopping time, is quoted in the text rather than tabulated here. Column groups differ only in the base classifier: the Cohen group is re-run through the same code path as the consistency group.

			Cohen et al.		Consistency	
Dataset	Cert.	Method	CA (%)	Median stop	CA (%)	Median stop
CIFAR-10	NP	Staged plug-in	56.0	100	57.9	100
		CS plug-in	55.9	39	58.0	37
		E-Cert-P	56.5	28	58.1	27
		E-Cert-R	56.4	23	58.1	23
	VC	Staged plug-in	59.6	250	63.9	250
		CS plug-in	60.6	65	65.1	63
		E-Cert-P	61.0	38	65.1	36
		E-Cert-R	61.0	27	65.2	27
	CDF	Staged plug-in	59.8	100	63.8	100
		CS plug-in	60.1	40	65.0	33
		E-Cert-P	59.9	34	64.5	33
		E-Cert-R	61.1	25	65.1	23
ImageNet	NP	Staged plug-in	49.9	100	52.7	100
		CS plug-in	50.1	39	52.8	36
		E-Cert-P	50.2	28	52.9	26
		E-Cert-R	50.2	24	52.8	22
	VC	Staged plug-in	54.8	250	56.7	250
		CS plug-in	55.9	66	57.5	57
		E-Cert-P	56.1	37	57.6	35
		E-Cert-R	56.1	27	57.3	26
	CDF	Staged plug-in	54.7	100	56.7	100
		CS plug-in	55.5	35	57.4	30
		E-Cert-P	55.1	34	56.9	32
		E-Cert-R	55.9	24	57.6	22

H.9 Fresh-noise replication

The ten seeds of Table 16 permute one cached score multiset per image, so their spread measures sensitivity to score *order* conditional on the draws. This subsection redraws the noise instead. With the paper's own sampler we drew three independent 10^4-draw caches per dataset on a fixed subset of 200 images, at noise seeds 2027, 2028 and 2029. The sampler takes every $\lfloor N/200 \rfloor$-th test image, so the subset sits inside the 1,000-image index set of the main tables. We evaluated every replicate with the same four arms, the same fixed-n plug-in and the same permutation seed 2026 as the published run. A fourth replicate is the paper's own cache restricted to those images; the generating script asserts that it reproduces all 4,800 published per-image stopping times exactly before summarizing anything. Table 21 reports the across-replicate mean and standard deviation of certified accuracy, of the median stop over the images each arm certifies, and of the mean over all 200 images with an uncertified image charged the 10,000 cap.

Monte Carlo noise is small at $n = 10^4$ and does not disturb the comparison. Certified accuracy moves by at most 1.0 percentage point between replicates, with standard deviation at most 0.5 pp and exactly zero in 17 of the 30 cells. That is roughly an order of magnitude below the ± 3.5 pp standard error of drawing 200 images from the test set, which is why we report intervals over images rather than across noise replicates. The certified-image medians move by at most 1.3 evaluations (3.4%) and the all-images means by at most 28 (0.5%). The ordering is stable. E-Cert-R has the lowest certified median in all 24 replicate–dataset–certificate cells, and the lowest all-images mean in 22; the two exceptions are ties with E-Cert-P within 1.3 evaluations. The small-budget advantage of Table 1 reproduces as well: at $n = 100$ E-Cert-R certifies 40.5–53.5% of images against the fixed-n

Table 20: The single-class arms at four certification thresholds, 1,000 images per dataset, $\sigma = 0.25$, $r = 0.125$, $\alpha = 10^{-3}$, certified accuracy (%) at $n = 10^4$. Only the null moves with τ; the generating script asserts that the $\tau = 0.5$ column reproduces the published per-image stopping times. Certificates are ordered VC, CDF, NP, the order in which the threshold matters.

Dataset	Cert.	Method	Certified accuracy (%) at $n = 10^4$			
			$\tau = 0.5$	$\tau = 0.6$	$\tau = 0.7$	$\tau = 0.8$
CIFAR-10	VC	Fixed-n plug-in	59.7	53.4	47.6	39.6
		Staged plug-in	59.6	53.3	47.4	38.8
		CS plug-in	60.6	54.7	48.7	40.6
		E-Cert-P	61.0	54.9	48.9	40.7
		E-Cert-R	61.0	54.9	49.1	40.7
	CDF	Fixed-n plug-in	60.1	53.5	47.2	37.8
		Staged plug-in	59.8	53.3	47.0	37.2
		CS plug-in	60.1	53.9	48.0	39.7
		E-Cert-P	59.9	53.1	47.7	40.1
		E-Cert-R	61.1	54.8	48.7	40.5
	NP	Fixed-n plug-in	56.3	50.9	44.0	36.2
		Staged plug-in	56.0	50.9	43.9	36.0
		CS plug-in	55.9	50.9	44.0	36.4
		E-Cert-P	56.5	50.9	44.2	36.5
		E-Cert-R	56.4	50.9	44.1	36.3
ImageNet	VC	Fixed-n plug-in	54.9	50.6	45.4	39.0
		Staged plug-in	54.8	50.3	45.3	38.8
		CS plug-in	55.9	51.3	45.8	40.4
		E-Cert-P	56.1	51.5	46.1	40.6
		E-Cert-R	56.1	51.5	46.1	40.6
	CDF	Fixed-n plug-in	54.9	50.1	44.8	37.1
		Staged plug-in	54.7	49.9	44.7	36.6
		CS plug-in	55.5	50.9	45.6	39.5
		E-Cert-P	55.1	51.3	45.7	40.0
		E-Cert-R	55.9	51.4	45.9	40.3
	NP	Fixed-n plug-in	50.0	45.9	41.1	33.2
		Staged plug-in	49.9	45.9	41.1	33.0
		CS plug-in	50.1	46.0	41.3	33.3
		E-Cert-P	50.2	46.1	41.4	33.3
		E-Cert-R	50.2	46.1	41.2	33.3

plug-in's 23.9–47.2%, with a replicate standard deviation of at most 1.2 pp. The one quantity with a visibly larger spread is the staged plug-in's all-images median, at standard deviation 62.5 and 125 on two CIFAR-10 cells. It can take values only on its ten-point grid, so one image moving across the median shifts it by a whole checkpoint, the quantization of § H.3.

These accuracies are computed on 200 images and therefore differ from the 1,000-image accuracies of Table 1; the comparison of interest is across replicates within this table, not against the main ones. Replication across base classifiers is § H.7; across training seeds of one method and across architectures it remains out of scope (§ J).

I MULTICLASS PAIRWISE-MARGIN CERTIFICATION

We give the pairwise-margin constructions, comparator definitions and multiclass experiments, followed by controls for working-law adaptation, error allocation, monitoring and target selection. Experiments use the same ground-truth-target, cached-stream protocol as the main text.

Table 21: Fresh-noise replication of the single-class certificates: four independent 10^4-draw replicates of the same 200 images per dataset (the paper's cache and three fresh noise seeds), each evaluated with the published arms at permutation seed 2026. Entries are across-replicate means $\pm$ standard deviations of certified accuracy, of the median stop over the images that arm certifies, and of the mean over all 200 images with an uncertified image charged the 10,000 cap. The fixed-n plug-in is a terminal analysis at $n = 10^4$ and has no stopping time ("–").

Dataset	Cert.	Method	CA (%)	Median (certified)	Mean (all images)
CIFAR-10	NP	Fixed-n plug-in	54.6 ± 0.3	–	–
		Staged plug-in	54.6 ± 0.3	100 ± 0	$4{,}706 \pm 4$
		CS plug-in	54.8 ± 0.3	38 ± 1	$4{,}583 \pm 18$
		E-Cert-P	54.9 ± 0.3	27 ± 0	$4{,}572 \pm 17$
		E-Cert-R	54.8 ± 0.3	23 ± 0	$4{,}569 \pm 18$
	VC	Fixed-n plug-in	56.5 ± 0.0	–	–
		Staged plug-in	56.5 ± 0.0	250 ± 0	$4{,}613 \pm 11$
		CS plug-in	56.7 ± 0.3	63 ± 1	$4{,}398 \pm 9$
		E-Cert-P	56.9 ± 0.3	36 ± 0	$4{,}367 \pm 10$
		E-Cert-R	57.1 ± 0.5	27 ± 0	$4{,}358 \pm 11$
	CDF	Fixed-n plug-in	56.6 ± 0.3	–	–
		Staged plug-in	56.5 ± 0.0	100 ± 0	$4{,}484 \pm 6$
		CS plug-in	56.5 ± 0.0	38 ± 1	$4{,}411 \pm 2$
		E-Cert-P	56.5 ± 0.0	33 ± 0	$4{,}383 \pm 4$
		E-Cert-R	57.2 ± 0.3	23 ± 0	$4{,}345 \pm 15$
ImageNet	NP	Fixed-n plug-in	43.0 ± 0.0	–	–
		Staged plug-in	43.0 ± 0.0	100 ± 0	$5{,}854 \pm 19$
		CS plug-in	43.0 ± 0.0	40 ± 0	$5{,}735 \pm 5$
		E-Cert-P	43.0 ± 0.0	29 ± 0	$5{,}726 \pm 4$
		E-Cert-R	43.0 ± 0.0	24 ± 0	$5{,}725 \pm 5$
	VC	Fixed-n plug-in	47.0 ± 0.0	–	–
		Staged plug-in	47.0 ± 0.0	250 ± 0	$5{,}581 \pm 1$
		CS plug-in	48.5 ± 0.0	68 ± 1	$5{,}257 \pm 10$
		E-Cert-P	48.5 ± 0.0	38 ± 0	$5{,}196 \pm 3$
		E-Cert-R	48.5 ± 0.0	28 ± 0	$5{,}189 \pm 7$
	CDF	Fixed-n plug-in	47.2 ± 0.3	–	–
		Staged plug-in	47.0 ± 0.0	100 ± 0	$5{,}480 \pm 17$
		CS plug-in	48.2 ± 0.3	37 ± 1	$5{,}311 \pm 14$
		E-Cert-P	47.4 ± 0.5	34 ± 0	$5{,}307 \pm 28$
		E-Cert-R	48.5 ± 0.0	25 ± 0	$5{,}187 \pm 4$

I.1 Single-class and multiclass certification targets

The single-class benchmark of § 7.1 tests one class at $\tau = 1/2$ using only its scalar score, the access model assumed in much of the smoothing literature. A smoothed probability above one half already implies the top-1 decision, because the remaining classes share less than one half. The single-class test therefore certifies the prediction while observing a single scalar per draw, either the indicator that the noisy prediction equals the target class (Cohen et al., 2019) or that class's softmax entry, and needs neither the identity of the runner-up nor a correction across competitors. This is the black-box regime in which pretrained classifiers are certified through their prediction interface: Salman et al. (2020) certify the Azure, Google, AWS and ClarifAI image-classification APIs this way, stating that their approach "applies to both the white-box and the black-box settings of the pretrained classifier". Voráček (2024), likewise, estimates the hard-label class probabilities.

We distinguish two model outputs. *Softmax* and *hardmax* name the *observable*: the probability vector the network returns, or that vector collapsed to the one-hot indicator of its argmax. A *hard label* is the $0/1$ scalar the hardmax observable yields once a target class is fixed, the indicator that the noisy prediction equals it, and its mean is the hard-label probability p_A. The distinction matters because the multiclass test on hardmax draws reads *class identities* rather than one hard label: it needs to know which class was predicted, not merely whether it was the target, and its per-pair margins take three

values rather than two. We therefore write "hardmax" for the observable, a cache or a column, and "hard label" only for that scalar and its mean.

Soft scores and hard labels are two different guarantees, not two estimates of one. On the hard label the single-class certificate tests $p_A = \Pr[\arg\max f(\boldsymbol{x}+\boldsymbol{\epsilon}) = A] > 1/2$: the majority vote survives the perturbation, which is Cohen et al. (2019)'s object. On the target's softmax score it tests $\mathbb{E}[f_A(\boldsymbol{x}+\boldsymbol{\epsilon})] > c$: the score-averaging classifier predicts A, and for $c > 1/2$ it certifies a *confidence level* as well as a label. That extra reach is why the soft certificate exists. Kumar et al. (2020a), whose CDF certificate we adopt, motivate it by observing that standard certificates "contain little information about how confident the user should be in the assigned label", and that the certified radii correlate very little with the base classifier's prediction scores. A guarantee at $c = 1/2$ alone therefore cannot separate an input the model is sure about from one it merely gets right. They accordingly report certified accuracy at a range of cutoffs c, and a margin variant alongside it. A hard label can be thresholded above one half too, but it then certifies a different quantity: that the *vote share* p_A stays above the threshold, not that the model's own score does (§ H.8 runs both).

Our experiments fix $c = \tau = 1/2$, where the soft certificate reduces to a label certificate for the score-averaging classifier and the two observables become directly comparable. Table 22 makes that comparison at the paper's operating point, alongside each certificate's *oracle ceiling*: the fraction of images whose certificate at the empirical law of all 10^4 draws clears τ. That is a plug-in reference, not a bound on what a valid test could reach, since the empirical certificate carries no coverage of its own. At this operating point the distinction is immaterial, because $\kappa_\rho = 0.53$ leaves the variance term little room to be understated; § B.2 shows that it is not immaterial at $\rho = 3$, where $\kappa_\rho = 90$. Three readings follow. First, the soft score is where our two new certificates live. On a $0/1$ stream the variance and the CDF are functions of the mean, so VC and CDF carry nothing the Bernoulli mean does not already carry and are not tabulated separately there; they remain computable, and can certify — at $p = 0.9$, $\rho = 0.5$ the VC value is $0.74 > \tau$. On soft scores they lift the ceiling by 4.4–7.0 points over soft NP. That ordering is Kumar et al. (2020a)'s own: their CDF certificate beats their mean-only baseline, and their appendix reports that in most cases it does so even when the baseline is granted the best possible bound on the mean. Second, at $c = 1/2$ the hard-label target is nevertheless the easier one on these checkpoints, ahead of the best soft certificate by 4.6–5.3 points. The base classifiers are usually right after the noise while their softmax is far from saturated: the median p_A is 0.935 against a median $\mathbb{E}[f_A]$ of 0.793 on Cohen et al. (2019)'s checkpoint, and $p_A > \mathbb{E}[f_A]$ on 83% of images. Third, that gap belongs to the target and not to the procedure. Every arm certifies within 1.0 point of its own ceiling, and on the hard label the published KT betting plug-in reaches the same ceiling as our e-process, within 0.2 points of it on every checkpoint. Nothing in the comparison turns on which sequential procedure is used. The fixed-n reference point says the same thing from the other side. Cohen et al. (2019)'s own rule is a one-sided Clopper–Pearson lower bound on p_A at the full budget, which is CERTIFY's estimation step with the target class prespecified as it is here. It certifies $66.8/69.9\%$, within $0.4/0.3$ points of the ceiling and so 0.1–0.4 points *above* both sequential arms. That ordering is expected, and it is the trade this paper is about. At the full budget a fixed-n binomial bound on a $0/1$ stream is close to optimal, and an anytime-valid procedure pays a little for the right to stop at any time. It recovers that many times over in evaluations: on the same streams our arm's median stop is 20 against the staged Clopper–Pearson rule's 25 (Figure 7). What the hard-label column does not admit is a VC or CDF certificate carrying information beyond the mean, which is the reason the paper reports the soft score as well.

Consistency training illustrates the distinction without changing the test: its median softmax score is 0.816 and median p_A is 0.993, compared with 0.793 and 0.935 for Cohen's checkpoint. It improves both targets, but the hard-label certificate still reaches 69.8% certified accuracy against at most 65.2% for the soft-score certificates. Thus a model comparison should specify whether it certifies vote probabilities or averaged soft scores; Table 19 evaluates the latter.

The available output and desired guarantee determine the certification target. Hard labels certify more images at $c = 1/2$ on these checkpoints; soft scores support confidence thresholds $c > 1/2$ and allow VC/CDF to use distributional information beyond the mean. The same bounded-witness, predictable-stake and futility constructions apply to soft scores (§ 7.1), hard labels (Table 22 and the single-class control of Table 23), and the pairwise margins below. We compare procedures within each target.

Table 22: Single-class certification on the same 1,000 CIFAR-10 images, $\sigma = 0.25$, $r = 0.125$, $\alpha = 10^{-3}$, from the target's softmax score and from the hard label. Certified accuracy (%) at $n = 10^4$ of E-Cert-R and, on the hard label, of two references beside it: Cohen et al. (2019)'s fixed-n Clopper–Pearson bound on p_A (CERTIFY's estimation step, the target class being prespecified here) and the published KT betting plug-in (§ I.5). Each certificate's oracle ceiling, the same certificate evaluated at the empirical law of the full stream, is beneath; it is a plug-in reference rather than a bound. VC and CDF carry no information beyond the mean on a $0/1$ stream and are therefore reported for the soft score only, which is also why the KT sequence has no soft-score column: it is a coin-betting estimator. The two population quantities are $\mathbb{E}[f_A]$, the mean softmax score, and p_A, the probability that the noisy prediction is the target class. The hard-label column is a fresh cache and the generic RIPr loop; the single-class control quoted below (66.2%) is the same test on the multiclass benchmark's own draws with the Bernoulli-specialised factors of § 5, which is why the two differ by two images.

Base classifier	Statistic	$\mathbb{E}[f_A]$	p_A	Soft score			Hard label, NP		
		(median)	(median)	NP	VC	CDF	Cohen CP	CS plug-in (KT)	E-Cert-R
Cohen et al.	Certified	0.793	0.935	56.4	61.0	61.1	66.8	66.2	66.4
	Oracle ceiling			57.2	61.6	61.6		67.2	
Consistency	Certified	0.816	0.993	58.1	65.2	65.1	69.9	69.8	69.8
	Oracle ceiling			58.5	65.5	65.5		70.2	

The argmax procedures of this appendix need strictly more: the realized top competitor at every draw, or the whole score vector. Where that is available they certify more, because beating the *realized* competitor is a weaker requirement than beating one half. At the same σ and radius E-Cert-R on the multiclass target certifies 68.7% of CIFAR-10 and 64.3% of ImageNet images, against the single-class control's 66.2% and 63.3% (Table 23 and Figure 11; the summary table carries the multiclass arms only). These settings use the same sequential-testing construction but require different model outputs and certify different claims.

I.2 PAIRWISE-MARGIN DETAILS

To compare all three certificate families under one multiclass objective, let $\mathbf{X}_t = (X_{t,1}, \ldots, X_{t,C}) \in \Delta_C$ be either the full softmax vector or the hardmax one-hot vector, and fix the target class A before sampling. For each competitor $B \neq A$, define the raw and normalized pairwise margins

$$\begin{aligned} D_t^{(B)} &= X_{t,A} - X_{t,B} \in [-1, 1], \\ Z_t^{(B)} &= \frac{1 + D_t^{(B)}}{2} \in [0, 1], \\ M_{A,B} &= \mathbb{E}[Z_t^{(B)}]. \end{aligned} \tag{13}$$

The affine normalization is needed only because the scalar NP, VC, and CDF results are stated for scores in $[0, 1]$. It maps a tie, $D_t^{(B)} = 0$, to $Z_t^{(B)} = 1/2$; it does *not* require the target–competitor score gap to exceed $1/2$.

To state the robust semantics explicitly, let $\mu_k(\delta) = \mathbb{E}_\epsilon[f_k(\boldsymbol{x} + \delta + \epsilon)]$ for $\epsilon \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$, and write $\rho = r/\sigma$ as in § 4. A valid scalar certificate applied to the unperturbed law $P_{A,B}$ of $Z_t^{(B)}$ satisfies

$$\mathrm{Cert}_j(P_{A,B}; \rho) \leq \inf_{\|\delta\|_2 \leq r} \frac{1 + \mu_A(\delta) - \mu_B(\delta)}{2}. \tag{14}$$

Therefore, $\mathrm{Cert}_j(P_{A,B}; \rho) > 1/2$ implies $\mu_A(\delta) > \mu_B(\delta)$ for every admissible δ. Requiring this for every competitor gives the sufficient multiclass condition

$$\min_{B \neq A} \mathrm{Cert}_j(P_{A,B}; \rho) > \frac{1}{2}, \qquad j \in \{\mathrm{NP}, \mathrm{VC}, \mathrm{CDF}\}. \tag{15}$$

For certificate family j, the associated global bad null is the union

$$H_{0,j}^{\mathrm{margin}}(\rho) = \bigcup_{B \neq A} \left\{ \mathrm{Cert}_j(P_{A,B}; \rho) \leq \frac{1}{2} \right\} = \bigcup_{B \neq A} \mathcal{P}_{0,j}(\rho, \tfrac{1}{2}), \qquad \mathcal{P}_{0,j}(\rho, \tfrac{1}{2}) := H_j(\rho, \tfrac{1}{2}), \tag{16}$$

reusing the marginal bad-null family $H_j(\rho,\tau) = \{Q : \mathrm{Cert}_j(Q;\rho) \le \tau\}$ of § 4: $\mathcal{P}_{0,j}(\rho, 1/2)$ is exactly the set of laws for $Z_t^{(B)}$ under which certificate j fails to certify that the target outranks B, independent of which competitor B is being tested. This is the set the "RIPr-guided pairwise e-process" paragraph and Prop. 8 below refer to. Let $\mathcal{K}_{t,j}^{(B)}$ be an e-process for the Bth null in (16), and define $\overline{\mathcal{K}}_{t,j}^{(B)} = \max_{s\le t} \mathcal{K}_{s,j}^{(B)}$. The global stopping rule is

$$\tau_{\mathrm{multi},j} = \inf\left\{ t : \min_{B\neq A} \overline{\mathcal{K}}_{t,j}^{(B)} \ge 1/\alpha \right\}. \tag{17}$$

This is an intersection–union decision: all $C-1$ pairwise bad nulls must be rejected, rather than averaged.

Prop. 4 in the main text records the resulting guarantee for this stopping rule, with its proof. One score vector updates every pair, so one round still costs one noisy model evaluation; only the arithmetic cost grows with C.

Throughout the remaining theory discussion, $V_{A,B} = \mathrm{Var}(Z_t^{(B)})$. For NP, (15) has a particularly transparent form. Since $M_{A,B} = (1+\mu_A-\mu_B)/2$ and, by (1), $\mathrm{Cert}_{\mathrm{NP}}(P_{A,B};\rho) > 1/2 \iff M_{A,B} > T_\rho(1/2) = \Phi(\rho)$,

$$\mathrm{Cert}_{\mathrm{NP}}(P_{A,B};\rho) > \frac{1}{2} \quad\Longleftrightarrow\quad \mu_A - \mu_B > 2\Phi(\rho) - 1. \tag{18}$$

This is exactly the radius-safe difference condition. It is sufficient for the curved exact condition $\Phi^{-1}(\mu_A) - \Phi^{-1}(\mu_B) > 2\rho$, but generally more conservative than it. Along the linear boundary $\mu_A-\mu_B = 2\Phi(\rho)-1$ the two conditions agree only at the symmetric point $\mu_A = \Phi(\rho)$, $\mu_B = \Phi(-\rho)$, where both sides equal exactly 2ρ. Among points on that boundary, the curved left-hand side is minimized at the symmetric point. Indeed, $(\Phi^{-1})'(u) = 1/\phi(\Phi^{-1}(u))$ is symmetric about $1/2$ and increases with distance from $1/2$, so its integral over an interval of fixed length is smallest when centered at $1/2$. Here ϕ is the standard normal density; Φ^{-1} is concave below $1/2$ and convex above it. VC instead uses

$$M_{A,B} - \kappa_\rho \sqrt{V_{A,B}} > \frac{1}{2}, \tag{19}$$

where $V_{A,B} = \mathrm{Var}(Z_t^{(B)})$, and CDF substitutes the CDF of $Z_t^{(B)}$ into (3). Both can exploit distributional information discarded by the mean-only condition.

Exact two-mean RIPr comparator. For comparison, the exact two-mean Gaussian condition tests the curved null

$$\begin{aligned} H_{0,B}^{\mathrm{exact}}(r): \quad & \mu_A \le T_{2\rho}(\mu_B), \\ & T_g(u) = \Phi\{\Phi^{-1}(u) + g\}. \end{aligned} \tag{20}$$

Predictably reverse-projecting a working law onto this null gives the exact two-mean RIPr e-process, the E-Cert-R (no refit) and E-Cert-R hardmax arms of § I.5, whose validity for this null is proved in § I.3. It uses the same all-pairs stopping rule (17), but its curved boundary is less conservative than the NP sufficient condition (18). On hardmax identities the pair law is two means, so this is the tightest two-class condition. The margin NP, VC and CDF nulls are what the softmax arms of Table 27 test instead. The exact two-mean null and the margin NP null define different hypotheses; the former is not an RIPr-optimized version of the latter.

The hardmax interpretation requires one distinction. A binary target indicator $\mathbf{1}\{Y_t = A\}$ is Bernoulli, so its variance and CDF are determined by one mean and VC/CDF add no independent information. The full pairwise hardmax margin in (13), however, takes values $0, 1/2, 1$ with probabilities $p_B, 1-p_A-p_B, p_A$. That is two free parameters, of which the mean pins only $p_A - p_B$, so the second is visible to VC and CDF but not to NP. VC can therefore certify strictly more than NP here in principle: at $\rho = 1/2$, for instance, $(p_A, p_B) = (0.30, 0)$ gives $M_{A,B} = 0.65 < \Phi(\rho) = 0.692$, so (18) fails, while the concentrated three-point law has $V_{A,B} = 0.053$ and clears (19). That regime is a moderate margin with most mass on the tie value $1/2$, meaning draws where neither A nor B wins. It is not where real smoothed classifiers put their binding competitor. We therefore exercise VC and CDF only on continuous softmax margins (§ I.10), where they can use distributional information beyond the mean, and use the hardmax margin only for NP.

Matched bottleneck design. For every pair, let $\widehat{P}_{t-1}^{(B)}$ be the 100-bin working law formed from completed margins, including a fixed Dirichlet pseudocount. At a predictable refit time, certificate j selects

$$\widehat{B}_{t,j} \in \arg\min_{B \neq A} \widehat{\mathrm{Cert}}_{t-1,j}^{(B)}(\rho), \qquad \widehat{Q}_{t,j} = \widehat{P}_{t-1}^{(\widehat{B}_{t,j})}, \tag{21}$$

with deterministic tie breaking. Thus $\widehat{B}_{t,j}$ and $\widehat{Q}_{t,j}$ are $\mathcal{F}_{t-1}$-measurable. The bottleneck law is used only to design the next factor: that same certificate-valid factor is applied to every unresolved pair. This shared design avoids solving up to $C-1$ numerical projections per refit and, more importantly, makes E-Cert-K and E-Cert-R use exactly the same working alternative. Both compute their design moments from the histogram bin centers, while capital is updated on the original continuous margins. Their stake caps remain different as defined below, so this comparison evaluates two complete designs; the equal-cap single-class control is reported in § F.1.

E-Cert-K: the quadratic-Kelly pairwise e-process. E-Cert-K uses the direct predictable construction already available for each scalar certificate. NP uses $U_{t,\mathrm{NP}}^{\mathrm{ord}}(z) = z - \Phi(\rho)$. VC places the boundary-normal witness in (4) at the plug-in mean of $\widehat{Q}_{t,\mathrm{VC}}$, clipped only to the feasible VC boundary. CDF linearizes its concave CDF functional at $\widehat{Q}_{t,\mathrm{CDF}}$ using the same tangent construction as (5). VC and CDF use a neutral factor whenever the working certificate does not exceed $1/2$. If $u_{\min,t}^{j} = \inf_{z\in[0,1]} U_{t,j}^{\mathrm{ord}}(z) < 0$, the quadratic-Kelly stake is the past-measurable half-Kelly rule

$$\eta_{t,j}^{\mathrm{ord}} = \left[\frac{\mathbb{E}_{\widehat{Q}_{t,j}}[U_{t,j}^{\mathrm{ord}}]}{\mathbb{E}_{\widehat{Q}_{t,j}}[(U_{t,j}^{\mathrm{ord}})^2]}\right]_+ \wedge \frac{1/2}{-u_{\min,t}^{j}}. \tag{22}$$

The resulting factor $e_{t,j}^{\mathrm{ord}}(z) = 1 + \eta_{t,j}^{\mathrm{ord}} U_{t,j}^{\mathrm{ord}}(z)$ is at least $1/2$ over $[0,1]$. The stake here is approximated rather than solved exactly, but the arm is still a predictable e-process, not a fixed-n plug-in confidence bound.

RIPr-guided pairwise e-process. RIPr keeps the same pairwise bad set $\mathcal{P}_{0,j}(\rho, 1/2)$ and the same working law $\widehat{Q}_{t,j}$, but replaces the plug-in factor by the log-growth-optimal factor described in § 5. Concretely, NP keeps $z - \Phi(\rho)$ and replaces the half-Kelly stake by the exact Kelly-optimal stake for that linear witness (§ 5). VC jointly selects its feasible boundary-normal witness and stake by

$$\begin{aligned} G_{t,\mathrm{VC}}(m,\eta) &:= \mathbb{E}_{\widehat{Q}_{t,\mathrm{VC}}} \log[1 + \eta\{D_m - Y_m(Z)\}], \\ (m_t^\star, \eta_t^\star) &\in \arg\max_{m,\eta} \; G_{t,\mathrm{VC}}(m,\eta), \\ &\text{over } m \in \mathcal{M}_{\mathrm{VC}}, \quad \eta \in \mathcal{A}(m). \end{aligned} \tag{23}$$

where $\mathcal{M}_{\mathrm{VC}}$ is the feasible VC boundary and $\mathcal{A}(m)$ enforces nonnegativity on $[0,1]$. CDF solves the same finite-bin reverse-KL projection as § B for $\widehat{Q}_{t,\mathrm{CDF}}$, uses the supporting tangent at the projected bad law rather than at the working law, and chooses the exact log-optimal stake. This is the design of § 5 applied to the pairwise margins, so we call it E-Cert-R here too; relative to E-Cert-K it changes only the stake for NP, and both the witness and the stake for VC and CDF.

Proposition 8 (Validity of shared bottleneck RIPr games)**.** *For $v \in \{\mathrm{ord}, \mathrm{RIPr}\}$, suppose the factor designed from $\widehat{Q}_{t,j}$ satisfies*

$$\sup_{P \in \mathcal{P}_{0,j}(\rho,1/2)} \mathbb{E}_P[e_{t,j}^{v}(Z) \mid \mathcal{F}_{t-1}] \le 1. \tag{24}$$

Then, for every competitor B, $\mathcal{K}_{t,j,v}^{(B)} = \prod_{s=1}^{t} e_{s,j}^{v}(Z_s^{(B)})$ is an e-process for its pairwise bad null, even though the factor was designed from the selected pair $\widehat{B}_{t,j}$. Applying (17) to either variant is therefore anytime-valid at level α without splitting α across classes.

Proof. Selection in (21) uses only completed scores, so $e_{t,j}^{v}$ is predictable. The supremum in (24) is over the entire pairwise bad set and therefore also bounds the conditional expectation under the law of any particular null pair B. Iterated conditioning proves the e-process claim, and Prop. 4 proves the global statement. □

Between predictable refits the chosen factors are frozen, while every new score vector updates all unresolved games. A failed numerical projection is replaced by the neutral factor one; this can reduce power but cannot invalidate the process.

The matched construction also clarifies what RIPr does and does not guarantee. For NP the two designs share the witness. The RIPr stake maximizes the working-law expected log-growth over all feasible stakes, of which the quadratic-Kelly stake is one, so at a refit

$$\mathbb{E}_{\widehat{Q}_{t,j}}[\log e_{t,j}^{\mathrm{RIPr}}(Z)] \geq \mathbb{E}_{\widehat{Q}_{t,j}}[\log e_{t,j}^{\mathrm{ord}}(Z)]. \tag{25}$$

For VC the same holds whenever the one-dimensional search over the boundary point reaches its maximum. For CDF, whose matched witness is anchored at the projected law rather than at the working law, the inequality is what the finite-alphabet log-optimality of the reverse projection delivers when the projection is exact. In the implementation both are numerical, and we do not rely on (25) anywhere. Even where it holds it is expected log-growth dominance under the fitted bottleneck law, not pointwise factor dominance, sample-path capital dominance, or a guarantee of earlier stopping under the unknown true law. Early histogram misspecification can make the exact Kelly factor more aggressive than the half-Kelly baseline, and a different competitor can become the true stopping bottleneck between refits. The paired stopping experiment must therefore measure, rather than assume, an RIPr advantage.

I.3 Projection and baseline details

This section gives the projection and baseline details used by § I.2 and implemented directly by the E-Cert-R (no refit) and E-Cert-R arms of § I.5, which both call the same projection and differ only in how often it is recomputed. Let $Y_t \in \{1, \ldots, C\}$ be the hardmax class identity from the tth Gaussian perturbation, let $p_k = \Pr(Y_t = k)$, and fix target class A before observing the noisy stream. Write $g = 2r/\sigma = 2\rho$ and

$$T_g(u) = \Phi\big(\Phi^{-1}(u) + g\big)\,. \tag{26}$$

The pairwise radius null for competitor B is

$$\mathcal{P}_{0,B}(r) = \{p \in \Delta_C : p_A \leq T_g(p_B)\}\,. \tag{27}$$

The derivative calculation in § A.4, with g in place of ρ, shows that T_g is increasing and concave for $g \geq 0$. Hence (27) is the hypograph of a concave function intersected with the simplex and is therefore convex.

I.3.1 Three-coordinate reverse projection

Fix a full-support working law $q \in \Delta_C$. For one competitor B, only the three aggregate coordinates

$$q_A, \qquad q_B, \qquad q_O = 1 - q_A - q_B$$

enter the projection; all classes outside $\{A, B\}$ retain their relative proportions. If $q_A \leq T_g(q_B)$, then q already lies in the pairwise bad set and we use the neutral factor one. Otherwise the active projection lies on $p_A = T_g(p_B)$. Parameterizing $p_B = b$ gives

$$p_A = T_g(b), \qquad p_O = 1 - T_g(b) - b, \qquad 0 < b < \Phi(-g/2), \tag{28}$$

where the upper endpoint is the solution of $T_g(b) + b = 1$. The scalar reverse projection minimizes

$$\begin{aligned} D_B(b) = q_A \log \frac{q_A}{T_g(b)} + q_B \log \frac{q_B}{b} \\ + q_O \log \frac{q_O}{1 - T_g(b) - b}. \end{aligned} \tag{29}$$

Using

$$T_g'(b) = \exp\left\{-g\Phi^{-1}(b) - \frac{g^2}{2}\right\}, \tag{30}$$

its first-order condition is

$$-\frac{q_A T_g'(b)}{T_g(b)} - \frac{q_B}{b} + \frac{q_O\{T_g'(b) + 1\}}{1 - T_g(b) - b} = 0. \tag{31}$$

The implementation solves this one-dimensional first-order condition by bisection and checks positivity, the active boundary, and the null expectation numerically. If $p_B^\star$ is the solution and $p_A^\star = T_g(p_B^\star)$, $p_O^\star = 1 - p_A^\star - p_B^\star$, the one-step factor is

$$e_t^{(B)}(y) = \begin{cases} q_A/p_A^\star, & y = A, \\ q_B/p_B^\star, & y = B, \\ q_O/p_O^\star, & y \notin \{A, B\}. \end{cases} \tag{32}$$

For completeness, reverse-projection optimality yields the same supporting inequality as the finite-alphabet first-order condition in § B: for every $p \in \mathcal{P}_{0,B}(r)$,

$$\mathbb{E}_p[e_t^{(B)}(Y_t) \mid \mathcal{F}_{t-1}] = \sum_{k=1}^{C} p_k \frac{q_k}{p_k^\star} \le 1. \tag{33}$$

Thus predictable choices of q give a pairwise e-process. One observed class identity determines (32) for every $B \neq A$, so all $C-1$ capitals are updated from the same model evaluation. The same factor extends to a softmax vector $x \in \Delta_C$ by affine interpolation,

$$e_t^{(B)}(x) = e_{t,O}^{(B)} + x_A\{e_{t,A}^{(B)} - e_{t,O}^{(B)}\} + x_B\{e_{t,B}^{(B)} - e_{t,O}^{(B)}\}. \tag{34}$$

Its expectation depends only on the mean softmax coordinates (μ_A, μ_B) and obeys the same supporting inequality over $\mu_A \le T_g(\mu_B)$. Hence the exact two-mean RIPr reference retained below is also an e-process for its curved two-mean null in the softmax setting.

I.3.2 Intersection–union validity

The global nonrobust null is $\mathcal{P}_0^{\text{multi}}(r) = \bigcup_{B \neq A} \mathcal{P}_{0,B}(r)$, and each per-pair process constructed above is an e-process for its own pairwise null, so Prop. 4 applies verbatim: rejecting when $\min_{B \neq A} \max_{s \le t} \mathcal{K}_s^{(B)} \ge 1/\alpha$ is level α without dividing by $C-1$. Averaging pairwise e-values would not implement this decision: the average can be large while a genuine runner-up null has not been rejected.

I.4 Plug-in counterparts of the min-test

Classwise bounds and intersection–union testing. The Bonferroni plug-in bounds the target from below at $\alpha/2$ and the *top observed* competitor from above at $\alpha/(2(C-1))$, the division paying for the union over which class is the maximum. Cast as an intersection–union test the division is unnecessary. Test each pairwise null $H_{0,B}$ with the same two Clopper–Pearson bounds at $\alpha/2$ each, and reject the union null only when every $H_{0,B}$ is rejected. Under the union null some $H_{0,B^\star}$ is true, so global rejection entails rejecting $B^\star$, which has probability at most α by the argument of Prop. 4, which nowhere uses that the component tests are e-processes (Delattre et al., 2025). The pairwise bound is monotone in the competitor's count, so rejecting every pair is the same as rejecting the top observed competitor. The min-test form of the classwise certificate is therefore the Bonferroni form with $\alpha/2$ in place of $\alpha/(2(C-1))$. E-processes additionally provide anytime-valid components, continuous monitoring, and direct certificate witnesses. Every table that reports the hardmax plug-in reports both forms, and the min-test form is the one the figures draw, since it is the stronger baseline. They differ by under 15 mean evaluations and 0.1 pp, with the median at the same 50-draw checkpoint, because the ten-point grid rounds both stopping times to the same checkpoint. At its best tested grid, the min-test form reaches 32 evaluations against the Bonferroni form's 36/41 (§ I.13). Continuous monitoring exposes the cost of this split more directly (§ I.9).

I.5 Real-data benchmark: Bonferroni plug-in, class partitioning, and E-Cert-R

Prop. 4 establishes that the min-test (17) is level-α under arbitrary dependence, with no multiplicity correction across the $C-1$ competitors. Prop. 8 establishes validity for the refit design, but says nothing about its speed relative to a E-Cert-R (no refit) design: the theory offers only working-law log-growth statements, not guarantees under the true law. Four original multiclass procedures run on the same real CIFAR-10 and ImageNet noisy-class-identity streams. Two are classical multiplicity corrections: the *Bonferroni plug-in* (Bonferroni-corrected classwise Clopper–Pearson), and the

class partitioning method *CPM* of Delattre et al. (2025). Two are instances of the categorical RIPr construction of § I.3 differing only in whether the working law is refit from data, namely *E-Cert-R (no refit)* and E-Cert-R (refit online; ours).

Baselines. The *Bonferroni plug-in* is classwise Clopper–Pearson: a one-sided lower bound on the target class at error $\alpha/2$, and a one-sided upper bound on the empirically largest competitor at error $\alpha/(2(C-1))$. This is the standard classwise construction, and it is worth being precise about how the three plug-in forms are related, because they are not three versions of one statement. Cohen CERTIFY bounds p_A from below and nothing else, resting on $p_B \le 1 - p_A$, and certifies when $\sigma\Phi^{-1}(\underline{p}_A) \ge r$: a *single-class* claim, $p_A > \Phi(\rho)$, which is sufficient for the argmax but stronger than it needs. The Bonferroni plug-in moves that bound to the pairwise claim – that A beats each B – by bounding every competitor explicitly, and pays $\alpha/(C-1)$ for the union over which one is the maximum. The *min-test plug-in* keeps that claim and drops the division, which § I.4 shows is unnecessary once the decision is read as an intersection–union test. Cohen's arm therefore belongs with the single-class hard-label results of Table 22, not alongside the pairwise arms. One upper-bound evaluation on whichever class is realized as the maximum stands in validly for all $C-1$ competitors, because each competitor's bound is valid at that level whichever one turns out to be the maximum. *CPM* (Delattre et al., 2025) instead reduces the number of corrected comparisons below C. An initial phase of $n_0 = 100$ samples fixes a bucket partition via Algorithm 1's greedy promotion rule, giving the empirically top class, the runner-up, zero or more promoted classes, and a meta-class for the rest, then Clopper–Pearson bounds at level $\alpha/c^\star$ are computed on the remaining samples. That n_0 matches Delattre et al.'s own Section 4 experiments, which fix $\alpha = 10^{-3}$, $n = 10^4$ and $n_0 = 100$. The initial phase is drawn from, and excluded from, the same evaluation stream every other method reads, so CPM is charged for it. Its staged checkpoints are corrected over the checkpoints at which it can decide: a checkpoint at or below n_0 leaves its estimation set empty, so its rejection region there is empty and contributes nothing to the union bound. At $n_0 = 100$ that is the six checkpoints beyond 100, so each bound is read at level $\alpha/(6c^\star)$. The Bonferroni plug-in needs no initial phase and is corrected over all ten. Both baselines are checked at every budget in the sweep, each with its own count of decidable checkpoints folded into its per-test level. That is the staged/anytime-monitoring convention used throughout this paper. Their reported stopping cost is therefore rounded up to the next checkpoint at or after the true crossing, not the exact crossing count. *E-Cert-R (no refit)* and E-Cert-R are both instances of the online, past-only categorical RIPr e-process developed in § I.3: a genuine e-process needs no Bonferroni factor for staged monitoring, so both are checked at their exact first-crossing evaluation count, with no checkpoint rounding. They differ only in the working law that builds the per-round factor. *E-Cert-R (no refit)* fixes it once, at the start, to the clean (noiseless) softmax vector, and never revisits it. E-Cert-R refits it at a doubling schedule from a predictable blend of that same clean vector and the running empirical class counts. The Bonferroni plug-in and CPM never see the clean vector at all; among these four arms only E-Cert-R (no refit) and E-Cert-R do, so any gap between *those two* isolates the effect of refitting specifically, with access to side information held fixed.

A fifth arm, the *CS plug-in*, is the multiclass form of the strongest single-class comparator of § 7.1. Betting confidence sequences with the predictable empirical-Bernstein stakes of § D.2 replace the two classwise Clopper–Pearson bounds: a lower sequence on p_A and an upper sequence on each competitor's p_B, at level $\alpha/2$ each. These are combined by intersection–union across competitors, so no division by $C-1$, and inspected after every draw, so no checkpoint correction. The image is certified at the first round at which $\Phi^{-1}(L_A) - \Phi^{-1}(\max_B U_B) \ge 2\rho$. Competitors never observed share one upper sequence, and the first round is located exactly by bisection, since both sequences are running intersections. It is reported in Table 23 and Table 24, and drawn in Figure 8. It certifies 68.1% and 63.9%, that is 0.6 and 0.4 pp below E-Cert-R, with a common-success median of 38 on both datasets against E-Cert-R's 19/20. It is earlier than E-Cert-R on 115 and 72 of the jointly certified images, against 559 and 563 the other way.

A sixth arm, the *CS plug-in (KT)*, is the same construction with the published estimator for this stream. A hard-label indicator is a coin, and for coins Voráček (2024) stakes the Krichevsky–Trofimov running mean $\hat{q}_t = (H_{t-1} + \frac{1}{2})/t$, the Kelly bet against that estimate, in the fair game that pays $1/p$ on heads and $1/(1-p)$ on tails. The stake does not depend on the candidate p, so one play prices every candidate at once. The log-wealth $\log W_t(p) = \log \mathrm{KT}(H_t, t) - H_t \log p - (t - H_t)\log(1-p)$ is convex in p, so the unrejected set is an interval and its running intersection is a confidence sequence (their Algorithm 2). We use it exactly as above: a lower sequence on p_A and an upper sequence

on each competitor at $\alpha/2$ each, intersection–union over competitors, the same certification event. The estimator allows one simplification: both endpoints come from the one wealth process, so a competitor's upper bound needs no complement stream. The Krichevsky–Trofimov mixture has the closed form $\mathrm{KT}(h,t) = B(h+\frac{1}{2}, t-h+\frac{1}{2})/B(\frac{1}{2},\frac{1}{2})$, so the whole sequence is a vector operation on the running counts; the arm costs 0.06/0.07 s per image, against the empirical-Bernstein arm's repeated bisections. It certifies 68.6% and 64.3%: 0.5 and 0.4 pp *above* the empirical-Bernstein arm, and 0.1 and 0.0 pp from E-Cert-R. Its common-success median is 27 on both datasets, and it stops earlier than the empirical-Bernstein arm on 554 and 583 of the jointly certified images, against 125 and 54 the other way. Against E-Cert-R the order reverses: it is later on 583 and 533 and earlier on 96 and 90. The two plug-ins differ in their stake construction; E-Cert-R uses a refitted working law to construct a direct test of the certificate. This arm is the one the tables and figures call CS plug-in (KT): the estimator is Voracek's, the certification rule is the intersection–union one this paper uses for every multiclass arm. Where the two need separating we write CS plug-in (KT, IUT) for it and CS plug-in (KT, OvR) for the rule of the next paragraph.

The comparison Voracek himself reports pairs the same estimator with a different certification rule: the *one-vs-rest* bound of Cohen et al. (2019), which bounds p_B by $1-p_A$ and so needs a single lower sequence at the full α: certify once $\sigma\Phi^{-1}(L_A(t)) \geq r$. We evaluate this published rule on the same streams, separately from our intersection–union adaptation. Not splitting α makes it faster on the images both rules certify (median 25 against 27, earlier on 544 and 590 of them). Bounding every competitor by $1-p_A$ makes it certify 66.2% and 63.3%, that is 2.4 and 1.0 pp fewer images. At $\tau = 1/2$ this rule is also exactly the single-class NP certificate on a 0/1 stream, which is how it enters the hard-label column of Table 22.

Two properties follow from the baseline constructions. First, CPM's initial n_0-sample phase identifies the "top class" purely empirically, with no knowledge of the true target. If that identification is wrong, CPM's certificate is for the wrong class, and the image can never certify under CPM at any budget. The Bonferroni plug-in and the two e-processes never re-identify the target class from data, so none of them has this failure mode. Synthetic examples with narrow margins exhibit this failure mode; its prevalence on real images is assessed below. Second, $n_0 = 100$ lands exactly on a checkpoint in the budget grid, so CPM's estimation phase is empty, and its certificate unattainable, at every checkpoint up to and including $n = 100$ itself. Its earliest possible certification occurs at $n = 250$, the next checkpoint, regardless of how easy the image is.

Setup. Both datasets use the same 1,000-image, $\sigma = 0.25$, $\rho = 0.5$ ($r = 0.125$) cached noisy-class-identity streams as the paper's other real-data experiments, matching the image count of Table 1. Each image contributes 10,000 hardmax class identities from a smoothed ResNet-110 (CIFAR-10, $C = 10$) or ResNet-50 (ImageNet, $C = 1{,}000$), plus the clean (noiseless) softmax vector. The target class is the ground-truth label, fixed before sampling. Budgets sweep $n \in \{10, 25, 50, 100, 250, 500, 1{,}000, 2{,}500, 5{,}000, 10{,}000\}$, matching the checkpoint grid used throughout this paper. $\alpha = 10^{-3}$ throughout. A numerical audit of every image's E-Cert-R (no refit) and E-Cert-R refits runs automatically, comparing the realized factors against their supporting inequality (33) on a 1,001-point grid, at the eight competitors whose factors deviate most from 1. It raises an error if the excess exceeds 2×10^{-5}. Across the 2,000 image records (both datasets), each audited under both variants, the worst excess was 6.4×10^{-15} (online) and exactly 0 (E-Cert-R (no refit), which projects only once per image). This is a numerical check on representative competitors in the hardmax benchmark, not a substitute for the analytic validity argument and not a sweep of every run in the paper. For every multiclass experiment, the target is the ground-truth label $A_i = y_i$, fixed before sampling. If the all-pairs procedure succeeds for image i, it certifies that A_i remains the robust winner, and hence certifies a correct robust classification. For any prespecified subset S of images, we therefore report

$$\mathrm{CA}_S(n) = \frac{1}{|S|}\sum_{i\in S} \mathbf{1}\{\tau_i \leq n\}, \tag{35}$$

where τ_i is the method's all-pairs stopping time, and S is the full set of 1,000 images. The level $\alpha = 10^{-3}$ applies to each prespecified input–certificate decision, not simultaneously to the 1,000 images. We report CA with 95% Wilson score intervals and median stopping costs with 95% percentile-bootstrap intervals (5,000 resamples), computed directly from the per-image results.

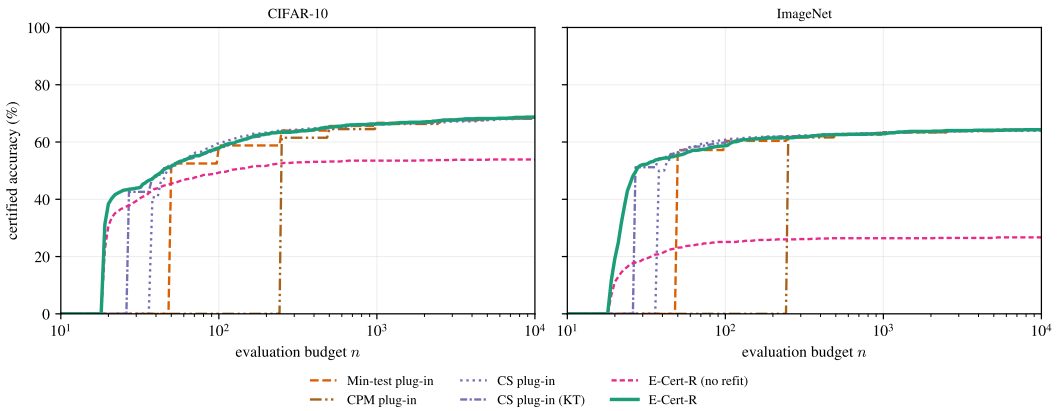


Figure 8: Certified accuracy against the evaluation budget n on the hardmax argmax streams, for six of the seven arms of Table 23, at every n rather than at the ten checkpoints. The staged plug-in is drawn in its min-test form, the stronger of the two: it leads the Bonferroni form by 2.3/3.5 pp at $n = 50$ and by 0.1 pp at $n = 10^4$, and the two share a colour and marker. The staged baselines may only reject at their checkpoints, which is visible as steps; CPM cannot decide before 250 because its $n_0 = 100$ prefix consumes the 100 checkpoint. The E-Cert-R (no refit) arm rises early but plateaus far below the rest. This benchmark uses a single score order, so there are no permutation bands.

Results. Figure 8 and Table 23 show the two staged plug-ins, CPM, and E-Cert-R converging to nearly identical certified accuracies by $n = 10^4$. The first two both reach 68.8% (95% Wilson CI $[65.9, 71.6]\%$), against E-Cert-R's 68.7% ($[65.8, 71.5]\%$) on CIFAR-10, and 64.3% ($[61.3, 67.2]\%$, all three identically) on ImageNet. The E-Cert-R (no refit) arm has lower terminal certified accuracy: 53.9% ($[50.8, 57.0]\%$) on CIFAR-10 and 26.7% ($[24.1, 29.5]\%$) on ImageNet. E-Cert-R (no refit) and E-Cert-R are the same construction differing only in whether the working law is refit, and both see the same clean-vector side information the Bonferroni plug-in and CPM never get. Refitting improves terminal certified accuracy, particularly on ImageNet. The initialization control below examines the source of this improvement.

At small budgets the Bonferroni plug-in and E-Cert-R are already close (57.9%/57.9% at $n = 100$ on CIFAR-10, 59.3%/58.7% on ImageNet). CPM is exactly 0% through $n = 100$ by construction, since $n_0 = 100$ leaves no estimation samples at or before that checkpoint, and it still trails at $n = 250$. By $n = 1{,}000$, the Bonferroni plug-in, CPM, and E-Cert-R are within one percentage point of each other on both datasets; E-Cert-R (no refit) remains well below throughout.

The difference between anytime and staged procedures is more pronounced in stopping cost. On the common-success set, E-Cert-R's median cost is 19 (CIFAR-10) and 19–20 (ImageNet, bootstrap CI) samples, against 50 for the Bonferroni plug-in and 250 for CPM on both datasets at this ten-point grid. The E-Cert-R (no refit) arm's median cost (20 CIFAR-10, 21–23 ImageNet) is close to E-Cert-R's *among the images it does certify*, and the theory promises no sample-path speedup from refitting, so the E-Cert-R (no refit) arm's large CA gap above is a ceiling effect, not a stopping-time effect.

These ratios include checkpoint-spacing effects, which § I.13 assesses separately. The Bonferroni plug-in's and CPM's costs here are checkpoint-quantized, with only 10 possible values. Their bootstrap intervals therefore collapse to a single point, $[50, 50]$ and $[250, 250]$. With 539 and 267 five-arm common successes, a strict majority of images cross between the same pair of checkpoints, so every bootstrap resample's median lands on the same grid value. Table 23's set intersects all seven arms and both grid variants; it has 538/267 and gives the same medians. How much of the gap this granularity accounts for, and how much CPM's pilot size accounts for, are quantified in § I.13 and § I.15, which credit each baseline at its own best tested configuration; the ratios quoted from here on are those. The comparison between E-Cert-R and its no-refit variant is not subject to this confound, since both are measured at their exact crossing time.

The figure shows both the checkpoint constraints on staged stopping and the lower terminal accuracy of the E-Cert-R (no refit) design. E-Cert-R combines early certification with the terminal accuracy attained by the staged baselines.

Table 23: The multiclass hardmax benchmark at $n = 10^4$, 1,000 images per dataset. CA uses all images. Medians use one common-success set (538/267 images), intersecting all seven arms and both grid variants. All-images costs, with uncertified inputs charged the cap, are in Figure 3 and Table 24. Staged costs: *ten-point/best tested grid* (§ I.13). Details: § I.

	CIFAR-10		ImageNet	
Method	CA	Median stop	CA	Median stop
Bonferroni plug-in	0.688	50 / 36	0.643	50 / 41
Min-test plug-in	0.689	50 / 32	0.643	50 / 32
CPM plug-in (Delattre et al., 2025)	0.689	250 / 133	0.643	250 / 133
CS plug-in	0.681	38	0.639	38
CS plug-in (KT) (Voráček, 2024)	0.686	27	0.643	27
E-Cert-R (no refit)	0.539	20	0.267	22
E-Cert-R	0.687	19	0.643	20

CPM's partition-selection failures do not reduce overall CA relative to the other procedures on these data. Its $n_0 = 100$ initial phase misidentifies the true target class on 21.8% of CIFAR-10 images (218/1,000) and 29.7% of ImageNet images (297/1,000). At $n_0 = 60$ the rate is 22.2%/29.6%, so it is not sensitive to that choice. On every one of those images CPM certifies 0% of the time at any budget, exactly as predicted. But the Bonferroni plug-in, E-Cert-R (no refit), and E-Cert-R *also* certify 0% of those same images by $n = 10^4$. Stratify by whether CPM's partition matches the target. On the matched subset the Bonferroni plug-in, CPM and E-Cert-R are statistically indistinguishable (88.0% Bonferroni plug-in/CPM and 87.9% E-Cert-R on CIFAR-10's 782 matched images; 91.5% for all three on ImageNet's 703). E-Cert-R (no refit) trails there too (68.9% CIFAR-10, 38.0% ImageNet). That is the same refitting gap as above, now independent of the CPM-specific stratification. The mismatched subset is 0% for all four methods on both datasets. In other words, the images whose noisy class-count law is unstable enough for a 100-sample prefix to name the wrong top class are, on this data, almost exactly the images none of the four procedures certifies at $r = 0.125$. The synthetic stress test above is different: it placed a genuinely certifiable point exactly where CPM's partition selection could fail.

I.6 Certified accuracy and speed on jointly certified images

Table 23 is the multiclass benchmark's compact summary: certified accuracy over all images, and the median stop over the images every arm certifies, a paired measure of how fast the arms reach the certificates they all reach. In a paired comparison, E-Cert-R, the Bonferroni plug-in, CPM and the CS plug-in (KT) differ on at most three CIFAR-10 images out of 1,000, and on no ImageNet image at all, indicating close agreement in which images are certified. The deployment measure that charges every image is in Figure 3 and § I.7.

I.7 Stopping cost over all images, and futility for the argmax

Table 23 reports medians on a common-success set, which measures how fast the arms certify the images they all certify. The single-certificate comparison of § 7.1 instead charges every image, an uncertified one at the cap, which measures the total evaluation cost. Table 24 applies the same accounting to the multiclass benchmark. It is the arms behind the multiclass hardmax column of Figure 3, listed one by one and with the Bonferroni form and the E-Cert-R (no refit) ablation the figure leaves out, and it adds the population futility rule of § 4 to every arm, as § H.4 does for the single-class certificates. The benchmark is one cached realization per image, so there are no across-permutation deviations to report.

Figure 9 is the complementary measure over the whole argmax design space, and the counterpart of Figure 7 for the single-class target: certified accuracy at $n = 10^4$ over all 1,000 images against the median stop on the images every arm of the panel certifies. Charging every image is what a deployment pays, but the third of the images nobody certifies dominates every mean and compresses each panel to a factor 1.0–1.2; on the jointly certified images the same arms spread by 4.0–12.5. Our arm is the cheapest of all eight cells.

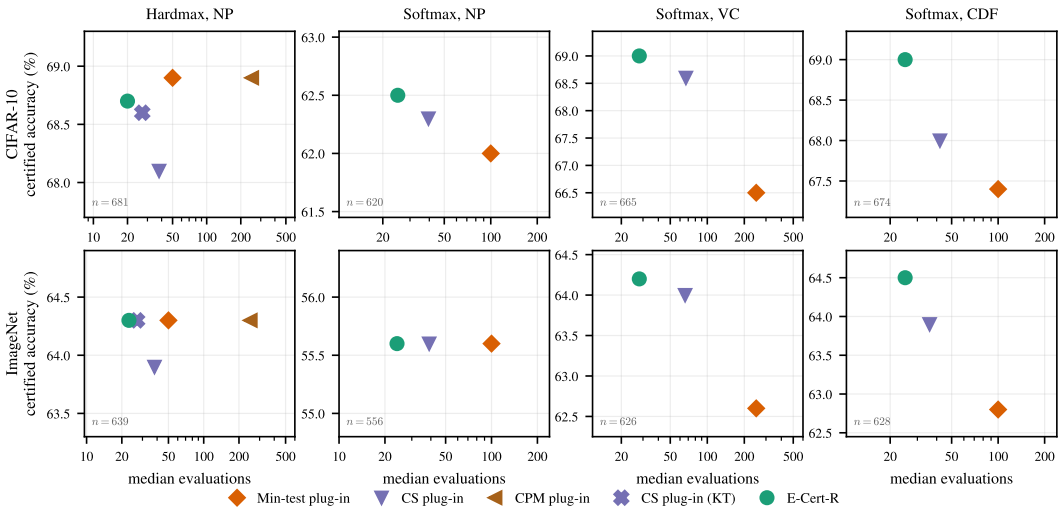


Figure 9: **Multiclass** certified accuracy at $n = 10^4$ over all 1,000 images, versus median stop on the images every displayed arm certifies (n per panel). Columns denote observable–certificate pairs; CPM applies only to hardmax counts. The common-success set contains 681/639 images, compared with 538/267 in Table 23, which also includes ablation arms; E-Cert-R's hardmax medians are therefore 20/23 here and 19/20 there. Fixed-budget and futility variants are omitted for the reasons given in Figure 7. E-Cert-R (no refit) and E-Cert-K are reported separately in Table 24 and Table 27; including the no-refit arm would restrict the common-success set substantially. Comparisons with Figure 7 also change the target: at $\tau = 1/2$, the single-class condition is strictly stronger than argmax certification (§ I.1).

The argmax futility rule mirrors the Bonferroni plug-in. The certified claim is that the target beats every competitor at radius r: $\Phi^{-1}(p_A) - \Phi^{-1}(p_B) > 2\rho$ for all $B \neq A$, equivalently that the radius $\frac{\sigma}{2}\{\Phi^{-1}(p_A) - \Phi^{-1}(\max_B p_B)\}$ exceeds r. The Bonferroni plug-in certifies from a *lower* Clopper–Pearson bound on p_A at $\epsilon/2$ and an *upper* bound on the top observed competitor at $\epsilon/(2(C-1))$. The futility rule runs the same two bounds the other way, an upper bound U_A on p_A and a lower bound $L_{\widehat{B}}$ on the top observed competitor $\widehat{B}$, and abstains once $\frac{\sigma}{2}\{\Phi^{-1}(U_A) - \Phi^{-1}(L_{\widehat{B}})\} < r$. Population protection follows as in the single-class case. Suppose the target truly beats every competitor at radius r. An erroneous trigger then needs $U_A < p_A$, of probability at most $\epsilon/2$, or $L_{\widehat{B}} > p_{\widehat{B}}$ for the data-chosen $\widehat{B}$, an event contained in $\{\exists B : L_B > p_B\}$ and therefore of probability at most $(C-1) \cdot \epsilon/(2(C-1)) = \epsilon/2$. The competitor split thus reappears on the futility side, since abstention has to be justified against every competitor at once, but it enters a Clopper–Pearson tail, so its effect enters through a $\log(C-1)$ term in the bound. The futility rule remains effective at $C = 1{,}000$ in Table 24. Looks and β_f are those of § H.4: $100, 250, 500, 1{,}000, 2{,}500, 5{,}000$ and $\beta_f = \alpha$ split equally. The futility time depends on the class-identity stream alone, so the two new arms combine the published crossing times with it exactly as before. The generating script asserts that disabling futility reproduces every published cell. It also asserts that, on real prefixes, the upper radius never falls below the Bonferroni plug-in's certified radius at the same level: the two bounds bracket the truth, so a checkpoint at which the plug-in certifies can never be futile.

Evaluation cost and futility outcomes. Six arms of the table reach almost the same terminal accuracy — the two plug-in forms, CPM, the two CS plug-ins and E-Cert-R, within 0.8 pp of one another — though not on an identical image set. They incur similar mean costs of 3,248–3,772 evaluations, because the 31–36% of images none of them can certify cost 10,000 each. The E-Cert-R (no refit) arm costs substantially more (4,644 and 7,352), because it also misses images the others certify. With futility E-Cert-R resolves 29.2%/33.8% of images by abstention, leaves 2.1%/1.9% at the cap, and its mean falls to 429 and 338 (7.6× and 10.7×), with no certificate lost. The rule depends on the class-identity stream alone, so the plug-ins take it too. The Bonferroni, min-test and CPM plug-ins fall to 487/473/614 on CIFAR-10 and 375/363/482 on ImageNet, again with no certificate lost. That leaves E-Cert-R 1.1–1.4× below them, rather than the 1.0–1.2× of the base

Table 24: Argmax certification on hardmax class identities (exact two-mean null, $\sigma = 0.25$, $r = 0.125$): stopping cost over *all* 1,000 images per dataset, an uncertified image charged the 10,000 cap, one cached realization. "+ futility" adds the argmax population futility rule of § I.7; every image then ends in one of three ways: certified (CA, %), abandoned by the rule (Futility, share of images), or still running at the cap (Cap, share of images). Lost counts certificates the base arm reaches within budget that the rule abandons first ("–": no futility rule). The Bonferroni and min-test rows are the two forms of the classwise plug-in (§ I.4). Bold marks the smallest median and mean in each block.

Dataset	Method	Median	Mean	CA	Futility	Cap	Lost
CIFAR-10	Bonferroni plug-in	50	3,306.3	68.8	–	31.2%	–
	Min-test plug-in	50	3,292.1	68.9	–	31.1%	–
	CPM plug-in	250	3,429.2	68.9	–	31.1%	–
	CS plug-in	47	3,304.8	68.1	–	31.9%	–
	CS plug-in (KT)	45	3,269.4	68.6	–	31.4%	–
	E-Cert-R (no refit)	117	4,644.5	53.9	–	46.1%	–
	E-Cert-R	**44.5**	3,248.4	68.7	–	31.3%	–
	Bonferroni plug-in + futility	50	487.1	68.8	29.2%	2.0%	0
	Min-test plug-in + futility	50	472.9	68.9	29.2%	1.9%	0
	CPM plug-in + futility	250	610.1	68.9	29.2%	1.9%	0
	CS plug-in + futility	47	485.7	68.1	29.2%	2.7%	0
	CS plug-in (KT) + futility	45	450.3	68.6	29.2%	2.2%	0
	E-Cert-R (no refit) + futility	100	1,825.3	53.9	29.2%	16.9%	0
	E-Cert-R + futility	**44.5**	**429.3**	68.7	29.2%	2.1%	0
ImageNet	Bonferroni plug-in	50	3,665.4	64.3	–	35.7%	–
	Min-test plug-in	50	3,653.3	64.3	–	35.7%	–
	CPM plug-in	250	3,772.0	64.3	–	35.7%	–
	CS plug-in	40.5	3,658.7	63.9	–	36.1%	–
	CS plug-in (KT)	**27**	3,627.5	64.3	–	35.7%	–
	E-Cert-R (no refit)	10,000	7,352.1	26.7	–	73.3%	–
	E-Cert-R	28	3,628.5	64.3	–	35.7%	–
	Bonferroni plug-in + futility	50	374.9	64.3	33.8%	1.9%	0
	Min-test plug-in + futility	50	362.8	64.3	33.8%	1.9%	0
	CPM plug-in + futility	250	481.4	64.3	33.8%	1.9%	0
	CS plug-in + futility	40.5	368.1	63.9	33.8%	2.3%	0
	CS plug-in (KT) + futility	**27**	**336.9**	64.3	33.8%	1.9%	0
	E-Cert-R (no refit) + futility	100	4,061.6	26.7	33.8%	39.5%	0
	E-Cert-R + futility	28	337.9	64.3	33.8%	1.9%	0

arms: the rule, not the test, removes most of the cap mass. The CS plug-in (KT) is the exception in both directions: it falls to 450 and 337, so it is $1.05\times$ above E-Cert-R on CIFAR-10 and $1.00\times$ below it on ImageNet. On this benchmark, once a plug-in uses the estimator built for a $0/1$ stream, the deployment cost of the two families is the same. The e-process's remaining advantage is the speed measured on jointly certified images. The E-Cert-R (no refit) arm gains much less and $16.9\%/39.5\%$ of its images stay at the cap. Refitting certifies 148 of those 169 and 376 of those 395, so most were reachable by a better design rather than futile; the remaining 21 and 19 stay undecided, since neither arm resolves them. These results distinguish the effects of futility and witness adaptation: futility reduces cost on uncertifiable inputs but does not resolve failures caused by inefficient designs.

Audit. As for the single-class rule (§ H.4), we place a categorical law exactly on the two-mean boundary ($p_B = 0.2$ and $p_A = \Phi(\Phi^{-1}(0.2) + 2\rho) = 0.563$, the remainder spread over the other classes) for $C = 10$ and $C = 1{,}000$, and run 20,000 streams of 5,000 draws through the deployed rule. No stream triggers at either C; the exact 95% upper limit on the trigger rate is 1.5×10^{-4} against $\beta_{\mathrm{f}} = 10^{-3}$.

Stake cap and capital preservation. E-Cert-K anchors its witness at the working law and caps its stake at half Kelly; E-Cert-R solves for the exact log-optimal stake under a feasibility guard of 10^{-8} and, for VC and CDF, also moves the witness to the projected law. Neither arm is run at a cap of 1: there the stake can reach exactly $1/(-u_{\min})$, and one worst-case draw sends the log capital to $-\infty$ for good. An exploratory run at that cap had lower medians than E-Cert-R but about 15 points lower

certified accuracy because some paths lost all capital. Sweeping the cap over $\{0.25, \ldots, 0.99\}$ on the hardmax margins changes no conclusion; the softmax control separates witness and stake effects (§ I.11), where the two arms can be matched.

The slope heuristic on the argmax arms. Run on the argmax streams, the cheaper rule of § H.4 repeats the single-class pattern: at the smallest minimum count it undercuts the population rule and discards certificates doing so, and by $n_{\min} = 2{,}500$ it is both dearer and lossy. Figure 3 draws it on E-Cert-R; the per-arm numbers are in the released results. We report the population rule in the tables because it is the one carrying (6).

I.8 Working-law initialization and refit frequency

In § 7.2, refitting increases certified accuracy from 0.539 to 0.687 on CIFAR-10 and from 0.267 to 0.643 on ImageNet. This comparison combines repeated adaptation with improvement over the initial clean-softmax working law. We vary initialization and refit frequency separately to assess these effects.

Four seeds are each held fixed for the entire stream: the paper's clean softmax vector; a uniform law; a law estimated from a 100-draw pilot prefix; and an oracle seed computed from the full stream. The pilot's 100 draws are charged to the budget, as CPM's are. The oracle is not deployable and serves as a retrospective reference. Independently, four refit counts: two, three and five geometrically spaced refits, and the paper's doubling schedule (15 designs). Two further arms run that schedule with the clean vector removed from the arm entirely, replacing it as the Dirichlet prior by the uniform law, once on its own and once after a charged 100-draw pilot. The self-check asserts that the clean-seed arm reproduces the published E-Cert-R (no refit) stops and the doubling arm the published E-Cert-R stops exactly on both datasets, and that the pilot prefix is charged.

Effect of working-law initialization on certified accuracy. Table 25 shows a single pilot-seeded design, never refit, reaching 0.677 and 0.636 against the doubling schedule's 0.687 and 0.643. Relative to the clean-seed accuracies 0.539 and 0.267, this closes 93.2% and 98.1% of the gap: $(0.677 - 0.539)/(0.687 - 0.539)$ and $(0.636 - 0.267)/(0.643 - 0.267)$, respectively. The oracle seed reaches 0.689 and 0.644, *above* the doubling arm on both datasets, showing that a single design can attain comparable terminal accuracy in these experiments. The uniform seed certifies nothing on either dataset — with one design and no refit, a working law whose mean is $1/C$ never leaves the null, so the Kelly stake stays at zero. That is a property of freezing the design, not of the clean vector: run the paper's doubling schedule with a uniform prior instead and it certifies 0.687 and 0.641 at medians of 22 and 21, against 0.687 and 0.643 at 19 and 20 with the clean prior. The clean forward pass is therefore worth 0–0.2 pp of accuracy and one to three evaluations here, which the plug-in baselines' lack of access to it (§ I.5) cannot explain the cost comparison by. Charging a 100-draw pilot on top, as CPM is charged for its partition, leaves accuracy at 0.688 and 0.643 and moves the median to 119: the pilot is pure overhead, and the marginal stopping time is unchanged. A paired count makes the residual precise: doubling beats the pilot seed on 10 images out of 1,000 on CIFAR-10 and 7 on ImageNet, with none the other way, yielding small paired differences ($p = 0.002$ and $p = 0.016$). Against the oracle seed doubling is indistinguishable (0 against 2, and 0 against 1; $p = 0.5$ and $p = 1.0$).

Evaluation cost of pilot initialization and online refitting. Pilot initialization incurs additional sampling cost: its median stop is 119 evaluations against the doubling schedule's 19 and 20, the difference being almost exactly the 100 pilot draws. Three refits already reach 0.675 and 0.632, five reach 0.686 and 0.642, and it is flat thereafter. A factor-1.5 geometric schedule with 24 designs reproduces the doubling arm's 0.687 and 19, 0.643 and 20 exactly, and so does a schedule refitting after every one of the first 200 rounds, with 207 designs. Online refitting therefore improves the working law while accumulating evidence, without a separate pilot phase. CPM also incurs such a prefix cost (§ I.15).

Clean-input predictions and working-law initialization on ImageNet. These base classifiers are trained on Gaussian-augmented inputs, so a clean forward pass evaluates them off their training distribution. On the same 1,000 images the clean argmax is correct on only 37.2% of ImageNet images while the smoothed target probability clears $\Phi(\rho)$ on 64.1%; The two disagree on 32.7%

Table 25: Seed quality and refit count varied separately. "Seed" arms fix one design for the whole stream; "refit" arms use the paper's clean seed and vary how often it is updated. The pilot arm's 100-draw prefix is charged to its stopping cost, which is why its median exceeds the others by about 100. The oracle seed uses the full stream and is not deployable; it provides a retrospective reference for a single fixed design. Median stop is over the images every arm certifies (undefined, "–", for the uniform seed, which certifies nothing). Every projection is audited against its supporting inequality; the worst excess over all arms and images is 6.4×10^{-15}.

Dataset	Method	CA	Median stop†
CIFAR-10	clean (paper)	0.539	20
	uniform seed	0.000	–
	pilot-100 seed	0.677	119
	oracle seed	0.689	19
	2 geometric refits	0.567	20
	3 geometric refits	0.675	20
	5 geometric refits	0.686	19
	doubling (paper)	0.687	19
	doubling, uniform prior	0.687	22
	doubling, pilot-100	0.688	119
ImageNet	clean (paper)	0.267	22
	uniform seed	0.000	–
	pilot-100 seed	0.636	119
	oracle seed	0.644	19
	2 geometric refits	0.487	20
	3 geometric refits	0.632	20
	5 geometric refits	0.642	20
	doubling (paper)	0.643	20
	doubling, uniform prior	0.641	21
	doubling, pilot-100	0.643	119

of images, most of them clean-input errors for which smoothing still supports certification. On CIFAR-10 the figures are 70.7% against 67.1%, with 17.8% disagreement. The E-Cert-R (no refit) arm of § 7.2 seeds its working law from exactly this clean vector and never revisits it, so its design is only as good as that agreement. That matches the direction and rough size of its shortfall (0.539 against 0.687 on CIFAR-10, 0.267 against 0.643 on ImageNet). These observations support the interpretation that inaccurate clean-vector initialization contributes to the E-Cert-R (no refit) arm's low accuracy on ImageNet.

The results support improving an inaccurate working law as the main benefit of refitting in this benchmark. Online updates do so without a separate sampling budget (§ 7.2, § 8).

I.9 The multiplicity split, its dependence on C, and per-round cost

Two further ablations. Their self-check asserts that the reimplemented refit loop reproduces the main benchmark's stopping loop exactly, on real data from both datasets; the split variant is that same audited loop run at a rescaled α, so no new estimator is introduced anywhere in this subsection.

Sampling cost of competitor-wise error allocation. § I.13 found that the $\alpha/(C-1)$ correction has little effect on certified accuracy. Its effect on sample cost is larger. A practitioner who wants the $C-1$ pairwise statements to hold *simultaneously*, as confidence statements rather than as one intersection–union decision, must Bonferroni-correct across competitors, requiring every pair to cross $(C-1)/\alpha$ rather than $1/\alpha$; the decision itself does not need the split (§ I.4). The last row of each block of Table 26 ($C' = C$) shows what that costs at equal accuracy: the median stopping cost rises from 20 to 26 evaluations on CIFAR-10 ($C = 10$) and from 23 to 42 on ImageNet ($C = 1{,}000$), factors of $1.3\times$ and $1.8\times$. The two measurements together give a quantitative comparison: the correction has little effect on certified accuracy but increases median sample cost by 30–80%.

Dependence of additional sample cost on $\log C$. The two ratios, $1.3\times$ at $C = 10$ and $1.8\times$ at $C = 1{,}000$, suggest a possible $\log C$ dependence. The two datasets also differ in classifier, confidence profile and image difficulty. To control for these differences, we sweep competitor count within each dataset.

For each image and each C', the certified claim is restricted to "the target beats every class in $S_{C'}$", where $S_{C'}$ is the target plus the $C' - 1$ competitors of largest **clean** softmax probability. Two features control the effect of subset selection. $S_{C'}$ comes from the noiseless forward pass, available before any noisy draw, so it is a predictable design choice and each restricted min-test is a valid certificate of a genuine (weaker) claim, not a data-dependent selection on the stream being tested. Ranking by clean probability is a heuristic for adding the hardest competitors first: it is not the noisy or the certificate ordering, and it does not guarantee that the bottleneck pair is present at $C' = 2$ — ImageNet's certified accuracy falls from 77.3% there to 64.3% at the full label space, so the restricted problem is genuinely easier. What the design does guarantee is that both arms run on the *same* predictable subset, so the split and no-split medians are compared on identical claims at every C'; the reversed and random orderings below test how much the ranking itself matters. The whole sweep costs one pass per image, since the per-pair capital paths depend on neither C' nor the threshold. First-crossing times are recorded against every threshold at once, and each cell is a maximum over a subset of pairs. The self-check asserts that the full-C' no-split arm reproduces the published stopping loop exactly.

Table 26 reports the results. The no-split median is *flat* in C', at 20 evaluations for every C' on CIFAR-10 and 23 at every $C' > 2$ on ImageNet, while the split median climbs from 20 to 26 and from 24 to 42. This is what the min-test of Prop. 4 predicts once the selection rule is read together with it. The no-split stopping time is the *maximum* over pairs of per-pair crossing times, so enlarging $S_{C'}$ can only raise it, and whether it does depends on how much slower the bottleneck pair is than the rest. On these images it is barely slower, and the flatness accordingly does not rest on the ordering. We reread the same capital paths with $S_{C'}$ filled instead by measured smoothed counts, by the clean order *reversed* (the hardest competitor added last, the adversarial case for a flatness claim), and by a seeded random permutation. That moves the no-split median by at most four evaluations anywhere in the sweep, and leaves the split-to-no-split ratio within 0.07 of the reported one at every C'. The split raises each pair's log-capital threshold by $\log(C' - 1)$, regardless of competitor ordering. Regressing the extra cost on $\log(C' - 1)$ within each dataset gives slopes of 2.65 and 2.73 evaluations per log-competitor with $R^2 = 0.998$ and 0.999, and intercepts of 0.22 and 0.22 against a predicted zero. The additional cost is approximately log-linear within each dataset, with slopes differing by less than 3%.

A log-growth approximation provides an interpretation of these slopes. The split raises the threshold by $\log(C' - 1)$, and a stream stops near threshold/growth rounds, so the slope should be the reciprocal of the bottleneck pair's per-round log-growth. The measured slopes imply 0.377 and 0.367 nats per round, which are the right order for these margins. These values are consistent with the proposed log-capital growth explanation.

Per-round cost. Prop. 4 costs one model evaluation per round no matter how large C is, but the arithmetic grows, so projection overhead must be considered alongside the reduction in model evaluations. One call to the hardmax projection routine takes 1.86 ms at $C = 10$ and 1.92 ms at $C = 1{,}000$. That call is vectorised across competitors, so the flatness measures the implementation rather than a projection count that is independent of C; on softmax margins it is the shared bottleneck design of (21) that solves one certificate-specific projection per refit, from the selected margin's working law, instead of one per pair; that law is the 100-bin margin histogram, not the three-cell law the hardmax construction projects. A complete image costs 31 ms at $C = 10$ against 121 ms at $C = 1{,}000$ of single-threaded analysis time. These measurements quantify projection and update overhead; they do not establish an end-to-end speedup over another early-stopping procedure. The single-class online experiments include sampling, batching and futility on both soft scores and hard labels, and find no consistent runtime advantage over the strongest early-stopping baselines (Table 4, § D.3, § D.4). The multiclass benchmark reports analysis overhead only.

I.10 Continuous softmax margins: plug-in against e-process

The plug-in-versus-e-process comparison of § I.7 is run on the continuous softmax pairwise margins $Z^{(B)} = \frac{1}{2}(1 + f_A - f_B)$ of § I.2, on the same $1{,}000$ images per dataset with the per-draw softmax

Table 26: Cost of the $\alpha/(C-1)$ split swept *within* each dataset, holding the base classifier and images fixed. C' is the number of classes the argmax claim is restricted to: the target plus the $C'-1$ competitors of largest clean softmax probability, a rule fixed before any noisy draw, so each row is a valid certificate of a weaker claim. ‡: median over the images both arms certify at that C'. At $C'=2$ there is no correction to make, and the ratio is 1.00 as it must be.

		Certified accuracy		Median stop‡		
Dataset	C'	no split	split	no split	split	ratio
CIFAR-10	2	0.720	0.720	20	20	1.00
	3	0.701	0.701	20	22	1.10
	5	0.690	0.690	20	24	1.20
	10	0.687	0.687	20	26	1.30
ImageNet	2	0.773	0.773	24	24	1.00
	5	0.731	0.731	23	27	1.17
	10	0.712	0.712	23	29	1.26
	50	0.672	0.672	23	34	1.48
	100	0.661	0.660	23	36	1.57
	500	0.648	0.645	23	40	1.74
	1,000	0.643	0.640	23	42	1.83

vectors of the same cached streams, for all three certificate families. The plug-in baseline is the fixed-n bounds of § 7.1: empirical-Bernstein for NP, empirical-Bernstein mean with the Maurer–Pontil variance bound for VC, and the one-sided DKW band for CDF. Each is applied to every pairwise margin $Z^{(B)} = \frac{1}{2}(1 + f_A - f_B)$ at level $\alpha/10$ per checkpoint, certifying at the first of the ten checkpoints at which the lower certificate of *every* competitor exceeds τ. The Bonferroni form divides by $C-1$ as well. The generating run asserts that the vectorized per-pair bounds reproduce the scalar fixed-n certificates on single pairs, that the hardmax Bonferroni form reproduces the published plug-in stops exactly, and that the min-test form is never later. The *CS plug-in* is the single-class comparator of § 7.1 applied to every pairwise margin and combined by intersection–union. It uses the betting test of $\mathbb{E}[Z^{(B)}] \le T_\rho(\tau)$ at level α for NP; the variance upper sequence at $\alpha/2$ feeding the betting test of the moving threshold $\tau + \kappa_\rho\sqrt{V_t}$ at $\alpha/2$ for VC; and the DKW-type sequence for CDF. Each pair is inspected after every draw, and the image certified when the last pair crosses. The two softmax e-process designs are E-Cert-K and E-Cert-R, the constructions of § I.2, run on the same margins. The population futility rule is added to each by bounding every competitor's certificate from above at $\epsilon/(C-1)$ per checkpoint and abstaining when the smallest such bound is at most τ, so that an erroneous abstention again requires a coverage failure.

Scope of the CPM comparator. CPM reads the *selection counts* of the argmax: it buckets the classes from an n_0-draw prefix and bounds a bucket by one Clopper–Pearson interval on the *sum* of its members' counts, which is what lets the correction run over $c^\star$ buckets instead of $C-1$ classes. The published construction applies to class counts (Delattre et al., 2025); continuous margins do not supply these counts. One could write an analogue that bounds $\sum_{B\in\text{bucket}} \mathbb{E}[f_B]$ and uses $\max_B \mathbb{E}[f_B] \le \sum_B \mathbb{E}[f_B]$, but it would be dominated by the baseline already run here: it pays $\alpha/c^\star$ *and* the slack of a sum standing in for a maximum, where the min-test plug-in pays neither. The ordering is visible on hardmax, where both are run. The min-test median is 50 against CPM's 250 (Table 24), so omitting CPM here leaves the stronger of the two plug-in forms as the comparator.

Table 27 charges every image. Against the staged min-test plug-in, the e-process arms certify 0–2.5 pp more images on every certificate and reach them 2–12× sooner in median, E-Cert-R 5–12×. Both plug-in forms are tabulated and the min-test one is quoted, being the stronger. These cells are where the split is most visible. § I.4 makes its cost a property of the grid rather than of the level, so it appears only when the extra level moves the crossing past a checkpoint; on the hardmax full budget it rarely does, leaving the two forms within 15 mean evaluations and 0.1 pp at a common median. These margins certify an order of magnitude earlier, so the crossing sits among the sparse early checkpoints and the split moves it in four of the six cells (500 against 250 on CIFAR-10 NP and ImageNet CDF, 1,000 against 500 on ImageNet NP and VC), worth up to 0.7 pp of certified accuracy. Futility does not account for the gap: it is charged to both forms off the one shared statistic, since it reads the

margins and not the level a form spends, and the same four cells separate after it (250 against 100, and 500 against 250 on ImageNet VC). The plug-in medians sit at the 250 and 500 checkpoints, the e-process medians at 40–113. The CS plug-in removes the checkpoint quantization and lands at medians of 73–124. That is level with E-Cert-K (0.8–1.3×) and 1.5–2.7× behind E-Cert-R, at certified accuracy 0–1 pp below it. On continuous margins the betting sequence and E-Cert-K are close substitutes. The two e-process arms differ in two ways at once, however. E-Cert-R takes the certificate-specific witness *and* solves for the exact log-optimal stake, bounded only by a 10^{-8} numerical guard, while E-Cert-K caps its stake at half Kelly. The matched control below separates these two changes. Means are dominated by the mass at the evaluation cap, and futility removes most of it with no certificate lost in any of the twenty-four cells. Applied to the plug-ins it does the same (means 722–1,831 for the staged and 563–1,423 for the CS plug-in against 527–1,326 for E-Cert-R, margins of 1.3–1.8× and 1.0–1.4×, while the base arms' means were within 1.3× of each other). VC futility leaves more at the cap (6–11%) than NP or CDF (3–5%): its upper bound subtracts a variance *lower* bound that is loose per pair at $\epsilon/(C-1)$, and a loose upper bound is the conservative direction.

I.11 Effects of witness selection and stake optimization

We compare three designs on the continuous margins of Table 27, holding the working law, shared-bottleneck selection, refit schedule and stake cap fixed. The first uses a working-law witness with a quadratic-Kelly stake. The second retains that witness and optimizes the stake exactly. The third retains the exact stake optimizer and uses the witness at the projected bad law. Thus the first comparison isolates stake optimization and the second witness selection. For NP, the witness is fixed by the half-space null, so the latter two designs coincide. All three are distinct for VC and CDF.

This is the multiclass counterpart of § F.1 and uses the same shared cap, 0.99. It is a control, not a deployed configuration: the published arms of Table 27 keep their own caps (half Kelly for E-Cert-K, the 10^{-8} feasibility guard for E-Cert-R), which is why the medians here do not reproduce that table's.

At the shared cap, exact stake optimization does not improve median stopping time: it increases the median by 1–2 evaluations (0.94–0.97×), with mean-cost ratios of 0.98–1.06× (Table 28). Replacing the witness reduces median cost by 1.24–1.44× and mean cost by 1.09–1.29× on VC and CDF. The NP ratio is exactly 1.00×, as required by the identical witness $z - \Phi(\rho)$ and shared stake optimizer. This column checks implementation consistency. Certified accuracy differs by at most 0.3 pp across the three arms in every cell.

Under these controls, the measured gain comes from witness selection for VC and CDF. The single-class control also finds little benefit from replacing the quadratic stake alone (Table 8).

These medians use the images all three arms certify. Table 29 reports all-images costs, charging unresolved inputs the 10^4 cap and applying the population futility rule of § 4 to all three arms. Futility times depend only on the margins, so we reuse the per-image times from § I.10. The generating script checks that this reconstruction reproduces the published cells.

The reduction in all-images mean cost is smaller because unresolved inputs are common to all three arms and dominate the mean. With paired image bootstrap intervals, the witness-change ratios are 1.019× [1.013, 1.026] on CIFAR-10 VC, 1.024× [1.017, 1.033] on CIFAR-10 CDF and 1.030× [1.019, 1.049] on ImageNet CDF. On ImageNet VC the ratio is 0.984× [0.959, 1.004]: the RIPr-guided witness certifies 642 images against the working-law witness's 645, and the three additional inputs charged at 10^4 reverse the mean-cost comparison.

The attribution is stable when the noise itself is redrawn. Table 30 repeats the control on three independent Gaussian replicates of one fixed 200-image subset, each with its own futility times. The witness-change ratio is 1.024 ± 0.011 on CIFAR-10 VC, 1.021 ± 0.003 on CIFAR-10 CDF, 1.017 ± 0.0004 on ImageNet CDF and 0.987 ± 0.008 on ImageNet VC across the three, reproducing the direction and the size of every cell of Table 29, and exactly 1.000 on NP in all three. Certified accuracy moves by at most 0.5 pp between replicates (a standard deviation of 0.29 pp).

Table 27: Softmax pairwise margins over all 1,000 images, an uncertified image charged the cap, per certificate. Plug-ins are the fixed-n bounds of § 7.1 per pair on a ten-checkpoint grid, in Bonferroni and min-test form; the two e-process designs are E-Cert-K and E-Cert-R, the constructions of § I.2. Columns as in Table 24; bold marks the smallest median and mean in each block.

Dataset	Cert.	Method	Median	Mean	CA	Futility	Cap	Lost
CIFAR-10	NP	Bonferroni plug-in	500	4,121.9	62.0	–	38.0%	–
		Min-test plug-in	250	4,063.5	62.0	–	38.0%	–
		CS plug-in	73.5	3,859.5	62.3	–	37.7%	–
		E-Cert-K	77	3,845.8	62.4	–	37.6%	–
		E-Cert-R	**49**	3,841.9	62.5	–	37.5%	–
		Bonferroni plug-in + futility	250	862.0	62.0	34.6%	3.4%	0
		Min-test plug-in + futility	100	803.5	62.0	34.6%	3.4%	0
		CS plug-in + futility	73.5	599.5	62.3	34.6%	3.1%	0
		E-Cert-K + futility	77	585.9	62.4	34.6%	3.0%	0
		E-Cert-R + futility	**49**	**581.9**	62.5	34.6%	2.9%	0
	VC	Bonferroni plug-in	500	3,940.8	66.3	–	33.7%	–
		Min-test plug-in	500	3,845.2	66.5	–	33.5%	–
		CS plug-in	105.5	3,309.3	68.6	–	31.4%	–
		E-Cert-K	92	3,246.6	69.0	–	31.0%	–
		E-Cert-R	**42**	3,208.4	69.0	–	31.0%	–
		Bonferroni plug-in + futility	250	1,818.1	66.3	24.7%	9.0%	0
		Min-test plug-in + futility	250	1,722.6	66.5	24.7%	8.8%	0
		CS plug-in + futility	100	1,186.7	68.6	24.7%	6.7%	0
		E-Cert-K + futility	92	1,123.9	69.0	24.7%	6.3%	0
		E-Cert-R + futility	**42**	**1,085.7**	69.0	24.7%	6.3%	0
	CDF	Bonferroni plug-in	250	3,692.8	67.0	–	33.0%	–
		Min-test plug-in	250	3,634	67.4	–	32.6%	–
		CS plug-in	102	3,383.6	68.0	–	32.0%	–
		E-Cert-K	102	3,232.5	69.0	–	31.0%	–
		E-Cert-R	**40**	3,194.8	69.0	–	31.0%	–
		Bonferroni plug-in + futility	100	1,025.2	67.0	28.0%	5.0%	0
		Min-test plug-in + futility	100	966.4	67.4	28.0%	4.6%	0
		CS plug-in + futility	100	716.0	68.0	28.0%	4.0%	0
		E-Cert-K + futility	100	564.9	69.0	28.0%	3.0%	0
		E-Cert-R + futility	**40**	**527.2**	69.0	28.0%	3.0%	0
ImageNet	NP	Bonferroni plug-in	1,000	4,773.5	55.4	–	44.6%	–
		Min-test plug-in	500	4,650.2	55.6	–	44.4%	–
		CS plug-in	110	4,491.4	55.6	–	44.4%	–
		E-Cert-K	112	4,491.7	55.6	–	44.4%	–
		E-Cert-R	**68.5**	4,477.0	55.6	–	44.4%	–
		Bonferroni plug-in + futility	250	845	55.4	41.3%	3.3%	0
		Min-test plug-in + futility	100	721.8	55.6	41.3%	3.1%	0
		CS plug-in + futility	100	562.9	55.6	41.3%	3.1%	0
		E-Cert-K + futility	100	563.2	55.6	41.3%	3.1%	0
		E-Cert-R + futility	**68.5**	**548.5**	55.6	41.3%	3.1%	0
	VC	Bonferroni plug-in	1,000	4,335	62.1	–	37.9%	–
		Min-test plug-in	500	4,142.9	62.6	–	37.4%	–
		CS plug-in	124	3,735.4	64.0	–	36.0%	–
		E-Cert-K	98	3,651.0	64.5	–	35.5%	–
		E-Cert-R	**46**	3,638.7	64.2	–	35.8%	–
		Bonferroni plug-in + futility	500	2,022.8	62.1	26.4%	11.5%	0
		Min-test plug-in + futility	250	1,830.8	62.6	26.4%	11.0%	0
		CS plug-in + futility	100	1,423.2	64.0	26.4%	9.6%	0
		E-Cert-K + futility	98	1,338.8	64.5	26.4%	9.1%	0
		E-Cert-R + futility	**46**	**1,326.5**	64.2	26.4%	9.4%	0
	CDF	Bonferroni plug-in	500	4,210.1	62.1	–	37.9%	–
		Min-test plug-in	250	4,033.7	62.8	–	37.2%	–
		CS plug-in	91.5	3,743.6	63.9	–	36.1%	–
		E-Cert-K	113	3,668.0	64.5	–	35.5%	–
		E-Cert-R	**43**	3,605.1	64.5	–	35.5%	–
		Bonferroni plug-in + futility	250	1,134.0	62.1	32.2%	5.7%	0
		Min-test plug-in + futility	100	957.5	62.8	32.2%	5.0%	0
		CS plug-in + futility	91.5	667.5	63.9	32.2%	3.9%	0
		E-Cert-K + futility	100	591.9	64.5	32.2%	3.3%	0
		E-Cert-R + futility	**43**	**529.0**	64.5	32.2%	3.3%	0

I.12 Certified accuracy against radius on the softmax margins

§ I.10 compares the arms at one operating point, $\sigma = 0.25$ and $r = 0.125$. The single-class certificates are swept over radii and smoothing levels (§ G) and the hardmax argmax arms over radii (§ I.14);

Table 28: Witness against stake on the softmax pairwise margins, all three arms at one shared stake cap (0.99), the same shared-bottleneck working law and the same refit schedule; 1,000 images per dataset. CA is certified accuracy over all images; †: median and mean over the images all three arms certify (556–690 per cell). The first two arms isolate the stake rule and the last two the witness. On NP the RIPr-guided witness *is* the fixed NP witness, so the last two are the same design there and coincide by construction.

Dataset	Cert.	Design	CA	Median†	Mean†
CIFAR-10	NP	Working-law witness / quadratic stake	0.623	24	119
		Working-law witness / exact stake	0.625	25	117
		RIPr witness / exact stake	0.625	25	117
	VC	Working-law witness / quadratic stake	0.689	35	182
		Working-law witness / exact stake	0.689	36	172
		RIPr witness / exact stake	0.690	29	144
	CDF	Working-law witness / quadratic stake	0.690	35	163
		Working-law witness / exact stake	0.690	37	155
		RIPr witness / exact stake	0.690	26	137
ImageNet	NP	Working-law witness / quadratic stake	0.556	23	66
		Working-law witness / exact stake	0.556	24	67
		RIPr witness / exact stake	0.556	24	67
	VC	Working-law witness / quadratic stake	0.645	34	98
		Working-law witness / exact stake	0.645	35	100
		RIPr witness / exact stake	0.642	28	92
	CDF	Working-law witness / quadratic stake	0.645	34	108
		Working-law witness / exact stake	0.645	36	110
		RIPr witness / exact stake	0.645	25	86

this subsection supplies the missing axis for the continuous margins. Only $\rho = r/\sigma$ enters the tests, so every radius reuses the cached softmax shards of Table 27 unchanged and only the certificate arithmetic is repeated. We use the radius grid of Table 33, so the two curves can be read against each other, and as there each radius is a separate prespecified decision at α. A self-check reproduces the published per-image stopping times of Table 27 exactly at the $r = 0.125$ column, for both plug-in forms.

The three certificates separate in the way § G finds for the single-class target, and for the same reason; Figure 10 draws the three side by side. **NP** is flat up to $\rho = 2$: the four arms lie within 1.7 points of one another at every $r \leq 0.5$, because the NP null is a half-space and its witness is fixed, so a design can only choose the stake (§ B.3). It stops being flat at $r = 0.75$, for a reason outside the certificate, which the $\rho = 3$ paragraph below separates from the stake cap. **VC** opens a gap that grows with ρ: at $r = 0.5$ E-Cert-R certifies 28.9% against the min-test plug-in's 17.9% on CIFAR-10, and 35.5% against 24.5% on ImageNet, with E-Cert-K at 28.3 and 34.5. **CDF** separates into three layers. At $r = 0.375$ the ordering is plug-in 42.6/46.4, E-Cert-K 49.2/52.8, E-Cert-R 49.9/55.1 (CIFAR-10/ImageNet). At $r = 0.5$ the Bonferroni form certifies *nothing at all* and the min-test form 10.7/14.3, against E-Cert-K's 33.2/40.2 and E-Cert-R's 41.3/49.6.

At $r = 0.75$ ($\rho = 3$), E-Cert-R is the only arm that certifies any image on either dataset, and on *two* certificates: CDF at 11.5%/20.3% and NP at 6.0/6.3% (CIFAR-10/ImageNet), with E-Cert-K and both plug-ins at zero on each. The two zeros have different causes, neither of them the certificate's. NP needs the threshold-aware design grid of § B.2: on the uniform grid no working law can leave the NP null at this ρ, so an arm reading its stake off one is silent by construction rather than by evidence. E-Cert-K stays at zero even on the resolved grid, because its half-Kelly cap bounds what a single draw contributes and puts the first possible crossing at 10,224 draws against a budget of 10^4. The single-class CDF sweep shows the same pattern at this ρ (§ G). In both targets, E-Cert-R retains positive certified accuracy where the compared procedures reach zero. By $r = 1$ every arm is at zero.

The sweep uses one smoothing level, because a second smoothing level needs its own Gaussian-augmented base classifier and therefore its own softmax cache, which for ImageNet is 38 GB per

Table 29: The witness-against-stake control of Table 28 charged over *every* image, an uncertified one at the 10^4 cap, without and with the population futility rule. CA is certified accuracy (%); Lost counts certificates the base arm reaches within the budget that the rule abandons first. Same streams, same working law, cap and refit schedule as Table 28; only the accounting differs. NP is the wiring check, where the last two arms are the same design.

			No futility		+ population futility		
Dataset	Cert.	Design	Mean	CA	Mean	CA	Lost
CIFAR-10	NP	Working-law witness / quadratic stake	3,844.2	62.3	584.2	62.3	0
		Working-law witness / exact stake	3,841.8	62.5	581.9	62.5	0
		RIPr witness / exact stake	3,841.8	62.5	581.9	62.5	0
	VC	Working-law witness / quadratic stake	3,235.7	68.9	1,113.0	68.9	0
		Working-law witness / exact stake	3,228.5	68.9	1,105.9	68.9	0
		RIPr witness / exact stake	3,208.4	69.0	1,085.8	69.0	0
	CDF	Working-law witness / quadratic stake	3,212.3	69.0	544.7	69.0	0
		Working-law witness / exact stake	3,207.3	69.0	539.7	69.0	0
		RIPr witness / exact stake	3,194.8	69.0	527.2	69.0	0
ImageNet	NP	Working-law witness / quadratic stake	4,476.6	55.6	548.1	55.6	0
		Working-law witness / exact stake	4,477.1	55.6	548.6	55.6	0
		RIPr witness / exact stake	4,477.1	55.6	548.6	55.6	0
	VC	Working-law witness / quadratic stake	3,617.0	64.5	1,304.8	64.5	0
		Working-law witness / exact stake	3,618.1	64.5	1,305.9	64.5	0
		RIPr witness / exact stake	3,638.8	64.2	1,326.6	64.2	0
	CDF	Working-law witness / quadratic stake	3,619.5	64.5	543.4	64.5	0
		Working-law witness / exact stake	3,621.2	64.5	545.1	64.5	0
		RIPr witness / exact stake	3,605.2	64.5	529.1	64.5	0

Table 30: The witness-against-stake control on three *independent* Gaussian noise replicates of one fixed 200-image subset, at the same σ, radius, working law, cap and refit schedule, each with the population futility rule and its own futility times. Mean charges every image, an uncertified one at the 10^4 cap; entries are mean $\pm$ standard deviation across the three replicates. The sampler separates the seed that selects images from the one that drives the draws, so the three differ only in the noise.

Dataset	Cert.	Design	Mean $\pm$ SD	CA (%) $\pm$ SD
CIFAR-10	NP	Working-law witness / quadratic stake	477 ± 22	64.3 ± 0.3
		Working-law witness / exact stake	475 ± 25	64.3 ± 0.3
		RIPr witness / exact stake	475 ± 25	64.3 ± 0.3
	VC	Working-law witness / quadratic stake	971 ± 41	70.2 ± 0.3
		Working-law witness / exact stake	968 ± 38	70.2 ± 0.3
		RIPr witness / exact stake	946 ± 31	70.3 ± 0.3
	CDF	Working-law witness / quadratic stake	549 ± 28	70.3 ± 0.3
		Working-law witness / exact stake	547 ± 26	70.3 ± 0.3
		RIPr witness / exact stake	536 ± 27	70.3 ± 0.3
ImageNet	NP	Working-law witness / quadratic stake	428 ± 3	59.5 ± 0.0
		Working-law witness / exact stake	429 ± 3	59.5 ± 0.0
		RIPr witness / exact stake	429 ± 3	59.5 ± 0.0
	VC	Working-law witness / quadratic stake	$1{,}409 \pm 16$	66.0 ± 0.0
		Working-law witness / exact stake	$1{,}409 \pm 15$	66.0 ± 0.0
		RIPr witness / exact stake	$1{,}428 \pm 4$	65.5 ± 0.0
	CDF	Working-law witness / quadratic stake	620 ± 2	65.7 ± 0.3
		Working-law witness / exact stake	621 ± 1	65.7 ± 0.3
		RIPr witness / exact stake	611 ± 2	65.7 ± 0.3

Table 31: Certified accuracy (%) on the continuous softmax pairwise margins against the perturbation radius, $\sigma = 0.25$, $n = 10^4$, 1,000 images per dataset, on the streams and grid of Table 33; each radius is a separate prespecified decision at α. Bold marks the best arm *at that radius*; a column where every arm is at zero marks nothing. $\rho = r/\sigma$ runs from 0.25 to 4.

Dataset	Cert.	Method	Perturbation radius r							
			0.0625	0.125	0.1875	0.25	0.375	0.5	0.75	1
CIFAR-10	NP	Min-test plug-in	69.7	62.0	54.3	48.2	34.9	24.8	0.0	0.0
		Bonferroni plug-in	69.7	62.0	54.3	48.1	34.5	24.5	0.0	0.0
		E-Cert-K	**69.8**	62.4	**54.9**	**49.1**	**35.3**	**25.9**	0.0	0.0
		E-Cert-R	**69.8**	**62.5**	54.8	**49.1**	**35.3**	25.8	**6.0**	0.0
	VC	Min-test plug-in	72.0	66.5	61.3	53.0	37.2	17.9	0.0	0.0
		Bonferroni plug-in	71.9	66.3	61.0	52.8	36.3	16.1	0.0	0.0
		E-Cert-K	72.6	**69.0**	**63.2**	**57.1**	**43.0**	28.3	0.0	0.0
		E-Cert-R	**72.7**	**69.0**	**63.2**	56.9	42.8	**28.9**	0.0	0.0
	CDF	Min-test plug-in	71.9	67.4	62.3	55.8	42.6	10.7	0.0	0.0
		Bonferroni plug-in	71.6	67.0	61.9	55.4	41.2	0.0	0.0	0.0
		E-Cert-K	**72.5**	**69.0**	**64.8**	60.2	49.2	33.2	0.0	0.0
		E-Cert-R	**72.5**	**69.0**	**64.8**	**60.3**	**49.9**	**41.3**	**11.5**	0.0
ImageNet	NP	Min-test plug-in	61.9	**55.6**	**50.4**	**45.3**	33.4	22.4	0.0	0.0
		Bonferroni plug-in	61.5	55.4	50.2	45.0	32.8	21.6	0.0	0.0
		E-Cert-K	**62.3**	**55.6**	**50.4**	**45.3**	33.5	22.9	0.0	0.0
		E-Cert-R	62.2	**55.6**	**50.4**	**45.3**	**33.6**	**23.1**	**6.3**	0.0
	VC	Min-test plug-in	66.3	62.6	59.3	54.8	43.9	24.5	0.0	0.0
		Bonferroni plug-in	66.0	62.1	58.2	53.2	42.1	17.2	0.0	0.0
		E-Cert-K	**67.8**	**64.5**	**62.2**	**58.2**	**49.0**	34.5	0.0	0.0
		E-Cert-R	67.2	64.2	61.7	57.9	**49.0**	**35.5**	0.0	0.0
	CDF	Min-test plug-in	65.7	62.8	59.6	55.5	46.4	14.3	0.0	0.0
		Bonferroni plug-in	65.0	62.1	58.5	53.9	43.3	0.0	0.0	0.0
		E-Cert-K	**67.0**	**64.5**	62.3	60.3	52.8	40.2	0.0	0.0
		E-Cert-R	**67.0**	**64.5**	**62.5**	**60.8**	**55.1**	**49.6**	**20.3**	0.0

level. And the ρ at which an arm falls to zero is a property of these streams, this budget and the stake cap, not a threshold of the certificate. Which of the three binds is worth settling before a zero is read as evidence: at $r = 1$ ($\rho = 4$) the cap alone does it, since the largest log-factor an NP draw can contribute puts the first possible crossing at 218,105 draws under E-Cert-R's cap and further under E-Cert-K's, so no budget within an order of magnitude of 10^4 would move that column.

I.13 Monitoring-grid sensitivity, multiplicity correction, and fixed-budget comparison

Three controls for § 7.2, all computed with the audited primitives of the main benchmark rather than a reimplementation. Their $m = 10$ rows reproduce the main benchmark's grid exactly, which cross-checks the two scripts against each other.

Staged-grid sensitivity. The staged baselines in § I.5 are monitored on the same ten-point grid used for reporting, so their stopping cost is quantized to ten values. Monitoring more often resolves the crossing better but costs a larger Bonferroni factor, so there is an interior optimum. Table 32 sweeps the grid size m, each variant paying α/m. The trade is mild on accuracy, at most 0.3 pp across the sweep, but material on cost. The Bonferroni plug-in improves from 50 to 36 (CIFAR-10) and 50 to 41 (ImageNet) at $m = 50$, the min-test plug-in from 50 to 32 on both at $m = 25$, and CPM from 250 to 133 at $m = 25$, all then flattening or worsening slightly as the correction grows. Crediting each baseline at its own best m, the comparison most favorable to them and the one the main text's speedup claim is stated against, leaves E-Cert-R ahead by $1.7\times/1.6\times$ over the min-test plug-in, the strongest of the three, $1.9\times/2.1\times$ over its Bonferroni form and $7.0\times/6.7\times$ over CPM. The CPM ratio is refined further in § I.15, which credits CPM at its own best initial-phase size as well and brings the lead down to $2.5\times$ and $2.2\times$; § 7.2 quotes CPM at that best n_0, whereas the ratio

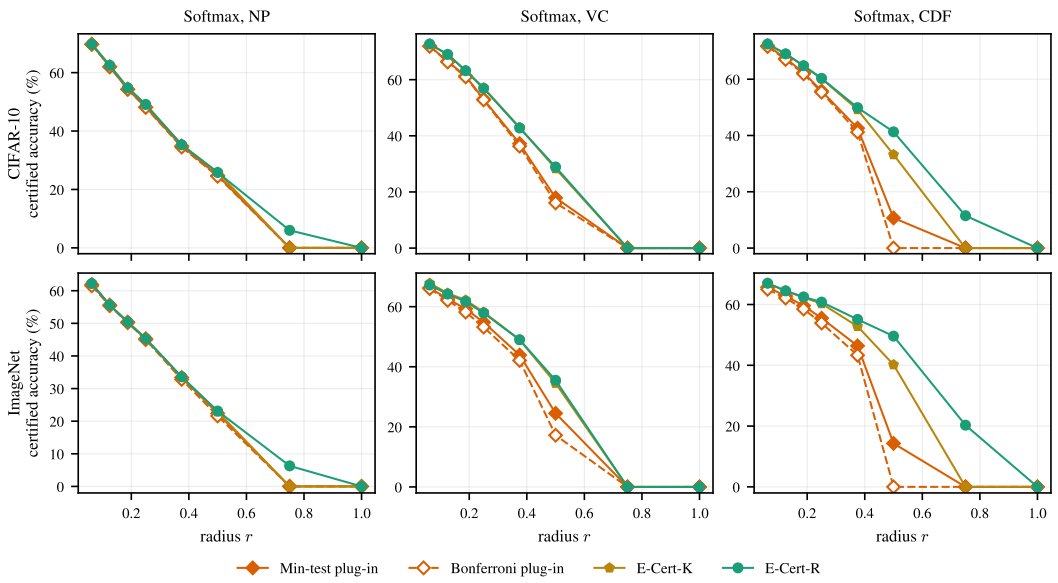


Figure 10: Table 31 drawn: certified accuracy against radius on the softmax pairwise margins, rows datasets and columns certificates. Colour and marker are the procedure as everywhere else; the two plug-in forms share an encoding, so the dashed line is the Bonferroni one. NP is flat across arms up to $r = 0.5$ and separates at $r = 0.75$, where only E-Cert-R certifies; VC opens a gap with ρ; and on CDF the Bonferroni form reaches zero at $r = 0.5$ where the min-test form still certifies 10.7/14.3%, both are at zero by $r = 0.75$, and E-Cert-R alone certifies there.

stated here holds CPM's n_0 at 100. E-Cert-R appears once, with no grid and no correction, since it is credited at its exact first crossing. The common-success set here is 537 (CIFAR-10) and 267 (ImageNet), slightly smaller than the five-arm benchmark's 539/267 (§ I.5) because it additionally intersects success across every grid size.

Effect of multiplicity correction on certified accuracy. Avoiding the $\alpha/(C-1)$ split is the theoretical point of § 6, and its empirical effect requires a controlled comparison. Hold each image's observed stream fixed and charge the Bonferroni plug-in only the correction a C'-class label space would impose. Certified accuracy then falls by 0.3 pp on CIFAR-10 between $C' = 10$ and $C' = 100{,}000$ (0.688, 0.687, 0.686, 0.686, 0.685 at $C' = 10^1, \ldots, 10^5$), and not measurably at all on ImageNet (0.643 at every C'). Clopper–Pearson bounds tighten like $\sqrt{\log(1/\alpha)/n}$, so an additional $\log C$ term has a small observed effect at $n = 10^4$. We therefore do not attribute the multiclass results *at the main radius* to the removed split. There its effect on accuracy is negligible, and the measured advantage comes from anytime monitoring. The picture changes at the margin of certifiability, where the split is decisive. § I.14 shows the split costing the Bonferroni plug-in 2.3 pp at $r = 0.75$ on both datasets, where it certifies 16.2% and 35.4% against its own min-test form's 18.5% and 37.7%. Removing the split remains worthwhile for structural reasons as well: neither a checkpoint grid nor a label-space size has to be fixed before sampling, and Prop. 4 needs no assumption about dependence among the $C-1$ margins. This is a sensitivity analysis of the correction alone; a genuinely larger label space would also change the competitor-count distribution, and we make no claim about that.

Fixed-budget reference. For a reference to standard fixed-budget certification, we run Cohen CERTIFY (Cohen et al., 2019) on the identical cached streams: select the top class from an $n_0 = 100$ prefix, take a one-sided Clopper–Pearson lower bound $\underline{p}_A$ on the remainder at level α, and certify when $\sigma\Phi^{-1}(\underline{p}_A) \geq r$. At $n = 10^4$ it certifies 0.666 (CIFAR-10) and 0.635 (ImageNet), against E-Cert-R's 0.687 and 0.643 (its staged form, checked at the ten checkpoints at $\alpha/10$, certifies 0.663 and 0.633; see § I.16, and note the single-checkpoint figure is the one to compare with the literature). Note this tests a strictly stronger population claim than the pairwise min-test, since it demands

Table 32: Staged monitoring-grid sensitivity. m is the realized number of checkpoints (requested sizes de-duplicate, so $100 \to 99$ and $250 \to 230$). The level column is each arm's own divisor: CPM spends $n_0 = 100$ draws before estimating, so a checkpoint at or below 100 leaves its rejection region empty and is not charged, and it pays over the checkpoints beyond 100 only. †: median stop on the common-success set. E-Cert-R needs no correction ("–" under Looks) and is credited at its exact crossing.

Dataset	Method	Looks m	α-level	CA	Median stop†
CIFAR-10	Bonferroni plug-in	10	$\alpha/10$	0.688	50
		25	$\alpha/25$	0.688	42
		50	$\alpha/50$	0.687	36
		99	$\alpha/99$	0.686	38
		230	$\alpha/230$	0.686	39
	Min-test plug-in	10	$\alpha/10$	0.689	50
		25	$\alpha/25$	0.689	32
		50	$\alpha/50$	0.688	36
		99	$\alpha/99$	0.688	35
		230	$\alpha/230$	0.687	36
	CPM plug-in	10	$\alpha/6$	0.689	250
		25	$\alpha/16$	0.688	133
		50	$\alpha/33$	0.688	146
		99	$\alpha/66$	0.686	142
		230	$\alpha/166$	0.686	136
	E-Cert-R	–	α	0.687	19
ImageNet	Bonferroni plug-in	10	$\alpha/10$	0.643	50
		25	$\alpha/25$	0.643	42
		50	$\alpha/50$	0.643	41
		99	$\alpha/99$	0.643	43
		230	$\alpha/230$	0.643	44
	Min-test plug-in	10	$\alpha/10$	0.643	50
		25	$\alpha/25$	0.643	32
		50	$\alpha/50$	0.643	36
		99	$\alpha/99$	0.643	35
		230	$\alpha/230$	0.643	36
	CPM plug-in	10	$\alpha/6$	0.643	250
		25	$\alpha/16$	0.643	133
		50	$\alpha/33$	0.643	146
		99	$\alpha/66$	0.643	142
		230	$\alpha/166$	0.643	136
	E-Cert-R	–	α	0.643	20

$p_A > \Phi(\rho)$ outright rather than only that A beat its strongest competitor, so it is an anchor for the absolute level, not a like-for-like competitor. Its selection phase names the true label on 78.2% and 70.3% of images, necessarily matching CPM's partition-match rate because both take an $n_0 = 100$ empirical argmax.

I.14 Certified accuracy at representative radii

The preceding multiclass results use one prespecified radius, $r = \sigma\rho$. We now evaluate sensitivity to the radius. We report certified accuracy at a grid of representative radii rather than the average certified radius (ACR). Sun et al. (2025) show that ACR is a poor metric for randomized smoothing:

trivial classifiers can drive it arbitrarily high, and it rewards gains on easy inputs while hiding losses on hard ones. The per-radius curve is what it was meant to summarize.

Each multiclass hardmax procedure of § I.5 is run on the same class-identity streams ($\sigma = 0.25$) at every radius of a fixed grid $G = \{0.0625, 0.125, 0.1875, 0.25, 0.375, 0.5, 0.75, 1.0\}$. Each column is one prespecified decision and every arm spends the full α on it, which is the claim the column makes and puts the table on the same footing as the single-class sweep it is read against (§ G). A certified *radius* would be different — it maximises over the grid using the data, so it needs the $\alpha/|G|$ correction — and the saved per-image radii keep it; we do not report them, for the reason above.

Certified-accuracy differences across radii. Table 33 shows the arms tracking each other closely out to $r = 0.5$. E-Cert-R stays within 0.9 points of the best plug-in at every radius on both datasets, and 8.7–42.2 points above the E-Cert-R (no refit) arm. The competitor split costs the Bonferroni form at most 0.6 pp up to $r = 0.5$, and 2.3 pp at $r = 0.75$ on both datasets, where the certificate is marginal and a factor of $C - 1$ decides images: the pattern matches the $\sigma \times r$ sweep of § G, where differences invisible at a conservative radius reopen as the radius is pushed, so agreement at $r = 0.125$ does not imply agreement at larger radii.

The empirical-Bernstein CS plug-in hits a ceiling at $\rho = 3$. At $r = 0.75$ it certifies *nothing*, on either dataset, while every comparator certifies 16–21% and 35–40% (the E-Cert-R (no refit) ablation, at 9.6 and 2.6%, is not one). This is not a property of the images but of the bound. Take the easiest stream there is, in which the target wins all 10^4 draws, and read off the two-mean condition at $\rho = 3$, which needs a gap of 6. The Clopper–Pearson pair at $\alpha/(10 \cdot 2)$ gives $L = 0.99901$, $U = 0.00168$ and a gap of 6.03, so the staged plug-in clears it; the Krichevsky–Trofimov sequence clears it too, but only at $n = 9{,}439$ of the 10^4 available; the empirical-Bernstein sequence does not cross inside the 10^4-draw budget at all: its gap is 5.93 at $n = 10^4$ and exceeds 6 by $n = 12{,}000$. The exact binomial tail is simply tighter than a betting confidence sequence at the boundary of the parameter space, where the mixture's regret term is what the sequence pays and the tail pays nothing. Inverting a mean bound therefore carries a budget ceiling here that a test of the certificate itself does not – a statement about this sequence, this stake and this budget, not about mean-CS inversion in general: E-Cert-R certifies 18.3% and 37.4% at the same radius. The effect appears only once ρ is large enough that the required gap approaches what $n = 10^4$ can supply at all: at $r = 0.5$ ($\rho = 2$, gap 4) the anytime-valid arms are within 0.9 points of the best plug-in, and at $r = 1$ ($\rho = 4$, gap 8) nothing certifies.

The fixed-budget row sits at or above every other arm at every radius, by at most 0.4 points, and the margin is widest where the certificate is marginal ($r = 0.75$: 18.9 against the staged min-test's 18.5 on CIFAR-10 and 37.8 against 37.7 on ImageNet). The fixed-budget procedure spends the full α on one pair of Clopper–Pearson bounds at $n = 10^4$, while the staged version divides it across ten checkpoints. E-processes retain validity at every stopping time and reach median stops of 19–23 evaluations, a benefit not captured by terminal accuracy alone.

The same sweep against the single-class target. Table 33 above and Table 31 vary the radius for the multiclass target; § G does it for the single-class one, and no table puts the two side by side. Figure 11 does, for E-Cert-R, over the radii the three sweeps share. Every curve, like both tables, spends the full α at each radius, so a vertical gap is not the reporting protocol. The near-zero point is evaluated at $\rho = 10^{-3}$, that is $r = 2.5 \times 10^{-4}$, since the VC design search divides by $\kappa = \sqrt{e^{\rho^2} - 1}$. There the multiclass condition is that the smoothed classifier is correct. As $\rho \downarrow 0$ the NP and VC population certificates both tend to the mean, so on a stream fixed at 1 each tends to 1, while the discretized CDF keeps a one-cell loss and reads 0.990. Two things separate the drawn point from that limit and neither is the level: the anchor sits at $\rho = 10^{-3}$ rather than at 0, where NP and VC need not equal the mean exactly, and a working law on the uniform grid represents that stream by its top cell centre 0.995, which is a property of the design grid rather than of the certificate.

Level-matched, the multiclass target is ahead or level in every cell up to $r = 0.5$ and the apparent crossings at $r = 0.75$ mostly disappear: hardmax ties to within 0.1 pp and softmax CDF turns positive (+0.7 and +1.3 pp). Softmax NP at $r = 0.75$ was the one cell that read the other way, and the cause was ours rather than the target's: the multiclass path evaluated on the uniform design grid where the single-class sweep applies the rule of § B.2, so at $\rho = 3$ its working law could not leave the NP null and the arm returned a structural zero (Prop. 5(i)) in place of a measurement. With the rule applied on both sides the multiclass arm certifies 6.0 and 6.3 pp against the single-class 5.9 and 5.5,

Table 33: Certified accuracy (%) of the multiclass hardmax procedures of § I.5 against the perturbation radius, $\sigma = 0.25$, $n = 10^4$, 1,000 images per dataset; each radius is a separate prespecified decision at α. The three staged rows (Bonferroni, min-test, CPM) spend α/m at each of $m = 10$ checkpoints, except that CPM is charged only the 6 checkpoints beyond its $n_0 = 100$ initial phase; the two CS rows and both E-Cert-R rows are inspected after every draw and pay no checkpoint factor. The last row is the same min-test bound spent in full, one decision at $n = 10^4$ with the whole α, so its gap to the staged min-test row is the price of the ten checkpoints and nothing else.

		Perturbation radius r							
Dataset	Method	0.0625	0.125	0.1875	0.25	0.375	0.5	0.75	1
CIFAR-10	Bonferroni plug-in	72.6	68.8	64.2	60.1	49.7	41.0	16.2	0.0
	Min-test plug-in	72.6	68.9	64.2	60.2	49.8	41.1	18.5	0.0
	CPM plug-in	72.6	68.9	64.1	60.2	49.6	41.0	18.6	0.0
	CS plug-in	72.4	68.1	63.7	59.3	49.3	40.2	0.0	0.0
	CS plug-in (KT)	72.6	68.6	63.9	59.9	49.5	40.4	16.3	0.0
	E-Cert-R (no refit)	57.9	53.9	48.7	44.4	36.7	28.1	9.6	0.0
	E-Cert-R	72.6	68.7	63.9	59.8	49.4	40.2	18.3	0.0
	Min-test plug-in (full budget)	72.6	68.9	64.3	60.2	49.8	41.1	18.9	0.0
ImageNet	Bonferroni plug-in	67.6	64.3	62.3	60.8	56.2	51.3	35.4	0.0
	Min-test plug-in	67.6	64.3	62.5	61.2	56.5	51.9	37.7	0.0
	CPM plug-in	67.6	64.3	62.5	61.2	56.5	51.7	37.7	0.0
	CS plug-in	67.2	63.9	62.3	60.3	56.3	51.2	0.0	0.0
	CS plug-in (KT)	67.6	64.3	62.3	60.7	56.2	51.2	35.4	0.0
	E-Cert-R (no refit)	31.3	26.7	23.1	19.2	14.4	9.4	2.6	0.0
	E-Cert-R	67.6	64.3	62.4	60.7	56.1	51.6	37.4	0.0
	Min-test plug-in (full budget)	67.6	64.3	62.5	61.2	56.5	51.9	37.8	0.0

which is the direction the null geometry predicts: the two conditions differ by $1 - p_A - p_B$, the mass on the non-runner-up classes, which is non-negative and shrinks to zero as $\Phi(\rho) \to 1$.

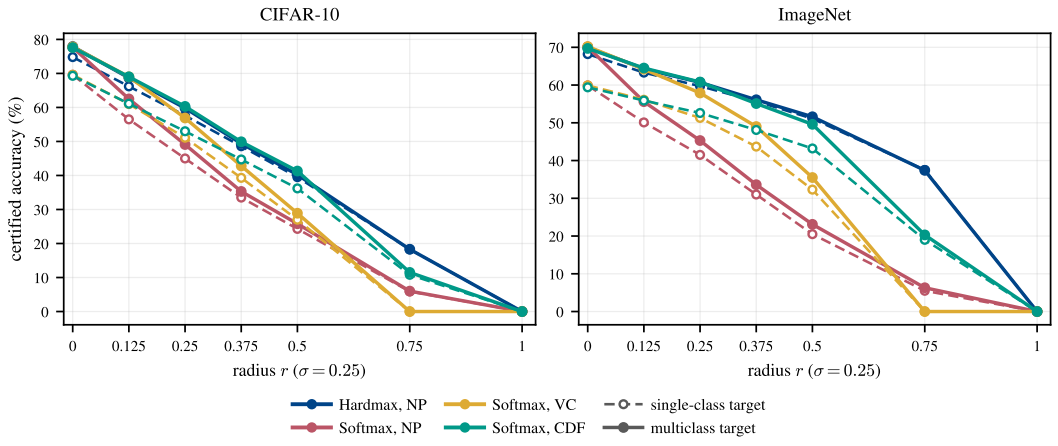


Figure 11: E-Cert-R alone, against the perturbation radius, $\sigma = 0.25$, $n = 10^4$, 1,000 images per dataset. Solid is the multiclass target and dashed the single-class one. Every curve is the same arm, so colour indexes the (observable, certificate) cell rather than the method here. All eight curves are at level α per radius. The multiclass intercepts at $r = 0$, 77.6–77.9 and 69.7–70.2, are the smoothed classifier's own accuracy (78.2 and 70.4) less the cost of proving it at $\alpha = 10^{-3}$. The zeros differ in kind: at $r = 0.75$ the multiclass softmax VC arm certifies 0 of 1,000 images, while NP certifies 6.0/6.3% and does so only on the threshold-aware grid of § B.2; by $r = 1$ every arm is at zero, where the stake cap puts any NP crossing past 2×10^5 draws.

Table 34: CPM swept over its initial-phase size n_0, with every other arm unchanged. "Partition match" is the fraction of images whose initial-phase argmax is the true label. †: median over the images every n_0 variant and E-Cert-R all certify, so the column is comparable across the sweep. $\star$ marks the value we credit CPM at in the main text; $n_0 = 100$ is the value Delattre et al. (2025) use.

Dataset	n_0	CA	Partition match	Median stop†	Ratio to E-Cert-R
CIFAR-10	$10^\star$	0.683	76.8	50	2.5
	25	0.686	78.2	100	5.0
	50	0.688	78.3	100	5.0
	100	0.689	78.2	250	12.5
	200	0.689	78.4	250	12.5
	400	0.688	78.4	500	25.0
ImageNet	$10^\star$	0.642	69.9	50	2.2
	25	0.642	70.0	100	4.3
	50	0.642	70.5	100	4.3
	100	0.643	70.3	250	10.9
	200	0.643	70.3	250	10.9
	400	0.643	70.2	500	21.7

I.15 Sensitivity to the CPM initial-phase size

The staged-grid control evaluates each baseline at its best tested monitoring grid. We also vary CPM's initial-phase size n_0, which is fixed at 100 in the default benchmark. That choice is faithful, since it is the value Delattre et al. (2025) use in their own experiments. It is not necessarily the best value *here*, and it has a mechanical consequence. With 100 draws spent before estimation begins, CPM cannot certify at or before the $n = 100$ checkpoint, so its first feasible certification is pinned to the next one. Sensitivity to n_0 must therefore be assessed when reporting relative stopping costs.

Table 34 sweeps $n_0 \in \{10, 25, 50, 100, 200, 400\}$, changing nothing else: every other arm is read from the main benchmark unchanged, and the self-check asserts that $n_0 = 100$ reproduces the benchmark's own CPM computation. CPM's median stopping cost decreases as n_0 decreases over the tested range, and is lowest at the smallest tested value. Its median stop falls from 250 at $n_0 = 100$ to 50 at $n_0 = 10$ on both datasets, while its certified accuracy moves by 0.6 pp on CIFAR-10 ($0.689 \to 0.683$) and 0.1 pp on ImageNet ($0.643 \to 0.642$). Crediting CPM there, our lead over it drops from $7.0\times$ and $6.7\times$ to $\mathbf{2.5}\times$ and $\mathbf{2.2}\times$ (50 against 20 and 23). § 7.2 credits CPM at this best n_0.

A smaller pilot lets CPM begin estimation earlier. Over the tested range, the partition-match rate decreases gradually: certified accuracy is 0.683 at $n_0 = 10$, against 0.689 at the default. Because $n_0 = 10$ is the smallest tested value, smaller pilots may improve CPM further. The reported speedups apply only to the tested configurations. Pilot size and monitoring grid are varied separately, holding the other at its default; their joint optimization has not been evaluated.

I.16 Target selection: removing the protocol asymmetry

The benchmark of § I.5 provides the ground-truth label to our procedures and to the Bonferroni plug-in, while CPM and Cohen CERTIFY identify the target class from an initial n_0 draws. That is the convention of the certified-accuracy literature, but it is an asymmetry in our favor and we assess its effect using two controls: a shared target-selection phase for all procedures, and ground-truth targets for all procedures.

In the *selection* variant every procedure spends a shared prefix of n_0 draws choosing $\widehat{A}$ as the empirical argmax of the class counts. It then runs its own test on the remaining stream with target $\widehat{A}$, and is credited only when $\widehat{A}$ is the true label. Cost is counted from the first draw, so the prefix is charged to everyone. CPM uses this same n_0 for its selection and grouping pilot, rather than a separate fixed 100-draw prefix. Validity is untouched: $\widehat{A}$ is measurable with respect to the first n_0 draws while the e-process consumes only later ones, so conditionally on $\widehat{A}$ it remains an e-process for "$\widehat{A}$ is not the argmax". We sweep n_0 rather than fixing it at 100, because a prefix of 100 over-charges

Table 35: Median *total* evaluations to certify, counted from the first draw, when a shared n_0-draw target-selection prefix is charged to every procedure, and (last column) when CPM and Cohen are handed the true target instead. Medians are over the images all seven procedures certify in that setting (534–537 CIFAR-10, 267 ImageNet).

		Shared selection prefix n_0				Oracle
Dataset	Method	10	25	50	100	target
CIFAR-10	Bonferroni plug-in	50	100	100	250	50
	Min-test plug-in	50	100	100	250	50
	CPM plug-in	50	100	100	250	250
	CS plug-in	48	63	88	138	38
	CS plug-in (KT)	37	52	77	127	27
	E-Cert-R (no refit)	30	45	70	120	20
	E-Cert-R	29	44	69	119	19
	Cohen CERTIFY	50	50	100	250	25
ImageNet	Bonferroni plug-in	50	100	100	250	50
	Min-test plug-in	50	100	100	250	50
	CPM plug-in	50	100	100	250	250
	CS plug-in	48	63	88	138	38
	CS plug-in (KT)	37	52	77	127	27
	E-Cert-R (no refit)	32	47	72	122	22
	E-Cert-R	30	45	70	120	20
	Cohen CERTIFY	50	50	100	250	25

a sequential test whose own median stop is around 20. In the *oracle* variant CPM and Cohen instead receive the true target: CPM still builds its bucket partition from its initial phase, but around the correct class, and Cohen skips selection and spends its whole budget on the bound.

Two findings emerge. First, certified accuracy changes little under the target-selection controls. Providing CPM and Cohen with the true target leaves their accuracies unchanged to within one image ($0.689 \to 0.689$ and $0.663 \to 0.663$ on CIFAR-10; $0.643 \to 0.643$ and $0.634 \to 0.633$ on ImageNet). Adding a selection prefix to our methods has little effect: E-Cert-R stays within three images of 0.687 at $n_0 = 10$, and within one image of 0.687 and 0.643 at every other n_0. The reason is the stratification already reported above: the images on which selection fails are, at $r = 0.125$, essentially the images no procedure certifies, so the target-selection protocol has little observed effect on the accuracy comparison.

Second, the stopping-cost advantage survives the correction, though its size moves with n_0. Charging everyone the same prefix (Table 35), E-Cert-R leads the Bonferroni plug-in by 1.4–$2.3\times$, and the min-test plug-in and CPM by the same 1.4–$2.3\times$, across $n_0 \in \{10, 25, 50, 100\}$ on both datasets, against 1.6–$2.1\times$ and 6.7–$7.0\times$ in the own-best-grid comparison of § I.13. The two CS plug-ins shrink with the prefix rather than jumping to the next checkpoint. Their margins are therefore smaller, and fall as n_0 grows: 1.2–$1.7\times$ for the empirical-Bernstein arm and 1.1–$1.3\times$ for the KT arm, the latter reaching $1.06\times$ at $n_0 = 100$, where a 100-draw prefix is four times either arm's own oracle stop and the prefix dominates the comparison. The oracle CPM column still pays for its 100-draw grouping pilot, with the competitor partition rebuilt around the supplied target; it is not a pilot-free oracle. The advantage persists because the staged baselines stay checkpoint-quantized. A selection prefix reduces their early estimation budget and pushes their first feasible certification to the next checkpoint, whereas an e-process is credited at its exact crossing. The weakest cell is $n_0 = 50$, where the Bonferroni plug-in reaches the 100-draw checkpoint while E-Cert-R needs 69–70.

I.17 Variability across draw orders and independent Monte Carlo replicates

Draw order in the multiclass panels. The single-class columns of Figure 3 average ten draw-order permutations and carry ± 1 SD bars; the multiclass columns read one cached order and carry none, so a reader is entitled to ask what the missing replication would have shown. We recompute the two arms cheap enough to replicate exactly, the min-test and Bonferroni plug-ins, on twenty permuted copies of the same class streams, permuted per image the way the single-class panels permute theirs. Certified

Table 36: Across-replicate mean and standard deviation over four independent Gaussian-noise replicates on a fixed 200-image subset per dataset (replicate one is the paper's cache restricted to that subset). The image set, labels, and clean softmax vectors are held identical; only the noise is redrawn. Median stop is taken over each method's own successes.

Dataset	Method	CA (mean $\pm$ SD)	Median stop (mean $\pm$ SD)
CIFAR-10	Bonferroni plug-in	0.701 ± 0.003	50.0 ± 0.0
	Min-test plug-in	0.701 ± 0.003	50.0 ± 0.0
	CPM plug-in	0.701 ± 0.003	250.0 ± 0.0
	CS plug-in	0.690 ± 0.004	38.0 ± 0.0
	CS plug-in (KT)	0.698 ± 0.003	27.0 ± 0.0
	E-Cert-R (no refit)	0.560 ± 0.000	20.0 ± 0.0
	E-Cert-R	0.700 ± 0.004	20.0 ± 0.0
ImageNet	Bonferroni plug-in	0.655 ± 0.000	50.0 ± 0.0
	Min-test plug-in	0.655 ± 0.000	50.0 ± 0.0
	CPM plug-in	0.655 ± 0.000	250.0 ± 0.0
	CS plug-in	0.655 ± 0.000	38.0 ± 0.0
	CS plug-in (KT)	0.655 ± 0.000	27.0 ± 0.0
	E-Cert-R (no refit)	0.275 ± 0.000	22.0 ± 0.0
	E-Cert-R	0.655 ± 0.000	22.0 ± 0.0

accuracy is invariant: it is constant across all twenty permutations in three of the four dataset–arm cells and moves by a single image in the fourth. That is what the certificate structure predicts, since the terminal checkpoint sees the whole cached multiset and only a lucky early checkpoint can depend on order. Mean cost over all images spans 0.94% of its mean (SD 0.22%), and the deployed order sits inside that spread in every cell. The gaps the figure is read for are three to two hundred times larger, so order is not what separates the arms. The e-process arms would need their capital paths replayed and are not covered here.

Fresh Gaussian draws. Every "seed" above permutes a single cached score multiset, so the reported standard deviations quantify sensitivity to score *order* conditional on the draws, not uncertainty across fresh draws. That limitation of a cached-stream protocol can be removed only by resampling, so we took independent noise replicates. The sampler separates the seed that selects images from the seed that drives the Gaussian draws, so a replicate holds the image subset exactly fixed and redraws only the noise, and the default equal-seed call reproduces the paper's existing cache bit-for-bit. Table 36 reports four replicates on a shared 200-image subset per dataset.

At $n = 10^4$ this source of uncertainty is negligible: certified accuracy has an across-replicate standard deviation of 0.004 or less on CIFAR-10 and is *exactly* constant on ImageNet, and the median stopping cost is identical across replicates for every method on both datasets. For $p \approx 0.70$ at 200 images the binomial standard error from *which images were sampled* is 0.032, about eight times the largest replicate standard deviation, so image sampling rather than Monte Carlo noise dominates the uncertainty in every certified-accuracy number here. This holds at the terminal budget; small-budget behavior is necessarily noisier.

J LIMITATIONS AND BROADER IMPACT

The theory assumes that the score stream obeys the stated conditional laws and that every tuning decision is made from the past. Violations of those assumptions, dependence induced by reusing randomness, distribution shift, and simultaneous post-hoc selection across inputs, radii, targets and certificate families each require separate analysis. An inaccurate working law is not among them: it bears on efficiency, not validity. RIPr log-growth optimality applies to the exact finite-alphabet construction under its stated regularity conditions. For continuous-score witnesses, § B.4 gives an oracle stopping-time bound, but neither growth optimality nor a bound for the fitted design. Our implementation uses full-support Dirichlet regularization for the working histogram, whose pseudocount is a design choice with no bearing on validity (§ B). Regardless of regularization, validity requires checking the relevant null-side constraint and factor nonnegativity over the complete score support.

Certified accuracy depends on both the target and the procedure. The single-class number depends on whether the target is the mean softmax score or the hard-label probability, and on these checkpoints the second is 4.6–5.3 points easier. A soft-score certified accuracy therefore reports in part how a base classifier spreads its softmax mass (Table 22). Comparisons in this paper are therefore always within one target.

Most cost comparisons count model evaluations. The actual online experiment in § D.3 additionally charges sampling, transfers, batch overshoot and analysis, and there the ordering can reverse: CDF projection overhead outweighs the sampling it saves, and on the much cheaper NP and VC projections RIPr still does not consistently beat the cheap witness. On the hard label the same measurement puts our arm 0–6% behind a staged Clopper–Pearson rule at equal certified accuracy (§ D.4). Fewer model evaluations therefore do not establish a deployment speedup, and these timings are specific to the batching and monitor implementations tested; our prefix-CS recomputes its VC/CDF bounds at chunk endpoints. Numerical CDF projection failures use a validity-preserving neutral factor; all affected runs remain included in the reported results.

At the default radius the equal-cap single-class controls find only modest incremental gains from RIPr, with occasional reversals across noise seeds. The same controls at $\rho = 3$ show that this is regime-dependent: there the CDF projection separates an arm that certifies from one that does not, while the VC control stays flat (§ F.1).

Order-seed variability in Table 16 measures sensitivity to permutations of fixed cached score multisets, not independent noise replication. Other experiments vary the noise draws, the smoothing level, the evaluation budget and the base classifier's training method: § H.9 and § I.17 redraw the Gaussian noise on a fixed image subset and find the Monte Carlo spread negligible at $n = 10^4$; § G varies the noise level with its own checkpoint and score cache per σ; and § H.7 repeats the single-class benchmark on consistency-regularized checkpoints, leaving the cost ratios between procedures in similar ranges. Variation across architectures and across training seeds of one method remains unassessed: apart from that appendix, every number comes from a single released checkpoint per dataset and noise level, and the compared training methods share an architecture within each dataset.

The multiclass extension of § I has an additional limitation: the intersection–union guarantee requires every pairwise e-process in the family to be valid and does not by itself address selecting the target class from the same noisy stream without correction.

Anytime-valid monitoring can reduce unnecessary model evaluations and can make robustness claims less sensitive to arbitrary terminal budgets. Conversely, the language of "anytime validity" may create false reassurance if the score stream is adaptively altered, if many targets are searched without correction, or if a population RS certificate is mistaken for a guarantee about all real-world shifts.

K ADDITIONAL RELATED WORK

Randomized smoothing: origins and training. Early smoothing certificates used connections to differential privacy (Lecuyer et al., 2019) and additive-noise robustness bounds (Li et al., 2019). Cohen et al. (2019) derived the tight Gaussian ℓ_2 radius. Subsequent work improved the base classifier through adversarial training of the smoothed model (Salman et al., 2019), direct optimization of the certified radius (Zhai et al., 2020), and consistency regularization across noise draws (Jeong &

Shin, 2020). Pretrained classifiers can also be adapted by prepending a denoiser (Salman et al., 2020) or a diffusion model (Carlini et al., 2023). These changes to the classifier are compatible with the statistical certification procedures studied here.

Certificate shape, other threat models, and limits. Dvijotham et al. (2020) use f-divergences to handle general smoothing measures, and Yang et al. (2020) give a framework for designing smoothing distributions. Certificates can incorporate first-order information (Mohapatra et al., 2020) or combine guarantees at nearby points (Cullen et al., 2022). For structured inputs, Scholten et al. (2023) smooth random subsets of entities; for patch attacks, Levine & Feizi (2020) replace random sampling with a deterministic certificate.

Input-dependent noise levels (Alfarra et al., 2022) introduce additional soundness requirements. Súkeník et al. (2022) identify gaps in existing guarantees and a curse of dimensionality, while also deriving conditions under which input-dependent smoothing is valid. More generally, the certifiable ℓ_p radius decreases with dimension for $p > 2$ (Kumar et al., 2020b; Blum et al., 2020). Smoothing can shrink decision regions and create class-wise accuracy disparities (Mohapatra et al., 2021), and finite-precision implementations need additional care to remain sound (Voráček & Hein, 2023). Li et al. (2023) survey and benchmark these methods. Our experiments address the statistical estimation step for NP, VC and CDF certificates. Extending the construction to other sampled certificates requires a witness for the corresponding null; deterministic certificates have no Monte Carlo estimation step to replace.

Reducing the sampling cost of certification. Chen et al. (2022) allocate samples by input difficulty and apply a fixed-n confidence bound at the chosen budget. Horváth et al. (2022) reduce sampling cost at a prescribed radius by lowering the variance of the classifier's predictions. Ugare et al. (2024) reuse a certificate for the original model when certifying an approximation to it. Li et al. (2022) tighten certificates using samples from a second smoothing distribution, and Delattre et al. (2025) construct less conservative confidence intervals.

Voráček (2024) uses confidence sequences to make sequential certification decisions for the hard-label probability. We use their Krichevsky–Trofimov sequence in the hardmax benchmark (§ I.5). Their comparisons include Clopper–Pearson intervals, a union-bound confidence sequence and the staged procedure of Horváth et al. (2022); our hardmax benchmark also includes an empirical-Bernstein betting sequence. These confidence-sequence methods are anytime-valid. They bound the smoothed quantity and substitute the bound into a certificate. Our VC and CDF constructions instead test the certificate's null directly, using a joint witness to avoid separate bounds on its components (§ 7.1).

Cullen et al. (2026) learn an image-specific betting prior for an e-process on the hard-label probability. They report roughly a twentyfold reduction in sampling relative to a fixed budget of 10^4, with an anytime-valid lower confidence bound. Their stopping rules use the gap between estimated and certified radii, wealth bankruptcy, an upper-confidence-bound exit, and a plateau in the certified radius. The UCB exit requires both $W_t(1/2) \geq 1/\alpha$ and $\widehat{p}_t < 1/2$, which they note implies an upper confidence bound below $1/2$. When the true hard-label probability p exceeds $1/2$ that exit places p outside their confidence set, so their anytime coverage bounds it by α; they do not state this for the exit rule itself.

For the single-class NP certificate at normalized radius $\rho > 0$ and threshold $\tau = 1/2$, the alternative is $p > \Phi(\rho) > 1/2$, so that UCB exit already protects every input in this alternative. When $1/2 < p \leq \Phi(\rho)$, the same exit still has probability at most α, although these inputs do not satisfy the requested certificate. Our population futility rule uses the requested certificate threshold, allowing early abstention on this interval while bounding erroneous abstention when the population certificate exceeds that threshold (Prop. 3). We also derive rules for VC, CDF and multiclass certificates. For their radius-gap, bankruptcy and plateau rules, anytime coverage preserves the validity of any certificate they report, but does not by itself bound the probability of abandoning an input certifiable at the requested radius; they note that the bankruptcy filter can produce rare false negatives near the decision boundary. Their analysis concerns the hard-label probability, without the VC/CDF score functionals or our intersection–union treatment of the argmax.

Certificate inversion and regression. Inverting the tests over certificate thresholds is a possible route to confidence sequences for the certificate value (Howard et al., 2021; Waudby-Smith & Ramdas,

2024); we study decisions at a fixed threshold. Regression requires a choice of smoothed target and a corresponding certificate. Chiang et al. (2020) use median smoothing to certify regression outputs and object detectors. Applying sequential tests to that setting would require matching the tested null to its median-based certificate; we do not establish that extension here.

Statistical verification of neural networks. Monte Carlo verification also arises outside randomized smoothing. Webb et al. (2019) estimate property-violation probabilities by multilevel splitting. Weng et al. (2019) give probabilistic certificates under a distributional perturbation model, and Tit et al. (2021) formulate local robustness as a hypothesis test with non-asymptotic sample-size guarantees. Baluta et al. (2021) use sequential tests to verify probabilistic properties of a fixed network. The object being tested in our work is a population smoothing certificate.

Anytime-valid inference and e-values. Confidence sequences provide intervals that are valid uniformly over time (Darling & Robbins, 1967; Robbins, 1970). Nonnegative supermartingales and time-uniform Chernoff bounds supply general constructions (Howard et al., 2020; 2021); Ramdas et al. (2020) establish the role of nonnegative martingales in admissible anytime-valid inference. E-values, their calibration and combination, and their use under optional continuation are studied by Vovk & Wang (2021), Grünwald et al. (2024) and Shafer (2021). Grünwald (2024) discusses reporting e-values in place of p-values. Ramdas et al. (2023) and Ramdas & Wang (2025) survey this literature, and Wasserman et al. (2020) construct finite-sample tests from likelihood ratios without the usual regularity conditions.

Our RIPr designs use results on log-optimal e-values for composite nulls (Koolen & Grünwald, 2022) and reverse information projection (Lardy et al., 2024; Larsson et al., 2025). For bounded means, our comparators use the betting confidence sequences of Waudby-Smith & Ramdas (2024). Related analyses use universal portfolios (Orabona & Jun, 2024) and study near-optimal sequential tests (Shekhar & Ramdas, 2023). Manole & Ramdas (2023) develop sequential estimators of divergences, including KL divergence. The need for validity under continuous monitoring also appears in A/B testing, where repeated inspection of fixed-sample tests is known as the "peeking" problem (Johari et al., 2017; 2022).

Sequential Monte Carlo testing and adaptive stopping. Gandy (2009) sequentially approximates a Monte Carlo p-value while controlling the probability of a different decision from the test based on the exact p-value. Fischer & Ramdas (2025) construct betting procedures that allow Monte Carlo resampling, including permutation tests, to stop at any time. These methods control error in a resampling test; our witnesses test a smoothing certificate. Mnih et al. (2008) use empirical-Bernstein bounds to stop sampling once a prescribed precision is reached. Our fixed-n comparators use bounds from the same literature (Maurer & Pontil, 2009; Audibert et al., 2009). The population futility rules use Hoeffding's inequality (Hoeffding, 1963) and the DKW inequality (Dvoretzky et al., 1956) with Massart's constant (Massart, 1990).

Futility and curtailment. Futility stopping is standard in group-sequential clinical trials (Jennison & Turnbull, 1999). Stochastic curtailment (Lan et al., 1982) and conditional-power rules (Lachin, 2005) use the probability of the final decision given the data collected so far. Our population rule uses an upper confidence bound on the certificate and has the protection guarantee in (6). The slope heuristic extrapolates recent capital growth without such a guarantee; § H.4 measures the certificates it discards.

Intersection–union tests. An intersection–union test rejects a union of sub-nulls exactly when every sub-null is rejected at level α (Berger, 1982; Berger & Hsu, 1996). It needs no multiplicity correction across these component tests. Our argmax min-test applies this construction to the $C-1$ pairwise nulls (§ 6). This gives the guarantee in Prop. 4 without dividing α by $C-1$ and also justifies the min-test plug-in in § 7.2.