# Recertifying decisions via e-closure: uniform improvements of conformal risk control and randomized e-BH

Tian Bai and Aaditya Ramdas

Stanford University
{tianbai,aramdas}@stanford.edu

**Abstract**

Statistical procedures often control expected loss at a prescribed level while leaving part of their risk budget unused. Building on the e-value representation of risk-controlled decisions and the e-closure principle for multiple testing, we formulate *recertification*, a general paradigm for improving procedures by exploiting computable slack in their validity certificates. In particular, a valid procedure's losses normalized by their risk budgets form a state-indexed e-value collection; conversely, decisions whose normalized losses are bounded by this collection inherit its risk guarantees. Recertification uses distributional structure to quantify slack, enlarge the certifying e-values while preserving validity, and construct less conservative decisions.

We develop two applications. For conformal risk control (CRC) with a known loss family, recertification redistributes unused risk budget among the held-out points of each hypothetical augmented sample. The resulting controller is valid under every finite exchangeable law and is no more conservative than CRC on any calibration sample. For convex losses on an interval of actions, repeated recertification converges geometrically to an admissible controller. For false discovery rate control with arbitrarily dependent e-values, recertification exploits unused conditional budgets in randomized e-BH. It yields a closed-form randomized procedure whose rejection set contains that of randomized e-BH for every realization of the shared randomization and whose expected number of rejections is maximal among randomized procedures satisfying the same conditional e-value budget constraints. Recertifying instead with the larger e-collection of closed e-BH gives a second closed-form procedure, whose rejection set contains that of the first. Experiments on CIFAR-100 hierarchical classification and multiple testing simulations demonstrate improved budget utilization and gains in informativeness and power over the corresponding baselines.

***Keywords:*** e-values, conformal risk control, false discovery rate, multiple testing, exchangeability, admissibility.

## 1 Introduction

Many statistical procedures turn data into an informed action, while guaranteeing that some expected loss stays below a prescribed level. Statistical decision theory formalizes this task through a decision rule mapping data to actions, a loss depending on the action and an unknown state of nature, and the resulting statewise expected loss, or risk [Wald, 1950, Berger, 1985]. The bound imposes a risk budget that must hold in every state. Among rules satisfying this constraint, one seeks the most useful decisions, as in the constrained problems of Hodges and Lehmann [1952].

Multiple testing is a decision problem of exactly this kind. With $K$ hypotheses, the action is a rejection set $R \subseteq [K]$, so the action space is the collection $2^{[K]}$ of all subsets of $[K]$, and the unknown state is the set $T \subseteq [K]$ of true null hypotheses. The false discovery proportion $|T \cap R|/(|R| \vee 1)$ is the loss, and false discovery rate (FDR) control bounds its expectation by $\alpha$. The procedure of Benjamini and Hochberg [1995] achieves this by comparing the ordered p-values with linearly increasing thresholds; its e-value analogue,

the e-BH procedure of Wang and Ramdas [2022], applies the reciprocal thresholds to e-values, which are nonnegative statistics with expectation at most one under their null hypotheses [Vovk and Wang, 2021, Ramdas and Wang, 2025]. The recent work of Xu et al. [2026b] introduces the e-closure principle, a general principle that uniformly improves many testing procedures based on e-values, including e-BH. Specifically, given an e-value $E_T$ for each candidate true-null set $T$, valid when $T$ is the set of true nulls, the principle certifies every rejection set $R$ that satisfies

$$\frac{|T \cap R|}{|R| \vee 1} \leq \alpha E_T \qquad \text{for every } T \subseteq [K]. \tag{1.1}$$

Under the actual null set, the left side is the false discovery proportion and the right side has expectation at most $\alpha$. The false discovery rate is therefore controlled even when the analyst chooses among the certified rejection sets after seeing the data. Conversely, Xu et al. [2026b] show that a procedure controls the false discovery rate simultaneously over the rejection sets it may report if and only if all of them are certified in this way by some family of e-values. The principle applies equally to other error rates that are expected losses, and it contains the classical closed testing principle for the familywise error rate [Marcus et al., 1976] as the case of e-values taking only the values 0 and $1/\alpha$.

Nothing in this argument depends on the actions being rejection sets. In a general decision problem, a state $s$ specifies a class of distributions for the data $X$, an action $a$ is chosen from an arbitrary set, and each of several losses $L_t(s, a; X)$, indexed by $t$, carries its own budget $\alpha_t(s)$. A family of e-values, one for each state and risk, certifies the actions whose losses fit within all the budgets it allows. Every data-dependent choice among the certified actions then controls every risk, and every procedure with this guarantee is certified by some such family [Grünwald, 2023, 2024, Andrews and Chen, 2025]. In Section 2 we review the e-closure principle at this level of generality, with menus of arbitrary actions, possibly continuously indexed states and risks, and losses that may depend on the data.

The correspondence can be read in both directions, and the direction from procedures to e-values suggests how to improve a procedure. Improving a decision rule directly is difficult: a more liberal rule may have higher power or utility, but it is rarely clear how far a given rule can be pushed before its guarantee fails. By contrast, improving an e-value is easier. If an e-value has expectation strictly below one, and the shortfall can be computed, for instance conditionally on a statistic whose conditional law is known under the null, then adding the shortfall back yields a larger e-value. Through the e-closure correspondence, larger e-values translate back into decisions. Starting from a valid procedure, we apply this construction to its canonical e-values: its losses divided by their risk budgets. Inverting the enlarged certificate yields a menu containing the original actions and potentially more liberal ones, all with the same guarantee. We call this step *recertification*. It generalizes the anchoring idea of Ramdas [2026], which keeps an existing false discovery rate procedure and spends the slack in its validity proof, and it can be repeated.

We apply recertification to two problems. The first is conformal risk control (CRC) [Angelopoulos et al., 2024], which extends conformal prediction from coverage to general monotone losses. Given $n$ exchangeable calibration points and a known family of losses indexed by a conservativeness parameter $\lambda$, CRC selects the least conservative $\lambda$ for which the calibration losses, together with a bound on the loss of the unseen test point, fit within the budget $(n+1)\alpha$. The selected controller typically leaves part of this budget unused. Recertification computes this unused budget on every such sample, redistributes it among the $n+1$ held-out positions, and inverts the resulting e-values. The new controller is valid under every finite exchangeable law, and for every calibration sample it is no more conservative than CRC. In this sense, for a fixed loss family and observation domain, it improves CRC uniformly. This recertification step can be repeated, with diminishing returns. For convex losses, the iterates converge geometrically to a controller that admits no further pointwise improvement, and on finite outcome spaces a coordinatewise sweep reaches such a controller without convexity (Section 3).

Figure 1 illustrates these gains in hierarchical classification on CIFAR-100 (Section 3.6), using rules that give more specific predictions when the classifier is more confident. At a target of 15%, CRC attains a mean test risk of 14.21%. Repeated recertification raises it to 14.95%, close to the target, and reduces the mean prediction-set size from 10.77 to 8.92 labels, a reduction of 17.2%.

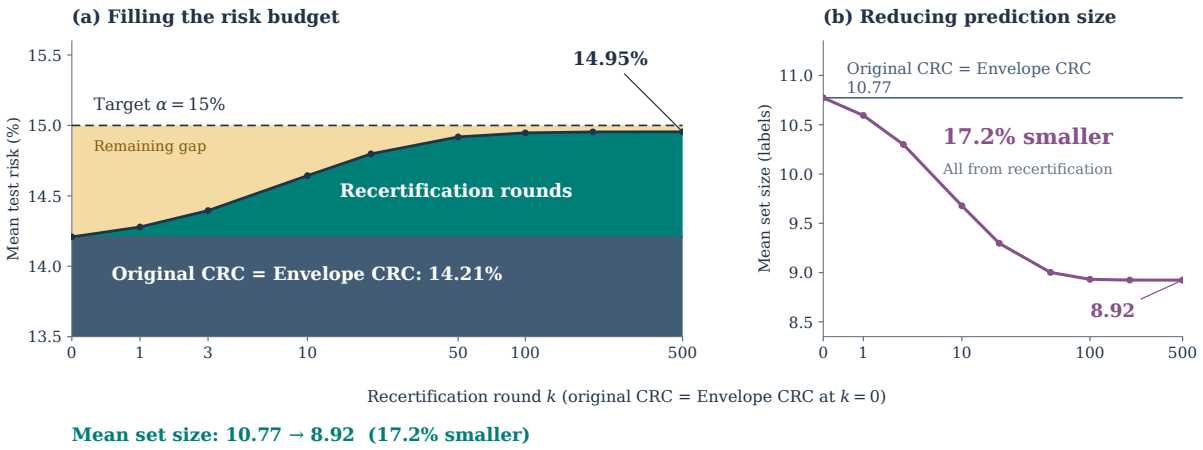



Figure 1: Risk-budget use and mean prediction-set size on CIFAR-100 with confidence-adaptive prediction rules ($n = 50$, $\alpha = 0.15$). Round 0 is CRC, and later rounds apply recertification. Results are averaged over 300 paired splits, each with 16 test images.

The second application returns to multiple testing. The e-BH procedure is valid under arbitrary dependence but can be conservative, and Xu and Ramdas [2023] showed that external randomization recovers part of the lost power. Their U-eBH procedure boosts the e-values by dividing them by a single uniform random variable and then applies e-BH. In this setting, we treat the conditional law of the rejection set given the e-values as a randomized action. In these terms, U-eBH is certified by the family of e-values behind e-BH, through a version of the self-consistency condition of Blanchard and Roquain [2008] that holds on average over the randomization, and it typically leaves part of the budget allowed by this family unused. Recertification spends this budget and yields a closed-form randomized rule, recertified U-eBH (RU-eBH), which rejects at least as many hypotheses as U-eBH for every realization of the randomization and has the largest expected number of rejections among all randomized procedures certified by the same family.

The same principle applies to closed e-BH, the improvement of e-BH by e-closure [Xu et al., 2026b], whose family of e-values is larger because it adapts to the number of true nulls. Recertifying with this family gives a second closed-form rule, recertified closed e-BH (RC-eBH), whose rejection set contains that of RU-eBH and which has the largest expected number of rejections among all randomized procedures certified by this family, closed e-BH included (Section 4). In simulations, both rules are more powerful than e-BH and U-eBH in every setting, and RC-eBH is the most powerful of the eight procedures we compare in nearly every setting.

## 1.1 Related work

**E-values and decisions.** E-values have become a standard tool for testing and multiple testing [Vovk and Wang, 2021, Ramdas and Wang, 2025]. Their link to decision making is developed by Grünwald [2024], who allows the significance level and the loss to be chosen after seeing the data, and by Koning [2023] and Chugg et al. [2026], who study post hoc testing and its admissible procedures. For a single risk, our certified menu is the set of actions whose e-posterior risk bound [Grünwald, 2023], with each state's loss divided by its budget, is at most one. Andrews and Chen [2025] study decisions reported together with certificates of their loss, and Koning and van Meer [2026] give evidence-weighted loss bounds for decisions based on conformal e-prediction sets. Decision-theoretic views of conformal prediction also study how prediction sets should inform downstream actions [Lekeufack et al., 2024, Kiyani et al., 2025]. These works allow general states and actions; our focus is the constructive use of the correspondence, which keeps the error budget fixed and recertifies an existing procedure.

**Closure principles.** Closed testing [Marcus et al., 1976] controls the familywise error rate, and every coherent procedure for that error rate is a closed testing procedure [Sonnemann, 2008]. Closed testing also yields simultaneous bounds on the false discovery proportion of every rejection set [Goeman and Solari, 2011], and only closed testing procedures are admissible for that purpose [Goeman et al., 2021]. E-closure [Xu et al., 2026b] extends the closure principle to error rates defined by expected losses. Its follow-ups include the uniform improvement of the Benjamini–Hochberg procedure [Goeman, 2026], anchoring [Ramdas, 2026], online versions [Xu et al., 2026a], and complete-class and admissibility results for e-value procedures [Sun and Wang, 2026]. We note that representation and admissibility are separate questions. Every valid procedure is certified by some family of e-values, while admissibility requires further conditions, and our admissibility results for CRC rely on the structure of that problem.

**Randomized multiple testing.** Randomization is classical in testing, where it makes tests based on discrete statistics exact [Lehmann and Romano, 2022]. For false discovery rate control with e-values under arbitrary dependence, Xu and Ramdas [2023] showed that stochastic rounding of the e-values and U-eBH increase the power of e-BH, and Xu et al. [2026b] combine stochastic rounding with e-closure. These methods randomize the e-values. Our procedures instead choose the distribution of the rejection set directly, subject either to per-hypothesis budgets that form a randomized version of the self-consistency condition of Blanchard and Roquain [2008] or to the constraints that certify closed e-BH. Other uniform improvements of e-BH include its minimally adaptive version [Ignatiadis et al., 2024], closed e-BH [Xu et al., 2026b], and conditional calibration [Fithian and Lei, 2022, Lee and Ren, 2024], which recalibrates the e-values given a sufficient statistic.

**Conformal risk control.** CRC already handles several, even infinitely many, risks that share a monotone action parameter [Angelopoulos et al., 2024, Section 4.3]. Risk-Controlling Prediction Sets (RCPS) and Learn then Test (LTT) instead provide high-probability statements about population risks [Bates et al., 2021, Angelopoulos et al., 2025]. Extensions of CRC treat non-exchangeable data [Farinhas et al., 2024], cross-validation [Cohen et al., 2024], non-monotonic losses [Angelopoulos, 2026], and end-to-end training [Yeh et al., 2025]. The leave-one-out characterization of validity that we use (Lemma 3.1) is the risk-control counterpart of the universality of conformal prediction [Vovk et al., 2022, Proposition 2.9]; see also Tibshirani et al. [2026], and Angelopoulos [2026] for its sufficiency direction. Conformal prediction with e-values is developed by Vovk [2025], Koning [2026], Gauthier et al. [2025], and Koning and van Meer [2026]. Apart from the sharper bound on the unseen test loss in envelope CRC [Yeh et al., 2025], which serves as our starting point, we are not aware of work on improving CRC controllers pointwise or on their admissibility. We use CRC to show concretely how large such improvements can be and what they cost in computation.

### 1.2 Organization

Section 2 formulates e-closure for general decision problems, including randomized actions, and introduces recertification. Section 3 develops the improvement of CRC, including its admissibility theory, worked examples and a real-data study. Section 4 returns to multiple testing, recertifies randomized e-BH with the e-collections of e-BH and closed e-BH, and reports simulations on Gaussian mean testing. Section 5 comments on the choice among certified actions and concludes.

## 2 Generalized e-closure and recertification

This section sets up the decision framework used throughout the paper. We first describe states, actions and losses, then establish the correspondence between e-values and valid decisions, and finally introduce recertification, which turns this correspondence into a way of improving existing procedures.

## 2.1 States, actions, and a family of risks

Suppose that a statistician observes data $X$ taking values in a space $\mathcal{X}$. The distribution of $X$ is governed by an unknown state $s \in \mathcal{S}$, and each state specifies a nonempty class $\mathcal{P}_s$ of distributions of $X$. After observing $X$, the statistician takes an action $a \in \mathcal{A}$, whose quality is measured by a family of known nonnegative losses

$$L_t(s, a; x), \qquad t \in \mathcal{T}.$$

The index $t$ ranges over the kinds of error that the statistician wishes to control, such as several error rates or a continuum of loss functions, and each loss may depend on the observation $x$ as well as on the state and the action. The sets $\mathcal{S}$, $\mathcal{A}$ and $\mathcal{T}$ are arbitrary and may be uncountable.

Each loss comes with a budget. Given fixed positive levels $\alpha_t(s)$, a decision rule $A : \mathcal{X} \to \mathcal{A}$ is *valid* if

$$\mathbb{E}_P[L_t(s, A(X); X)] \le \alpha_t(s) \quad \text{for every } s \in \mathcal{S},\ P \in \mathcal{P}_s,\ t \in \mathcal{T}. \tag{2.1}$$

In words, whatever the state and whichever distribution within it generates the data, the expected loss of the selected action stays within its budget, separately for every loss in the family. The expectation is unconditional: it averages over the data and hence over the action that the rule selects. Throughout, we assume that all decision rules and all losses that we evaluate are measurable, and when we discuss sets of actions we also assume that their statewise loss envelopes, defined below, are measurable. Appendix A gives sufficient conditions and discusses the uncountable intersections that arise.

Validity is certified by e-values. Recall that an e-value for a class of distributions is a nonnegative statistic whose expectation is at most one under every distribution in the class. Since (2.1) involves every state and every loss, we use one e-value for each such pair. An *e-collection* is a family of nonnegative measurable functions $E_{s,t}$ satisfying

$$\mathbb{E}_P E_{s,t}(X) \le 1 \qquad (s \in \mathcal{S},\ P \in \mathcal{P}_s,\ t \in \mathcal{T}). \tag{2.2}$$

In other words, an e-collection is a set of e-values, one for each combination of state and loss. Because an e-collection will be used to certify actions, we also refer to it as a certificate.

## 2.2 From e-collections to decisions and back

We now make precise the correspondence between e-collections and valid decisions described in the introduction. It is convenient to work with sets of actions rather than with single rules. A *menu* assigns to each observation $x$ a nonempty set $\mathcal{M}(x) \subseteq \mathcal{A}$ of actions that the statistician is prepared to take. After seeing the data, the statistician may choose any action from $\mathcal{M}(X)$, for instance by maximizing a utility or following a scientific preference. A decision rule is the special case of a menu with a single action at each observation. We first show how an e-collection produces a menu, then how a menu produces an e-collection, and finally that the two constructions are compatible.

**From e-collections to menus.** An e-collection certifies an action at the observation $x$ when every loss of that action fits within its budget, scaled by the corresponding e-value. The *certified menu* of $E$ is therefore

$$\mathcal{C}_\alpha(E; x) = \{a \in \mathcal{A} : L_t(s, a; x) \le \alpha_t(s) E_{s,t}(x)\ \forall s, t\}. \tag{2.3}$$

The constraint for state $s$ caps the loss an action may incur if $s$ is true, in proportion to $E_{s,t}(x)$. Intuitively, a large e-value indicates that the data are unlikely under $s$, so the menu may include actions that would perform poorly under $s$; a small one restricts the menu to actions that perform well under $s$.

**From menus to e-collections.** Conversely, a menu determines the largest loss it can incur in each state. For a menu $\mathcal{M}$, define its loss envelopes and their normalized versions by

$$g^{\mathcal{M}}_{s,t}(x) = \sup_{a \in \mathcal{M}(x)} L_t(s, a; x), \qquad E^{\mathcal{M}}_{s,t}(x) = \frac{g^{\mathcal{M}}_{s,t}(x)}{\alpha_t(s)}. \tag{2.4}$$

The menu is valid if $\mathbb{E}_P\, g^{\mathcal{M}}_{s,t}(X) \le \alpha_t(s)$ for every $s \in \mathcal{S}$, $P \in \mathcal{P}_s$ and $t \in \mathcal{T}$. This requires (2.1) to hold for every selection from the menu simultaneously, and it is equivalent to $E^{\mathcal{M}}$ being an e-collection, which we call the *canonical e-collection* of $\mathcal{M}$. Applying the first construction to it gives the closure of $\mathcal{M}$,

$$\overline{\mathcal{M}}(x) = \mathcal{C}_\alpha(E^{\mathcal{M}}; x) = \{a : L_t(s, a; x) \le g^{\mathcal{M}}_{s,t}(x)\ \forall s, t\}, \tag{2.5}$$

which consists of all actions whose losses never exceed those already incurred by the menu.

The following theorem, which we call generalized e-closure, states that the two constructions describe the same objects: certified menus are valid, and every valid menu is certified by its canonical e-collection.

**Theorem 2.1.** *Let $E$ be an e-collection. Every measurable selection $A(X) \in \mathcal{C}_\alpha(E; X)$ satisfies (2.1). More strongly, for every state–loss pair $(s, t)$ and every $P \in \mathcal{P}_s$,*

$$\mathbb{E}_P\left[\sup_{a \in \mathcal{C}_\alpha(E;X)} L_t(s, a; X)\right] \le \alpha_t(s). \tag{2.6}$$

*Further, a menu $\mathcal{M}$ is valid if and only if $\mathcal{M} \subseteq \mathcal{C}_\alpha(E)$ for some e-collection $E$. The loss closure (2.5) is extensive, monotone under inclusion of menus and idempotent, and it preserves every loss envelope. In particular,*

$$\mathcal{M} = \mathcal{C}_\alpha(E) \quad \Longrightarrow \quad \mathcal{M} = \mathcal{C}_\alpha(E^{\mathcal{M}}). \tag{2.7}$$

*Proof.* For every $a \in \mathcal{C}_\alpha(E; X)$, definition (2.3) gives $L_t(s, a; X) \le \alpha_t(s) E_{s,t}(X)$, so the supremum in (2.6) is at most $\alpha_t(s) E_{s,t}(X)$. Taking expectations under $P \in \mathcal{P}_s$ and using (2.2) proves (2.6), and the validity of every measurable selection follows. If $\mathcal{M} \subseteq \mathcal{C}_\alpha(E)$, then $g^{\mathcal{M}}_{s,t} \le \alpha_t(s) E_{s,t}$, so $\mathcal{M}$ is valid. Conversely, if $\mathcal{M}$ is valid, then $E^{\mathcal{M}}$ is an e-collection, and every action in $\mathcal{M}(x)$ satisfies the constraints defining $\mathcal{C}_\alpha(E^{\mathcal{M}}; x)$, so $\mathcal{M} \subseteq \overline{\mathcal{M}}$. Every action in $\overline{\mathcal{M}}(x)$ has losses at most $g^{\mathcal{M}}_{s,t}(x)$, so $\overline{\mathcal{M}}$ has the same envelopes as $\mathcal{M}$; idempotence follows, and monotonicity follows from the monotonicity of the envelopes. Finally, if $\mathcal{M} = \mathcal{C}_\alpha(E)$, then $\mathcal{M} \subseteq \mathcal{C}_\alpha(E^{\mathcal{M}})$, and $g^{\mathcal{M}}_{s,t} \le \alpha_t(s) E_{s,t}$ gives the reverse inclusion $\mathcal{C}_\alpha(E^{\mathcal{M}}) \subseteq \mathcal{C}_\alpha(E) = \mathcal{M}$. □

By (2.6), the guarantee holds simultaneously over all actions in a certified menu. The statistician may therefore choose an action after seeing the data, for instance by maximizing a utility or applying a scientific filter, and every such choice from the menu of a fixed e-collection controls every risk. This is closely related to the post hoc validity of e-values in hypothesis testing, where the significance level, and more generally the loss, may be chosen after the data are observed [Grünwald, 2024, Koning, 2023, Chugg et al., 2026]. We also note that the e-values may be arbitrarily dependent, because for each loss the proof uses only the corresponding e-value under the true state.

**Multiple testing as a special case.** The e-closure principle for multiple testing is recovered as follows. Take both the state space and the action space to be $2^{[K]}$, let $\mathcal{P}_T$ contain the distributions under which $T$ is the set of true null hypotheses, and use the loss $L(T, R) = |T \cap R|/(|R| \vee 1)$. The certified menu (2.3) then consists of the rejection sets certified by (1.1), which is the e-closure principle for the false discovery rate [Xu et al., 2026b], and the loss $L(T, R) = \mathbb{1}\{T \cap R \neq \varnothing\}$ gives its familywise error rate version. E-values for intersection hypotheses, which are valid whenever $T$ is contained in the set of true nulls, are in particular valid for the exact-state classes $\mathcal{P}_T$ used here. Section 4 reviews this setting in more detail. The general framework requires no structure on the actions, such as unions or an ordering, and no monotonicity of the losses. For instance, the false discovery proportion can decrease when a rejection is added.

**Randomized actions.** Actions may also be randomized. A randomized action is a probability measure $Q$ on $\mathcal{A}$: once $Q$ is chosen, an action is drawn from it using randomization independent of the data, and the loss of $Q$ is the average $\bar{L}_t(s, Q; x) = \int L_t(s, a; x)\, Q(da)$. Probability measures on $\mathcal{A}$ form an action space like any other, so Theorem 2.1 applies with these averaged losses. Every data-dependent choice of $Q$ from the certified menu $\{Q : \bar{L}_t(s, Q; x) \le \alpha_t(s) E_{s,t}(x)\ \forall s, t\}$ is therefore valid, with a guarantee that averages

over the subsequent draw; it does not cover choices made after the draw is seen. This menu can contain randomized actions whose possible realizations are not individually certified, so randomization can enlarge what a fixed e-collection certifies. Section 4 uses this to improve randomized multiple testing.

### 2.3 Recertification: spending the unused budget

A valid menu can be enlarged in two ways, and the distinction between them is central to our approach. The first is *closing*, which adds every action whose losses lie beneath the current envelopes of the menu. By Theorem 2.1, closing preserves every envelope: the closed menu offers more choice, which in multiple testing is the simultaneity gain described by Xu et al. [2026b], but it uses exactly the budget that the original menu uses. Any budget that the original procedure leaves unused therefore remains unused after closing. The second way, which we call *recertifying*, first replaces the envelopes by larger ones that still respect the budgets and then closes with respect to the new envelopes. The recertified menu can therefore contain actions whose losses exceed those of every action in the original menu, and this is how recertification spends the unused budget.

To recertify, we need to quantify the budget that the baseline leaves unused. We do so conditionally on a statistic, chosen for each state–loss pair so that the required conditional expectations are known. When the unused budget can be computed without conditioning, the statistic is simply taken to be constant. Fix a valid baseline menu with envelopes $g_{s,t}$, and let $V_{s,t}$ denote the conditioning statistic for the pair $(s,t)$. Suppose that known nonnegative functions $H_{s,t}$ and $B_{s,t}$ satisfy, for every $P \in \mathcal{P}_s$,

$$\mathbb{E}_P[g_{s,t}(X) \mid V_{s,t}] \le H_{s,t}(V_{s,t}) \le B_{s,t}(V_{s,t}), \qquad \mathbb{E}_P B_{s,t}(V_{s,t}) \le \alpha_t(s). \tag{2.8}$$

The function $H_{s,t}$ bounds the conditional expected loss of the baseline, while $B_{s,t}$ is a conditional budget whose expectation does not exceed the original level. Their difference $B_{s,t} - H_{s,t}$ is therefore an allowance that the baseline leaves unused, conditionally on $V_{s,t}$. The following proposition shows that any nonnegative increment within this allowance can be added to the envelopes.

**Proposition 2.2** (Conditional recertification)**.** *If measurable functions $W_{s,t}(X) \ge 0$ satisfy*

$$\mathbb{E}_P[W_{s,t}(X) \mid V_{s,t}] \le B_{s,t}(V_{s,t}) - H_{s,t}(V_{s,t}) \quad (P \in \mathcal{P}_s), \tag{2.9}$$

*then $(g_{s,t} + W_{s,t})/\alpha_t(s)$ is an e-collection. Its certified menu contains the original menu and every action whose losses are at most $g_{s,t} + W_{s,t}$ for all $s$ and $t$.*

*Proof.* Conditionally on $V_{s,t}$, the expectation of $g_{s,t} + W_{s,t}$ is at most $H_{s,t} + (B_{s,t} - H_{s,t}) = B_{s,t}$. Taking expectations once more, the unconditional expectation is at most $\alpha_t(s)$. The two inclusions follow from $W_{s,t} \ge 0$ and from the definition (2.3) of the certified menu. □

The increments $W_{s,t}$ may depend on the data in an arbitrary way; only their conditional expectations are constrained by (2.9). How the allowance is computed depends on the problem. In CRC (Section 3), the state is a hypothetical augmented sample whose law is known by exchangeability, so no further conditioning is needed, and the allowance is shared equally among the held-out points. For randomized e-BH (Section 4), the input e-values supply the budget directly, and conditional on the e-values, the averaged losses are known because the law of the external randomization is known. Proposition 2.2 also describes an existing method, the conditional calibration of e-values [Fithian and Lei, 2022, Lee and Ren, 2024], which improves e-BH by recalibrating each e-value given a sufficient statistic; there the sufficient statistic provides the conditional expectations (Remark 4.3). Finally, the recertified e-collection is again an e-collection, so its certified menu is valid by Theorem 2.1 and can serve as the baseline for another round of recertification.

# 3 Improving conformal risk control

We now apply the framework of Section 2 to conformal risk control. Section 3.1 sets up the problem and two baseline procedures, and Section 3.2 characterizes validity exactly through the losses on hypothetical augmented samples. This characterization exposes the budget that a controller leaves unused, which recertification spends in Section 3.3. Repeating the step leads to admissible controllers, which we illustrate on a small example (Section 3.4). Section 3.5 extends the construction to multiple risks, nonconvex losses and test features, and discusses computation, and Section 3.6 reports a real-data study on hierarchical image classification.

## 3.1 Setup and two valid starting rules

Conformal risk control [Angelopoulos et al., 2024] calibrates a family of actions so that a prescribed expected loss is controlled at a new test point. The actions are indexed by a parameter $\lambda$ in a finite ordered grid or a compact interval $\Lambda = [\lambda_-, \lambda_+]$, where larger values are more conservative. For example, $\lambda$ may be a threshold that determines a prediction set, with larger $\lambda$ giving larger sets. The quality of action $\lambda$ on an observation $z \in \mathcal{Z}$ is measured by a known loss $\ell(z, \lambda)$, such as the indicator that the prediction set misses the label, or the fraction of true labels that it misses. Following the original work, we assume that

$$0 \le \ell(z, \lambda) \le B, \qquad \ell(z, \lambda) \text{ is non-increasing and right-continuous in } \lambda, \qquad \ell(z, \lambda_+) \le \alpha, \tag{3.1}$$

where $0 < \alpha < B$ is the target level. Thus the losses are bounded, a more conservative action never incurs a larger loss, and the most conservative action always meets the target.

We observe calibration data $D = (Z_1, \dots, Z_n)$ and must choose an action for a test observation $Z_{n+1}$, whose loss we cannot see. We assume only that $(Z_1, \dots, Z_{n+1})$ is exchangeable; if the loss involves a trained prediction model, the model is held fixed, and exchangeability is assumed conditionally on its training data. The goal is a data-dependent action $A(D) \in \Lambda$ satisfying

$$\mathbb{E}\, \ell(Z_{n+1}, A(D)) \le \alpha$$

under every finite exchangeable law on $\mathcal{Z}^{n+1}$, and, subject to this constraint, as liberal as possible. We call such a rule $A$ a controller. The observation space and the loss family are fixed throughout, and controllers are always compared for the same observation space and loss family.

Standard CRC bounds the unobserved test loss by $B$ and selects the smallest $\lambda$ for which the calibration losses, together with this bound, fit within the budget $(n+1)\alpha$:

$$A_{\mathrm{CRC}}(D) = \min\left\{ \lambda : \sum_{i=1}^{n} \ell(Z_i, \lambda) + B \le (n+1)\alpha \right\}. \tag{3.2}$$

The bound $B$ is the same for every $\lambda$, whereas the loss that the test point can incur at a given $\lambda$ may be much smaller, particularly for conservative actions. Yeh et al. [2025] therefore allow the bound to depend on $\lambda$. Given a known non-increasing, right-continuous envelope $U$ with

$$\sup_{z \in \mathcal{Z}} \ell(z, \lambda) \le U(\lambda) \le B,$$

envelope CRC selects

$$A_{\mathrm{env}}(D) = \min\left\{ \lambda : \sum_{i=1}^{n} \ell(Z_i, \lambda) + U(\lambda) \le (n+1)\alpha \right\}. \tag{3.3}$$

In both rules, the safe endpoint $\lambda_+$ is used if the displayed set is empty. Both rules are valid [Angelopoulos et al., 2024, Yeh et al., 2025], and $U \le B$ gives $A_{\mathrm{env}} \le A_{\mathrm{CRC}}$ pointwise. The construction below improves any valid controller, envelope CRC included, and since the goal is an expected-loss bound, it fits the framework of Section 2: the e-values that certify a controller show where it can be improved.

### 3.2 The exact risk budget on a hypothetical augmented sample

Validity under exchangeability can be checked one augmented sample at a time. For an augmented multiset $S = (z_1, \dots, z_{n+1})$ of $n+1$ points, let $S_{-i}$ denote $S$ with one occurrence of $z_i$ removed, and define

$$r_i^A(S) = \ell(z_i, A(S_{-i})), \qquad T_A(S) = \sum_{i=1}^{n+1} r_i^A(S), \qquad \Delta_A(S) = (n+1)\alpha - T_A(S). \tag{3.4}$$

Here $r_i^A(S)$ is the loss incurred when $z_i$ plays the role of the test point and the remaining $n$ points form the calibration data, $T_A(S)$ is the total of these held-out losses, and $\Delta_A(S)$ is the part of the budget $(n+1)\alpha$ that they leave unused. Repeated observations are counted with their multiplicities.

**Lemma 3.1** (Exact leave-one-out characterization)**.** *A symmetric controller $A$ is valid under every finite exchangeable law if and only if*

$$T_A(S) \le (n+1)\alpha \qquad \textit{for every deterministic augmented multiset } S. \tag{3.5}$$

*Proof.* By symmetry of $A$ and exchangeability, the test loss has the same distribution as each held-out loss, so $\mathbb{E}\ell(Z_{n+1}, A(Z_{1:n})) = (n+1)^{-1}\mathbb{E}T_A(Z_{1:n+1})$, which proves sufficiency. For necessity, fix $S$ and let the data be a uniformly random ordering of $S$. This law is exchangeable, and its risk is exactly $T_A(S)/(n+1)$. □

The lemma is closely connected to the orbit-average characterization of exchangeability e-values [Vovk, 2025, Koning, 2026], and it is the risk-control counterpart of the universality of conformal prediction [Vovk et al., 2022, Proposition 2.9]. Its necessity part uses all finite exchangeable laws, which is why we require validity under exchangeability rather than under i.i.d. sampling alone (Appendix B.1).

For a valid controller, $\Delta_A(S) \ge 0$ is an exact account of the unused budget of the augmented sample $S$: it is the total amount by which the $n+1$ held-out losses on $S$ could increase without violating (3.5). It can be computed from the controller and the known loss once every point of $S$ is specified (see the example in Section 3.4). Finally, we remark that the original CRC also connects to this view: its validity can be derived from conformal e-values [Vovk, 2025, Koning, 2026] that are allowed to depend on the action being certified (Appendix B.1).

### 3.3 A baseline-preserving refinement

Starting from any valid symmetric controller $A$, recertification spends $\Delta_A(S)$ by distributing it among the held-out positions. For example, with equal sharing,

$$b_i^A(S) = r_i^A(S) + \frac{\Delta_A(S)}{n+1}, \qquad \sum_{i=1}^{n+1} b_i^A(S) = (n+1)\alpha. \tag{3.6}$$

The new budgets still sum to $(n+1)\alpha$ on every augmented sample, and each exceeds the corresponding held-out loss of $A$ by $\Delta_A(S)/(n+1)$. The recertified controller selects the smallest action whose loss, for every possible test point $z$, stays within its new budget:

$$(\mathcal{R}A)(D) = \min\left\{\lambda \le A(D) : \ell(z,\lambda) - \ell(z, A(D)) \le \frac{\Delta_A(D \cup \{z\})}{n+1} \ \forall z \in \mathcal{Z}\right\}. \tag{3.7}$$

The baseline action $A(D)$ satisfies every constraint, so the feasible set is nonempty, and on a finite grid, or on a compact interval under (3.1), its minimum is attained. We assume that the minimizer is implemented measurably. More generally, any feasible action no larger than $A(D)$ enjoys the following guarantee, so exact minimization is optional.

**Theorem 3.2** (Uniform improvement)**.** *For every valid symmetric baseline $A$, the controller $\mathcal{R}A$ is valid and satisfies $\mathcal{R}A \le A$ pointwise. In particular, with $A^{(0)} = A_{\mathrm{env}}$ and $A^{(k+1)} = \mathcal{R}A^{(k)}$,*

$$A^{(k+1)}(D) \le A^{(k)}(D) \le A_{\mathrm{env}}(D) \le A_{\mathrm{CRC}}(D) \quad \text{for every } D, k, \tag{3.8}$$

*and every iterate controls the expected loss at level $\alpha$.*

*Proof.* Fix $S$. The constraint in (3.7) with $D = S_{-i}$ and $z = z_i$ shows that each held-out loss of $\mathcal{R}A$ exceeds that of $A$ by at most $\Delta_A(S)/(n+1)$. Summing over $i$ gives

$$\sum_{i=1}^{n+1} \ell(z_i, (\mathcal{R}A)(S_{-i})) \le T_A(S) + \Delta_A(S) = (n+1)\alpha,$$

so $\mathcal{R}A$ is valid by Lemma 3.1. Since $A(D)$ is feasible in (3.7), $\mathcal{R}A \le A$, and the chain (3.8) follows by induction from $A_{\mathrm{env}} \le A_{\mathrm{CRC}}$. □

**Recertified CRC as an instance of generalized e-closure.** The improved controller $\mathcal{R}A$ is obtained exactly by applying Proposition 2.2 and Theorem 2.1 to the CRC problem, with unordered augmented samples as states. Let the data be the calibration multiset $D$, and let a state be an augmented multiset $S = (z_1, \dots, z_{n+1})$. The class $\mathcal{P}_S$ consists of a single law, under which $D = S_{-I}$ for a position $I$ drawn uniformly from $\{1, \dots, n+1\}$. Conditioning a finite exchangeable law on the multiset of its $n+1$ values yields such a law, and each such law arises from the exchangeable law that orders $S$ uniformly at random. When $D \subseteq S$, the held-out point $S \setminus D$ is determined by $S$ and $D$, and the loss is

$$L(S, \lambda; D) = \ell(S \setminus D, \lambda),$$

with $L(S, \lambda; D) = 0$ when $D$ cannot arise from $S$. Validity in state $S$ is then exactly the constraint (3.5) for that $S$, so validity in every state is equivalent to validity under every finite exchangeable law.

The canonical e-collection of a valid controller $A$ is $E_S^A(D) = L(S, A(D); D)/\alpha$, whose expectation under $\mathcal{P}_S$ is $T_A(S)/((n+1)\alpha)$. Its shortfall from one is $\Delta_A(S)/((n+1)\alpha)$, the unused budget of (3.4). Since each state specifies a single known law, recertification fills this shortfall without conditioning. In Proposition 2.2 we take $V_S$ constant, the baseline envelope $g_S(D) = L(S, A(D); D)$, the bounds $H_S = T_A(S)/(n+1)$ and $B_S = \alpha$, and the increment $W_S(D) = \Delta_A(S)/(n+1)$ for $D \subseteq S$ (and zero otherwise). The recertified e-collection is

$$E'_S(D) = \frac{g_S(D) + W_S(D)}{\alpha} = \frac{b_i^A(S)}{\alpha} \qquad \text{for } D = S_{-i},$$

which has expectation exactly one by (3.6). At an observed $D$, only the states $S = D \cup \{z\}$ with $z \in \mathcal{Z}$ impose constraints, and the certified menu (2.3) becomes

$$\mathcal{C}_\alpha(E'; D) = \left\{ \lambda \in \Lambda : \ \ell(z, \lambda) - \ell(z, A(D)) \le \frac{\Delta_A(D \cup \{z\})}{n+1} \ \ \forall z \in \mathcal{Z} \right\}.$$

Each constraint bounds a non-increasing function of $\lambda$, and $A(D)$ satisfies all of them, so the menu is an upper set of $\Lambda$ that contains $A(D)$, and its smallest element is $(\mathcal{R}A)(D)$. The restriction $\lambda \le A(D)$ in (3.7) is thus automatic: the recertified controller selects the least conservative certified action, and by Theorem 2.1 any other selection from the menu would be valid as well.

**Remark 3.3.** Alternatively, one can view the augmented sample as a conditioning statistic rather than as the state. Take the full sample $X = (Z_1, \dots, Z_{n+1})$ as data, of which only $D = (Z_1, \dots, Z_n)$ is observed, and use a single state whose class contains all finite exchangeable laws, with loss $\ell(Z_{n+1}, \lambda)$. This class is composite, but conditioning on the multiset $V$ of the values $Z_1, \dots, Z_{n+1}$ removes the dependence on the law. Proposition 2.2 then applies with $H(V) = T_A(V)/(n+1)$, $B(V) = \alpha$ and $W(X) = \Delta_A(V)/(n+1)$. The resulting certified menu depends on the unobserved test point, and requiring the action to be certified for every possible value of $Z_{n+1}$ recovers (3.7).

**Algorithm 1** One recertification of a known-loss CRC controller

**Input:** A valid symmetric controller $A$, loss $\ell$, calibration multiset $D$, and level $\alpha$.

Compute $a_0 = A(D)$.
**for** each hypothetical $z \in \mathcal{Z}$ **do**
    Compute $T_A(D \cup \{z\}) = \ell(z, a_0) + \sum_{j=1}^{n} \ell(Z_j, A(D_{-j} \cup \{z\}))$.
    Set $s_A(D, z) = \alpha - T_A(D \cup \{z\})/(n+1)$.
**end for**
Return the smallest $\lambda \le a_0$ such that $\ell(z, \lambda) - \ell(z, a_0) \le s_A(D, z)$ for every $z$.

Algorithm 1 summarizes one round of recertification at an observed calibration multiset $D$. Each round evaluates the current controller on all replacement samples $D_{-j} \cup \{z\}$. Equal sharing is the simplest allocation of the unused budget. Other permutation-respecting allocations are also valid, and every valid pointwise improvement of a controller with monotone losses arises from some allocation (Appendix B.2); which allocation works best in one step depends on the problem.

## 3.4 Iteration to admissibility under convexity

Since recertification preserves validity, it can be repeated, and each round can only improve the selected action. For convex losses, the iterates converge to an admissible controller that cannot be improved further, in the following pointwise sense.

**Definition 3.4** (Pointwise admissibility for CRC)**.** *A valid symmetric deterministic controller $A$ is pointwise admissible if no other measurable symmetric deterministic controller $C$ is valid under every finite exchangeable law and satisfies $C(D) \le A(D)$ for every $D$, strictly at some $D$. The observation domain, loss family, sample size and action space remain fixed.*

Admissibility is thus a pointwise notion: controllers are compared on every calibration sample. This is meaningful because Lemma 3.1 imposes a constraint on every augmented multiset, with no exceptional null sets, so changing a controller at even a single calibration multiset can affect its validity. The notion concerns only how conservative the selected actions are. An admissible controller need not be optimal for a particular utility, and several admissible controllers may coexist, as the example at the end of this section shows.

**Theorem 3.5** (An admissible limit)**.** *Suppose $\Lambda = [\lambda_-, \lambda_+]$ and every $\ell(z, \cdot)$ in (3.1) is convex. Suppose the exact inversions in (3.7) are measurably implemented. From any valid baseline, the iterates decrease to a valid pointwise-admissible controller $A^{(\infty)}$. With $\rho = n/(n+1)$,*

$$0 \le A^{(k)}(D) - A^{(\infty)}(D) \le \rho^k (A^{(0)}(D) - A^{(\infty)}(D)) \le \rho^k (\lambda_+ - \lambda_-). \tag{3.9}$$

*A valid controller is admissible if and only if it is an exact fixed point of $\mathcal{R}$ in this setting.*

The proof rests on a one-step comparison. Suppose that a valid controller $C$ satisfies $C \le A$. On every augmented sample, each held-out loss of $C$ exceeds that of $A$ by a nonnegative amount, and by Lemma 3.1 these increases sum to at most $\Delta_A(S)$, so each is at most $\Delta_A(S)$. By convexity, the action $\rho A(D) + C(D)/(n+1)$ increases each loss by at most a fraction $1/(n+1)$ of this amount, so it is feasible in (3.7), and therefore

$$\mathcal{R}A \le \rho A + C/(n+1). \tag{3.10}$$

The iterates decrease to a limit, which is valid by the right continuity of the losses. Applying (3.10) along the iterates and passing to the limit shows that no valid controller strictly dominates the limit, and taking $C = A^{(\infty)}$ gives the rate (3.9). Appendix B.3 gives the complete proof.

We note that convexity is used only for the admissibility of the limit, and every finite round of recertification is valid without it. The limit theorem also uses an interval of actions, since on a grid the interpolated

action in (3.10) may be unavailable; Appendix B.5 gives an example. Many common prediction losses, such as miscoverage indicators, are step functions of $\lambda$. For such losses the iteration still improves the baseline validly, and Section 3.5 describes a different completion that reaches an admissible controller using only monotonicity.

**A two-type example.** A small example makes the iteration concrete. Let the observation space consist of two points, $\mathcal{Z} = \{H, L\}$, with a high and a low loss,

$$\ell(H, \lambda) = 1 - \lambda, \qquad \ell(L, \lambda) = \tfrac{1}{2}(1 - \lambda), \qquad \lambda \in \Lambda = [0, 1], \tag{3.11}$$

and take $n = 2$ and $\alpha = 1/2$. The probabilities of the two points, and the dependence between observations, are arbitrary apart from exchangeability. We write multisets of observations as strings: $HL$ is the calibration sample with one point of each kind, and $HLL$ is an augmented sample, that is, a state in the formulation of Section 3.3. The sharp envelope is $U(\lambda) = 1 - \lambda$, and Table 1 lists each controller on the three possible calibration samples.

Table 1: The two-type example with $n = 2$ and $\alpha = 1/2$. Rows are the possible calibration samples and columns are controllers; entries are the selected actions, and smaller is preferred. The iterates start at envelope CRC. Every comparison holds for all exchangeable distributions on $\{H, L\}^3$.

| Calibration | CRC | Envelope $A^{(0)}$ | $A^{(1)}$ | $A^{(2)}$ | $A^{(\infty)}$ |
|---|---|---|---|---|---|
| $HH$ | 3/4 | 1/2 | 1/2 | 1/2 | 1/2 |
| $HL$ | 2/3 | 2/5 | 23/60 | 17/45 | 3/8 |
| $LL$ | 1/2 | 1/4 | 1/5 | 31/180 | 1/8 |

For instance, at $D = LL$ the envelope action is $1/4$. Appending $H$ gives the augmented sample $HLL$, whose held-out losses $3/4$, $3/10$ and $3/10$ total $27/20$. The equal allowance is $1/2 - (27/20)/3 = 1/20$, so the constraint for $z = H$ reads $1/4 - \lambda \le 1/20$, that is, $\lambda \ge 1/5$, while appending $L$ permits every $\lambda \ge 0$. Thus $A^{(1)}(LL) = 1/5$. The entries converge geometrically, as Theorem 3.5 predicts with $\rho = 2/3$. At the limit, the validity constraints for $HHH$, $HHL$ and $HLL$ bind, which makes the limit admissible, whereas the constraint for $LLL$ is slack: an admissible controller may leave budget unused on some augmented samples. Appendix B.5 gives the exact iteration formulas. Admissible controllers need not be unique either. The controller with entries $(1/2, 2/5, 1/10)$ on $(HH, HL, LL)$ is also admissible and is incomparable with the limit, and the pointwise minimum of the two is invalid. Which admissible controller is preferable thus generally depends on the specific problem setting.

## 3.5 Extensions and computation

We now extend the construction to multiple risks, to nonconvex losses and to controllers that use test features, and finally discuss its computational cost.

**Multiple risks.** In many applications a single action must control multiple losses at once. A prediction set, for example, may be required to control both its miscoverage and the fraction of true labels it misses, or a whole family of losses indexed by a continuous severity level. CRC handles such families when the losses share the same action [Angelopoulos et al., 2024, Section 4.3], and recertification extends in the same way.

Let $\ell_t(z, \lambda)$, $t \in \mathcal{T}$, be a family of losses indexed by an arbitrary, possibly uncountable, set $\mathcal{T}$, each satisfying (3.1) with its own level $\alpha_t > 0$ and the same action space $\Lambda$. A controller $A$ is *jointly valid* if $\mathbb{E}\,\ell_t(Z_{n+1}, A(D)) \le \alpha_t$ for every $t \in \mathcal{T}$ under every finite exchangeable law. By Lemma 3.1, applied to each risk, this holds if and only if

$$T_{A,t}(S) = \sum_{i=1}^{n+1} \ell_t(z_i, A(S_{-i})) \le (n+1)\alpha_t \qquad \text{for every augmented multiset } S \text{ and every } t \in \mathcal{T}.$$

Since all losses are non-increasing in $\lambda$, the most conservative of the separate CRC choices, that is, their supremum over $t$, is jointly valid whenever it is measurable, and the same holds for envelope CRC. For a jointly valid $A$, recertification shares each loss-wise unused budget equally among the held-out positions and imposes the resulting constraints for all risks simultaneously:

$$\begin{aligned}\Delta_{A,t}(S) &= (n+1)\alpha_t - T_{A,t}(S),\\ (\mathcal{R}A)(D) &= \min\left\{\lambda \le A(D) : \ell_t(z,\lambda) - \ell_t(z, A(D)) \le \frac{\Delta_{A,t}(D\cup\{z\})}{n+1} \quad \forall z\in\mathcal{Z},\ t\in\mathcal{T}\right\}.\end{aligned} \tag{3.12}$$

**Proposition 3.6** (Multiple risks)**.** *For every jointly valid symmetric controller $A$, the controller $\mathcal{R}A$ in (3.12) is jointly valid and satisfies $\mathcal{R}A \le A$. If, moreover, $\Lambda = [\lambda_-, \lambda_+]$ and every $\ell_t(z,\cdot)$ is convex, the conclusions of Theorem 3.5 hold for the iterates of (3.12), with the same rate $\rho = n/(n+1)$, regardless of the index set $\mathcal{T}$.*

The proof of Theorem 3.2 simply applies to each risk separately; see Appendix B.3. The number of risks thus enters neither the rate nor the admissibility conclusion. It matters only for attaining the minimum in (3.12) measurably. Appendix B.6 works out a simple two-risk example with a continuum of observations, in which recertification converges to the largest valid rule dominating the joint envelope baseline, a strict improvement at every interior point.

**Nonconvex losses.** Theorem 3.5 relies on convexity, which fails for many prediction losses, such as miscoverage indicators. For general monotone losses, admissibility can instead be characterized one calibration multiset at a time. Let $\nu_D(z)$ denote the number of copies of $z$ in $D$. Lowering the controller only at $D$, from $A(D)$ to $a < A(D)$, changes the held-out losses only on the augmented samples $D\cup\{z\}$, in each of which $z$ is held out $\nu_D(z)+1$ times. By Lemma 3.1, the change preserves joint validity if and only if

$$(\nu_D(z)+1)\{\ell_t(z,a) - \ell_t(z, A(D))\} \le \Delta_{A,t}(D\cup\{z\}) \qquad \forall z\in\mathcal{Z},\ t\in\mathcal{T}. \tag{3.13}$$

**Proposition 3.7.** *A jointly valid symmetric controller $A$ is pointwise admissible if and only if (3.13) fails for every calibration multiset $D$ and every $a < A(D)$. If the observation space $\mathcal{Z}$ is finite, the following sweep produces a pointwise-admissible controller that is no larger than $A$: visit the calibration multisets once in a fixed order, and at each visit replace $A(D)$ by the smallest $a$ satisfying (3.13), holding all other entries fixed.*

The sweep uses only monotonicity: lowering one entry can only shrink the allowances of the others, so an entry that cannot be lowered when it is visited cannot be lowered later (Appendix B.4). Its cost is a single pass over the controller table, which has one entry per calibration multiset.

**Test features.** In prediction problems, the feature $x$ of the test point is observed and only its label $y$ is unknown, so the controller may depend on it, $A(D,x)$. Validity then means $\mathbb{E}\,\ell_t(Z_{n+1}, A(D, X_{n+1})) \le \alpha_t$, a guarantee that is marginal over test features and labels, as for CRC itself. On an augmented sample, the held-out loss becomes $\ell_t(z_i, A(S_{-i}, x_i))$, and the recertified controller at the observed $(D,x)$ is

$$(\mathcal{R}A)(D,x) = \min\left\{\lambda \le A(D,x) : \ell_t((x,y),\lambda) - \ell_t((x,y), A(D,x)) \le \frac{\Delta_{A,t}(D\cup\{(x,y)\})}{n+1} \quad \forall y, t\right\}.$$

The constraints range only over the possible labels $y$, while each unused budget $\Delta_{A,t}$ evaluates the current controller at the feature of every held-out point. Lemma 3.1 continues to hold with these held-out losses, since its proof uses only the symmetry of $A$ in $D$ and exchangeability. The constraint for the label $y$ at $(D,x)$ is the constraint for the held-out copy of $(x,y)$ in $D\cup\{(x,y)\}$, so the proofs of Theorems 3.2 and 3.5 carry over with $D$ replaced by $(D,x)$.

**Computation.** The risk-control guarantees hold after any number of rounds, so the number of rounds is a computational choice, whose cost depends on how the controller depends on the data.

A single round is usually inexpensive. With $K$ candidate outcomes, $(\mathcal{R}A)(D)$ requires $A(D)$ and the totals $T_A(D \cup \{z\})$, that is, the controller evaluated on the $nK$ replacement samples $D_{-j} \cup \{z\}$, each a recalibration with the prediction model held fixed. For CRC and envelope CRC, these recalibrations share most of their work. A replacement changes a single term of the calibration loss $\sum_{i=1}^{n} \ell(Z_i, \lambda)$, so the sum is computed once and adjusted for each replacement, and each replacement action is the smallest $\lambda$ satisfying a constraint that is monotone in $\lambda$. Such an action is found by a binary search for general losses and directly for structured ones, such as piecewise-linear or step losses.

Repeated rounds are more demanding, because the replacements compound. Round $k$ evaluates $A^{(k-1)}$ on replacement samples, each of which requires $A^{(k-2)}$ on samples with a second point replaced, and so on. Hence $A^{(k)}(D)$ depends on the baseline at every calibration multiset obtained from $D$ by replacing up to $k$ of its points with candidate outcomes. The cost of many rounds is governed by how many distinct multisets arise and whether they can be shared across calibration samples and test points, and both depend on how the baseline depends on the data.

- *Finitely many types.* Suppose that the loss depends on an observation only through one of $K$ types and that the baseline depends on the calibration data only through the type counts. Every iterate then inherits this property, so all the multisets above belong to a single table indexed by count vectors. The table has $\binom{n+K-1}{K-1}$ entries, polynomially many in $n$ for fixed $K$, and each round is one pass over it. It does not depend on the observed data, so it is computed once and reused for every calibration sample and test point.
- *Individual scores or features.* When the baseline depends on individual observations, for instance through continuous scores or test features, no such table exists. The replacement samples must then be tracked separately for each calibration sample and test point, and their number grows exponentially in the number of calibration points whose loss is not identically zero. Exact reductions, such as grouping points with identical scores, can still make many rounds feasible (Appendix B.7).
- *Uncountable outcomes or risks.* Here even a single round requires care. The constraints in (3.12) must hold for a continuum of $z$ or $t$, and checking them requires a certified optimization or a uniform approximation bound, since finitely many sampled outcomes do not certify feasibility.

Two observations reduce the cost in every case. First, the iterates decrease and the losses are non-increasing, so stopping at a data-dependent round no later than a fixed maximum incurs at most the loss of the maximal round and preserves validity. Computation can therefore stop once a computable bound shows that further rounds cannot change the action appreciably. Second, replacing the allowances $\Delta_{A,t}$ by conservative lower bounds preserves validity and the comparison with the baseline, although the limit may then fall short of admissibility. Exact recertification thus trades computation for a uniform improvement.

## 3.6 Real-data study

We evaluate recertification on hierarchical image classification, one of the applications of CRC in Angelopoulos et al. [2024], and compare it with ordinary CRC and with envelope CRC [Yeh et al., 2025].

### 3.6.1 Hierarchical classification

We consider the hierarchical image classification setting, where each observation is $Z_i = (X_i, Y_i)$. Here, each $X_i \in \mathcal{X}$ is an image and $Y_i \in \mathcal{Y} = \{1, \ldots, K_1\}$ is its fine-class label. The labels are partitioned into $K_2$ superclasses $G_1, \ldots, G_{K_2}$, which form a tree of depth two with a root, the superclasses and the fine-class leaves. We identify each node with the set of its descendant labels, so that the root is $\mathcal{Y}$.

A fixed, pretrained classifier outputs class probabilities $p(x) = (p_1(x), \ldots, p_{K_1}(x))$. Let $f(x) = \arg\max_k p_k(x)$, with ties broken deterministically, and let $g(x)$ be the superclass containing $f(x)$. For an image $x$, predictions are restricted to the three nodes $\mathcal{Y}$, $g(x)$ and $\{f(x)\}$ on this path. For a reported node

$v$ and a true class $y$, the hierarchical loss is

$$L(v, y) = \frac{1}{2} \min_{u \in \mathrm{anc}(y)} d(v, u),$$

where $d$ is the tree distance and $\mathrm{anc}(y)$ consists of the leaf $\{y\}$ and its ancestors. Thus $L(v, y) = 0$ when $y \in v$, and otherwise $L(v, y)$ is half the number of upward edges needed to reach a node containing $y$.

To obtain a continuous action space, we allow randomization between adjacent depths. For $a \in [0, 2]$, let $N_a(x)$ report the root, the superclass and the leaf with probabilities $((1 - a)_+, 1 - |a - 1|, (a - 1)_+)$, where $r_+ = \max\{r, 0\}$. Its mean depth is $a$, and its expected loss is

$$\ell((x, y), a) = \mathbb{E}[L(N_a(x), y)] = \begin{cases} 0, & y = f(x), \\ (a - 1)_+/2, & y \in g(x) \setminus \{f(x)\}, \\ a/2, & y \notin g(x). \end{cases} \tag{3.14}$$

Given calibration data $D = (Z_1, \ldots, Z_n)$ and a target $\alpha \in (0, 1)$, a procedure chooses a depth rule $\widehat{a} : \mathcal{X} \to [0, 2]$ and reports $N_{\widehat{a}(x)}(x)$ for a test image $x$. The aim is to make predictions as specific as possible subject to

$$\mathbb{E}\, \ell\big(Z_{n+1}, \widehat{a}(X_{n+1})\big) \leq \alpha,$$

where $Z_{n+1}$ is a test observation exchangeable with $D$, the expectation is over the calibration and test data, and the randomization of the prediction is already integrated into $\ell$. This is the setting of Section 3 with test features, in which the depth rule is the controller.

We consider two families of depth rules $a_t$, indexed by a parameter $t$. The first uses a common randomized depth, $a_t(x) = t$ for $t \in [0, 2]$: every image receives the same distribution over depths, although the reported superclass and leaf still depend on the image through $g(x)$ and $f(x)$. The second adapts the depth to the confidence of the classifier. Write $p_f(x) = p_{f(x)}(x)$ for the probability of the most likely class and $p_g(x) = \sum_{k \in g(x)} p_k(x)$ for the total probability of its superclass. For a threshold $t \in [0, 1]$, set

$$a_t(x) = \mathbb{1}\{t > 1 - p_g(x)\} + \mathbb{1}\{t > 1 - p_f(x)\}.$$

This rule reports the root when $p_g(x) \leq 1 - t$, the superclass when $p_f(x) \leq 1 - t < p_g(x)$, and the leaf when $p_f(x) > 1 - t$, so more confident images receive more specific predictions.

In both families, larger values of $t$ are less conservative, so the parameter $\lambda$ of Section 3 corresponds to $-t$. CRC selects the largest $t$ satisfying

$$\sum_{i=1}^{n} \ell(Z_i, a_t(X_i)) + 1 \leq (n + 1)\alpha,$$

where 1 is the largest possible loss. Envelope CRC replaces this bound by the sharper envelope $U(t)$, which equals $t/2$ for common depth and 1 for every $t > 0$ for confidence-adaptive scores; in the latter family, envelope CRC therefore coincides with CRC. Starting from envelope CRC, we apply one round and 500 rounds of recertification with equal sharing. Each round recomputes the unused budgets on the augmented samples and increases the depth as far as they permit. In the common-depth family, the recertified rule still assigns one randomized depth to all images and depends on the calibration data only through the counts of the three cases in (3.14), so it is a finite table of the kind discussed in Section 3.5. In the confidence-adaptive family, the recertified rule depends on the test image and may randomize between adjacent depths beyond the deterministic threshold prediction.

### 3.6.2 Results

We use CIFAR-100, which has $K_1 = 100$ fine classes and $K_2 = 20$ superclasses of five classes each, so the root, superclass and leaf predictions contain 100, 5 and 1 labels. The classifier is a frozen ResNet-18, and the data

pool consists of its cached predictions on the 10,000 test images [González-González, 2026]. We report the mean hierarchical loss (risk), the mean node size $|v|$, and the correct specificity $\mathbb{1}\{y \in v\}[1 - \log|v|/\log K_1]$, which rewards predictions that are both correct and informative. All three metrics are averaged analytically over the randomization of the prediction, and gains are absolute differences in correct specificity, in percentage points.

We fix $n = 50$ and $\alpha \in \{0.02, 0.05, 0.10, 0.15\}$, and average over 300 random splits of the pool, each with 50 calibration images and 16 disjoint test images. The same splits are used for all families, methods and targets, and standard errors are computed across splits. Appendix B.7 gives the implementation details.

Table 2 reports the results. All methods have empirical risk below the target in every setting. In the common-depth family, envelope CRC is much less conservative than CRC, particularly at small targets, and leaves little of the risk budget unused. Recertification therefore adds only a small gain on top of it, which grows with the target from none at $\alpha = 0.02$ to 0.44 percentage points at $\alpha = 0.15$. In the confidence-adaptive family, envelope CRC coincides with CRC, and the entire improvement comes from recertification. At $\alpha = 0.10$, 500 rounds raise correct specificity from 67.57% to 68.24% and reduce the mean node size from 19.40 to 17.03 labels; at $\alpha = 0.02$, the gain is 1.71 percentage points.

Table 2: Hierarchical classification on CIFAR-100 with $n = 50$, averaged over 300 paired splits. One round and 500 rounds denote recertification of envelope CRC. CS is correct specificity (%) and size is the mean number of labels in the prediction. Time is in milliseconds per build of the reusable common-depth controller and per test image for confidence-adaptive scores.

| | | Common randomized depth | | | | Confidence-adaptive scores | | | |
|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | Method | Risk | CS (%) | Size | Time | Risk | CS (%) | Size | Time |
| 0.02 | CRC | 0.00044 | 0.282 | 99.50 | 9.347 | 0.01219 | 41.903 | 55.06 | 0.003 |
| | Envelope CRC | 0.01961 | 12.538 | 77.96 | 9.962 | 0.01219 | 41.903 | 55.06 | 0.003 |
| | One round | 0.01961 | 12.538 | 77.96 | 9.907 | 0.01324 | 42.205 | 54.45 | 0.004 |
| | 500 rounds | 0.01961 | 12.538 | 77.96 | 194.5 | 0.01814 | 43.613 | 51.58 | 2.885 |
| 0.05 | CRC | 0.03415 | 21.766 | 61.76 | 8.867 | 0.04052 | 60.032 | 32.97 | 0.003 |
| | Envelope CRC | 0.04884 | 30.989 | 45.62 | 8.997 | 0.04052 | 60.032 | 32.97 | 0.003 |
| | One round | 0.04887 | 30.994 | 45.62 | 9.341 | 0.04175 | 60.211 | 32.50 | 0.004 |
| | 500 rounds | 0.04907 | 31.025 | 45.61 | 198.0 | 0.04803 | 61.127 | 30.06 | 5.127 |
| 0.10 | CRC | 0.08748 | 49.556 | 15.46 | 8.877 | 0.09052 | 67.572 | 19.40 | 0.003 |
| | Envelope CRC | 0.09645 | 52.883 | 10.65 | 8.937 | 0.09052 | 67.572 | 19.40 | 0.003 |
| | One round | 0.09700 | 52.956 | 10.64 | 9.275 | 0.09156 | 67.657 | 19.10 | 0.004 |
| | 500 rounds | 0.09898 | 53.220 | 10.58 | 195.2 | 0.09880 | 68.235 | 17.03 | 28.2 |
| 0.15 | CRC | 0.13944 | 61.441 | 3.72 | 9.325 | 0.14208 | 70.550 | 10.77 | 0.003 |
| | Envelope CRC | 0.14462 | 62.279 | 3.37 | 9.157 | 0.14208 | 70.550 | 10.77 | 0.003 |
| | One round | 0.14543 | 62.393 | 3.35 | 9.755 | 0.14279 | 70.597 | 10.59 | 0.004 |
| | 500 rounds | 0.14780 | 62.722 | 3.28 | 201.3 | 0.14955 | 71.014 | 8.92 | 138.2 |

Timings use an Intel Core i7-12700KF with Python/Numba and exclude model inference and compilation warm-up. Common-depth builds and the faster adaptive methods average 20 repetitions; 500-round adaptive timings cover all 4800 test images, including table construction.

Figure 2 shows how the gains beyond envelope CRC accumulate over rounds. A single round recovers between 10% and 26% of the gain after 500 rounds. At $\alpha = 0.10$, for example, one round gains 0.073 percentage points for common depth and 0.085 for confidence-adaptive scores, compared with 0.337 and 0.663 after 500 rounds. The first 50 rounds recover at least 95% of the final gain in every setting, and later rounds change little.

The two families locate the conservativeness of CRC in different places. In the common-depth family,

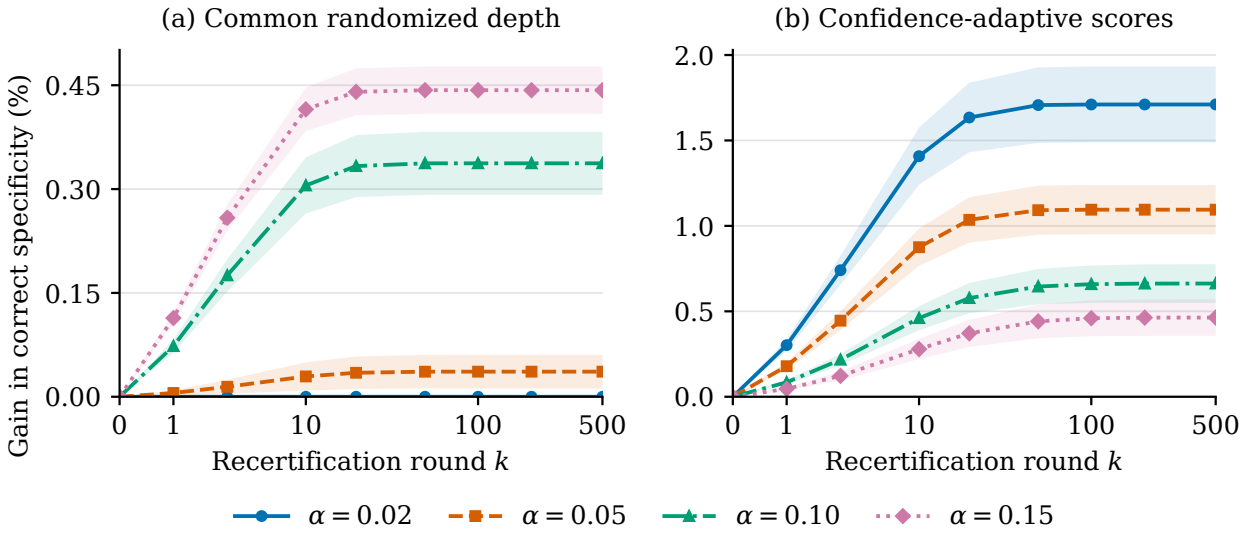



Figure 2: Gains in correct specificity of recertification beyond envelope CRC at $n = 50$. Each panel shows mean paired gains over 300 splits, with shaded pointwise 95% Student-$t$ intervals. The horizontal axis is logarithmic in $k + 1$, where $k$ is the number of rounds.

most of it lies in the bound on the unseen loss: the envelope $U(t) = t/2$ is much smaller than one at shallow depths, and envelope CRC removes most of the slack. In the confidence-adaptive family, the envelope equals the worst-case bound, yet recertification still produces more informative predictions, using budget that no choice of envelope can recover.

The possible increase in attained risk is bounded by the gap between the target level and the baseline risk. Since the risk of a controller $A$ equals $\mathbb{E}\, T_A(S)/(n+1)$ for an augmented sample $S$, the gap between $\alpha$ and the risk is exactly the expected unused budget $\mathbb{E}\, \Delta_A(S)/(n+1)$, which is what recertification spends. In the common-depth family, envelope CRC leaves only about 2% to 4% of the target unused, and at $\alpha = 0.02$ its risk is within sampling error of the target, so there is almost nothing to recover. In the confidence-adaptive family, CRC leaves between 5% and 39% unused, and the gains are correspondingly larger. In this family, the gains also shrink as the target grows, partly because more predictions are already as specific as possible: at $\alpha = 0.15$, 86.3% of the baseline predictions are leaves.

The runtimes reflect the two regimes of Section 3.5. The common-depth controller is a single table, built in about 0.2 seconds for 500 rounds and reused across splits and test images. Confidence-adaptive recertification is computed separately for each test image, and 500 rounds take from about 3 to 140 milliseconds per image, with the cost growing in the target.

# 4 Improving randomized e-BH

We now return to multiple testing, the setting of the e-closure principle that inspired this work [Xu et al., 2026b], and recertify a randomized version of the e-BH procedure. Let $E = (E_1, \ldots, E_K)$ be e-values for hypotheses $H_1, \ldots, H_K$, with arbitrary dependence among them. In the framework of Section 2, the state is the unknown set $T \subseteq \{1, \ldots, K\}$ of true nulls, whose class $\mathcal{P}_T$ consists of all laws under which $\mathbb{E}_P E_j \leq 1$ for every $j \in T$, and errors are measured by the false discovery proportion $\mathrm{FDP}_T(R) = |R \cap T|/(|R| \vee 1)$ of a rejection set $R$, with budget $\alpha$. Section 4.1 recalls how e-BH fits this framework. Section 4.2 casts randomized e-BH as an instance of generalized e-closure, with randomized decisions as actions, and identifies the budget it leaves unused. Section 4.3 recertifies it with the e-collection of e-BH, Section 4.4 with the larger e-collection of closed e-BH, and Section 4.5 reports simulations.

## 4.1 e-BH and self-consistency

For an observed vector $e$, let $e_{[1]} \geq \cdots \geq e_{[K]}$ be its ordered values, let $r(j)$ be the rank of $e_j$, and let $R_k(e)$ be the set of the $k$ hypotheses with the largest e-values, with ties broken by a fixed convention. The e-BH procedure of Wang and Ramdas [2022] rejects $R_k(e)$ for the largest $k$ with $k\, e_{[k]} \geq K/\alpha$. With rejection sets as actions, its validity is explained by the e-collection

$$E_T^{\mathrm{eBH}} = \frac{1}{K} \sum_{j \in T} E_j, \tag{4.1}$$

whose expectation under $\mathcal{P}_T$ is at most $|T|/K \leq 1$. Certification generally imposes one constraint for each state, here one for each of the $2^K$ sets of true nulls, but for this e-collection the constraints collapse to $K$ simple ones. Since both sides of $\mathrm{FDP}_T(R) \leq \alpha E_T^{\mathrm{eBH}}$ are sums over the hypotheses in $T$, the condition holds for every $T$ if and only if it holds for every singleton $T = \{j\}$, that is, if and only if $e_j \geq K/(\alpha|R|)$ for every $j \in R$. This is the self-consistency condition of Blanchard and Roquain [2008], which Wang and Ramdas [2022] used to establish the validity of e-BH under arbitrary dependence. The certified sets are thus exactly the self-consistent ones, and e-BH is the largest of them (Appendix C.1).

## 4.2 Randomized e-BH as generalized e-closure

The e-BH procedure can be conservative, because an e-value contributes only once it crosses one of the thresholds $K/(\alpha k)$, and an e-value just below its threshold is wasted. Xu and Ramdas [2023] showed that external randomization recovers part of this loss. The most powerful variant in their simulations, U-eBH, draws a single $U \sim \mathrm{Unif}(0,1)$ independent of $E$, boosts the e-values to $(E_1/U, \ldots, E_K/U)$, and applies e-BH to the boosted values; that is, it rejects $R_{\kappa_0}(E)$ with

$$\kappa_0(e, U) = \max\big(\{k : U \leq \alpha k\, e_{[k]}/K\} \cup \{0\}\big).$$

U-eBH rejects at least the e-BH set for every value of $U$ and controls the FDR at level $\alpha$. Its validity is not a consequence of applying e-BH to e-values, since $E_j/U$ has infinite expectation whenever $\mathbb{P}(E_j > 0) > 0$, but it becomes transparent once the randomization is made part of the action. A second variant, J-eBH, divides each e-value by its own uniform random variable drawn independently of the others and of $E$, and applies e-BH to the boosted values; it is also valid (Appendix C.3).

**Actions and losses.** Like e-BH, U-eBH rejects the hypotheses with the largest e-values, and only the number of rejections is random. We therefore consider *monotone* rules, which reject $R_\kappa(e)$ for a random number $\kappa$ of rejections, and take an action to be a distribution $p = (p_0, \ldots, p_K)$ of $\kappa$. A *boosting rule* chooses such a distribution $p(e)$ after observing the e-values, draws $\kappa \sim p(e)$ independently of the data, and rejects $R_\kappa(e)$. The loss of the action $p$ in state $T$ is the false discovery proportion averaged over this draw,

$$\bar{L}(T, p; e) = \sum_{k=1}^{K} p_k \, \frac{|T \cap R_k(e)|}{k}. \tag{4.2}$$

This loss depends on the observation through the ranking of the e-values, which the framework of Section 2 allows, and a boosting rule is valid in the sense of (2.1) exactly when the procedure it implements controls the FDR at level $\alpha$. Randomization can thus also be viewed as part of the action, with probability distributions as the actions, as described in Section 2.2. U-eBH is the boosting rule whose distribution $p^0(e)$ has the tails

$$\mathbb{P}(\kappa_0 \geq r \mid E = e) = \min\left\{1, \frac{\alpha}{K} \max_{k \geq r} k\, e_{[k]}\right\}, \qquad r = 1, \ldots, K. \tag{4.3}$$

Indeed, dividing every e-value by the same $U$ preserves their ranking, so e-BH applied to the boosted values rejects $k$ hypotheses when $k$ is the largest number satisfying $k\, e_{[k]}/U \geq K/\alpha$, that is, $U \leq \alpha k\, e_{[k]}/K$. Hence $\kappa_0 \geq r$ exactly when $U \leq \alpha k\, e_{[k]}/K$ for some $k \geq r$, equivalently when $U \leq (\alpha/K) \max_{k \geq r} k\, e_{[k]}$, and because $U$ is uniform on $(0,1)$ and independent of $E$, this event has the conditional probability (4.3).

**The certified menu.** The hypothesis of rank $r$ is rejected exactly when $\kappa \geq r$, and it then receives the share $1/\kappa$ of the false discovery proportion, if it is null. The loss (4.2) therefore splits over the hypotheses in $T$,

$$\bar{L}(T, p; e) = \sum_{j \in T} c_{[r(j)]}(p), \qquad c_{[r]}(p) = \sum_{k \geq r} \frac{p_k}{k},$$

where $c_{[r]}(p)$ is the expected share charged to the hypothesis of rank $r$, should it be null. Both sides of $\bar{L}(T, p; e) \leq \alpha E_T^{\mathrm{eBH}}(e)$ are again sums over the hypotheses in $T$, so the reduction of Section 4.1 applies verbatim, and the certified menu of $E^{\mathrm{eBH}}$ at $e$ is

$$\mathcal{C}_\alpha(E^{\mathrm{eBH}}; e) = \left\{ p : \; c_{[r]}(p) \leq \frac{\alpha}{K} e_{[r]} \;\; \text{for } r = 1, \ldots, K \right\}. \tag{4.4}$$

Each hypothesis may thus be charged at most its own share of the budget, a self-consistency condition that need only hold on average over the randomization. The menu is a polytope of distributions, and by Theorem 2.1 every boosting rule that selects from it has FDR at most $\alpha|T|/K \leq \alpha$.

U-eBH selects from the menu (4.4), which proves its validity, but its distribution usually satisfies the constraints with strict inequality. An intuitive explanation is that the common boosting factor $U$ does not respect the different slack structures. Indeed, dividing every e-value by the same $U$ ties the hypotheses together: U-eBH rejects at least $r$ hypotheses when e-BH, applied to the boosted values, selects some set $R_k(e)$ with $k \geq r$, which happens when $U \leq \alpha k \, e_{[k]}/K$ and therefore depends only on the smallest e-value $e_{[k]}$ in that set. As a result, a hypothesis can be charged as if its e-value were that of a lower-ranked hypothesis, and part of its budget then goes unused (Appendix C.1).

## 4.3 Recertifying U-eBH

Recertification spends this unused budget. Like every valid rule, U-eBH is certified by its canonical e-collection, its averaged loss divided by the budget, $\bar{L}(T, p^0(e); e)/\alpha$. As shown above, this e-collection lies below $E^{\mathrm{eBH}}$, hypothesis by hypothesis, and the gap is exactly the unused budget. Recertification replaces the canonical e-collection by the larger e-collection $E^{\mathrm{eBH}}$. This step involves no unknown conditional expectation: the input e-values supply the budget directly, and the averaged losses are known because the law of the external randomization is known (Appendix C.1 spells out the correspondence with Proposition 2.2). The certified menu of the recertified e-collection is (4.4). It contains U-eBH, but also distributions that do not arise from boosting all e-values by a common factor, and any of them may be selected. We select the one that rejects the most.

To find this distribution, we turn the budget constraints in (4.4), which bound the charges of individual hypotheses, into bounds on the quantities that measure power: the probabilities $\mathbb{P}(\kappa \geq r)$ of rejecting at least $r$ hypotheses. Recall that the hypothesis of rank $j$ is rejected exactly when $\kappa \geq j$, and then receives the share $1/\kappa$, so its charge is $c_{[j]}(p) = \mathbb{E}[\mathbb{1}\{\kappa \geq j\}/\kappa]$. On an event $\{\kappa = k\}$ with $k \geq r$, the shares of the $k$ rejected hypotheses sum to one: the top $r$ hypotheses receive $r/k$ together, and each hypothesis of rank $j \in \{r+1, \ldots, k\}$ receives $1/k$. Averaging over the events $\{\kappa = k\}$ with $k \geq r$, the first part contributes $r\,\mathbb{E}[\mathbb{1}\{\kappa \geq r\}/\kappa] = r\, c_{[r]}(p)$, and the second contributes $\sum_{j>r} \mathbb{E}[\mathbb{1}\{\kappa \geq j\}/\kappa] = \sum_{j>r} c_{[j]}(p)$, because a hypothesis of rank $j > r$ is rejected only on these events. Hence

$$\mathbb{P}(\kappa \geq r) = r\, c_{[r]}(p) + \sum_{j>r} c_{[j]}(p).$$

Bounding each charge by its budget now bounds the probability of rejecting at least $r$ hypotheses directly. The top $r$ hypotheses can contribute at most $r$ times the budget of the hypothesis of rank $r$, and each lower-ranked hypothesis at most its own budget. Every distribution in the certified menu (4.4) therefore obeys

$$\mathbb{P}(\kappa \geq r \mid E = e) \leq q_r^*(e) = \min\left\{ 1, \frac{\alpha}{K}\left( r\, e_{[r]} + \sum_{j>r} e_{[j]} \right) \right\}, \qquad r = 1, \ldots, K. \tag{4.5}$$

**Algorithm 2** Recertified U-eBH

**Input:** E-values $e_1, \dots, e_K$, level $\alpha$, and $U \sim \mathrm{Unif}(0,1)$ drawn independently of the e-values.

Order the e-values as $e_{[1]} \geq \cdots \geq e_{[K]}$, with ties broken by a fixed convention.
Compute the partial sums $s_r = \sum_{j=r}^{K} e_{[j]}$ for $r = 1, \dots, K$, and set $s_{K+1} = 0$.
**for** $r = 1, \dots, K$ **do**
    Set $q_r^* = \min\{1, (\alpha/K)(r\, e_{[r]} + s_{r+1})\}$.
**end for**
Set $\kappa^* = \max(\{r : U \leq q_r^*\} \cup \{0\})$.

**Output:** The rejection set $R_{\kappa^*}(e)$, consisting of the $\kappa^*$ hypotheses with the largest e-values.

These bounds depend on the e-values alone, and a single distribution attains all $K$ of them at once; when $(\alpha/K)\sum_j e_j \leq 1$, it makes every constraint in (4.4) tight (Appendix C.2). The recertified rule selects this distribution and implements it with the seed of U-eBH, as summarized in Algorithm 2. Computation requires only sorting the e-values and forming partial sums. U-eBH also selects from the menu, so its tails (4.3) lie below $q_r^*$ – they value the $k$ rejected hypotheses as if each had the smallest e-value among them.

The following proposition collects the properties of the recertified rule, which we regard as recertified randomized e-BH or, more precisely, recertified U-eBH (RU-eBH).

**Proposition 4.1** (Recertified U-eBH)**.** *RU-eBH has the following properties.*

*(i) Its FDR is at most $\alpha|T|/K \leq \alpha$ in every state $T$ and under every law in $\mathcal{P}_T$.*

*(ii) It contains the U-eBH rejection set, and hence the e-BH rejection set, for every $e$ and $U$.*

*(iii) Among boosting rules that select from (4.4), it has the stochastically largest number of rejections: $\mathbb{P}(\kappa \geq r \mid E = e) \leq \mathbb{P}(\kappa^* \geq r \mid E = e)$ for every $r$ and $e$.*

In particular, the probability of rejecting at least one hypothesis rises from $(\alpha/K)\max_k k\, e_{[k]}$ to $(\alpha/K)\sum_j e_j$, capped at one. RU-eBH draws on the budget of every hypothesis, whereas U-eBH draws on that of a single rejection set. We also note that restricting attention to monotone rules costs nothing in the expected number of rejections: for every $e$, no randomized rule over arbitrary rejection sets that is certified by $E^{\mathrm{eBH}}$ rejects more hypotheses in expectation than RU-eBH, although such rules need not be stochastically smaller (Proposition C.3 in Appendix C.3). Finally, RU-eBH can be written as e-BH applied to the randomized e-values $(K/\alpha)\,\mathbb{1}\{j \in R_{\kappa^*}\}/(|R_{\kappa^*}| \vee 1)$ (Corollary C.2).

## 4.4 Recertifying closed e-BH

The optimality in Proposition 4.1 is relative to the e-collection $E^{\mathrm{eBH}}$, and a larger e-collection permits a further improvement. A simple example exploits the fact that the bound $\alpha|T|/K$ in Proposition 4.1(i) is strictly below $\alpha$ unless every hypothesis is null. The minimally adaptive principle [Solari and Goeman, 2017, Ignatiadis et al., 2024] uses this slack: it runs a procedure at the inflated level $K\alpha/(K-1)$, but only when the procedure at level $\alpha$ rejects at least one hypothesis, which keeps the FDR under the global null at most $\alpha$. In our framework, this corresponds to replacing $1/K$ by $1/(K-1)$ in (4.1) for every state other than the full set, which still gives an e-collection because such a state contains at most $K-1$ true nulls. Applying recertification to this collection would therefore yield a more powerful rule, and, as a result, RU-eBH is not admissible among all FDR procedures (Appendix C.5).

A much larger e-collection is introduced by Xu et al. [2026b],

$$E_T^{\overline{\mathrm{eBH}}} = \frac{1}{|T|}\sum_{j \in T} E_j \qquad (T \neq \varnothing), \tag{4.6}$$

whose expectation under $\mathcal{P}_T$ is at most one and which certifies the rejection set of closed e-BH. It normalizes by the number of true nulls rather than by $K$, and thus adapts to the unknown proportion of true nulls.

For a monotone rule, the averaged false discovery proportion in state $T$ is still the sum $\sum_{j\in T} c_{[r(j)]}(p)$ of the charges, but the budget $\alpha E_T^{\overline{\mathrm{eBH}}}$ no longer splits into budgets for individual hypotheses. The certified menu of $E^{\overline{\mathrm{eBH}}}$ at $e$ therefore consists of the distributions $p$ that satisfy all $2^K-1$ constraints

$$\sum_{j\in T} c_{[r(j)]}(p) \le \frac{\alpha}{|T|}\sum_{j\in T} e_j \qquad (T\neq\varnothing), \tag{4.7}$$

which do not reduce to $K$ separate ones.

Nevertheless, the menu has enough structure that the rule with the largest expected number of rejections can be found in closed form. The key is to describe a monotone rule by its cumulative charges from the bottom. Let $S_s(p)=\sum_{r\ge s} c_{[r]}(p)$ be the total charge of the $K-s+1$ hypotheses with the smallest e-values. A hypothesis is rejected whenever a lower-ranked one is, with the same share, so the charges are non-increasing in the rank and $s\mapsto S_s(p)$ is convex. Conversely, every non-increasing convex sequence with $S_{K+1}=0$ and $S_1\le 1$ arises from a unique distribution $p$. The constraint (4.7) for the set $\{s,\dots,K\}$ caps $S_s(p)$ at $\alpha m_s$, where $m_s=(K-s+1)^{-1}\sum_{j\ge s} e_{[j]}$ is the mean e-value of that set, and $S_1(p)=\mathbb{P}(\kappa\ge 1)$ is also at most one. Let $\beta_1=\min\{1,\alpha m_1\}$, $\beta_s=\alpha m_s$ for $2\le s\le K$ and $\beta_{K+1}=0$. These caps need not form a convex sequence, but among the convex sequences below them there is a pointwise largest one, their greatest convex minorant $S^\sharp$, which is the lower convex hull of the points $(s,\beta_s)$. Since the expected number of rejections, $S_1(p)+2\sum_{s\ge 2} S_s(p)$, increases with every $S_s(p)$, we take the rule whose cumulative charges are $S^\sharp$; Proposition 4.2 shows that it satisfies all the other constraints in (4.7) as well. Its tails are

$$q_r^\sharp = r\,S_r^\sharp-(r-1)\,S_{r+1}^\sharp, \qquad r=1,\dots,K. \tag{4.8}$$

Geometrically, $q_r^\sharp$ is the value at $s=1$ of the line that extends the segment of the minorant over $[r,r+1]$; the same construction applied to the convex sequence $(\alpha/K)\sum_{j\ge s} e_{[j]}$ of cumulative budgets of $E^{\mathrm{eBH}}$ gives the bounds $q_r^*$ of (4.5) before capping. We call the resulting procedure *recertified closed e-BH* (RC-eBH), since it is certified by the e-collection of closed e-BH. It replaces $q_r^*$ by $q_r^\sharp$ in Algorithm 2, that is, it rejects $R_{\kappa^\sharp}(e)$ with $\kappa^\sharp=\max(\{r: U\le q_r^\sharp\}\cup\{0\})$. Since the points $(s,\beta_s)$ are ordered in $s$, the minorant is computed in $O(K)$ operations once the e-values are sorted.

**Proposition 4.2** (Recertified closed e-BH)**.** *RC-eBH has the following properties.*

*(i) Its FDR is at most $\alpha$ in every state $T$ and under every law in $\mathcal{P}_T$.*

*(ii) Its tails satisfy $q_r^\sharp\ge q_r^*$ for every $r$, so it contains the rejection set of RU-eBH, and hence those of U-eBH and e-BH, for every $e$ and $U$.*

*(iii) Among all randomized rules that select from the certified menu of $E^{\overline{\mathrm{eBH}}}$, monotone or not, it has the largest expected number of rejections given $E=e$. In particular, it rejects at least as many hypotheses in expectation as closed e-BH.*

The proof, which we defer to Appendix C.4, shows that, for the charges of the minorant, the hardest constraints of each size are those for the sets of lowest-ranked hypotheses, so the constraints for these sets imply all others in (4.7). Unlike the menu of $E^{\mathrm{eBH}}$, the menu of the mean e-collection generally has no stochastically largest element, which is why part (iii) concerns the expected number of rejections. Whereas the FDR of RU-eBH is bounded by $\alpha|T|/K$, that of RC-eBH is bounded by $\alpha$.

**Remark 4.3** (Conditional calibration)**.** Conditional calibration [Fithian and Lei, 2022, Lee and Ren, 2024] is another way of improving e-BH, and it follows the same pattern with a different source of budget. We only recast the e-BH version of Lee and Ren [2024] in the present terms without improving or modifying it. Its baseline is the part of each e-value that e-BH actually uses. With $R_0$ the e-BH rejection set, the pruned e-value $h_j=K\,\mathbb{1}\{j\in R_0\}/\{\alpha(|R_0|\vee 1)\}\le E_j$ leads e-BH to the same rejections. Suppose that under $H_j$ a statistic $S_j$ makes the conditional law of the data given $S_j$ known. The budget left unused by the pruned e-value, $\mathbb{E}[E_j-h_j\mid S_j]$, is then known, and Lee and Ren [2024] spend it by raising $h_j$ to a larger e-value whose conditional expectation given $S_j$ does not exceed that of $E_j$. This is Proposition 2.2 in e-value

units, with conditioning statistic $S_j$ and budget $\mathbb{E}[E_j \mid S_j]$, and e-BH applied to the new e-values rejects at least $R_0$. Conditional calibration thus spends unused budget across the values of the data that share the same $S_j$, whereas RU-eBH spends it across the values of the external seed. Appendix C.6 gives the exact correspondence.

**Remark 4.4** (Selective inference)**.** RU-eBH also applies to settings where the set of true nulls is allowed to be random. Every distribution in the menu (4.4) satisfies $\bar{L}(T, p; e) \leq (\alpha/K) \sum_{j \in T} e_j$ for every $T$, so with a seed independent of the data and of $T$, RU-eBH controls the FDR whenever $\mathbb{E}[E_j \mathbb{1}\{j \in T\}] \leq 1$ for every $j$. This is the setting of conformal selection under dependence [Jin and Candès, 2023, Bai and Jin, 2024]. There, deterministic pruning is e-BH applied to such e-values, and the same calculation shows that pruning with a common uniform or with independent uniforms is U-eBH or J-eBH, respectively. RU-eBH applied to the same e-values therefore selects a superset of U-eBH for every realization of the uniform and, since J-eBH is also certified by $E^{\mathrm{eBH}}$, at least as many test points as J-eBH in expectation (Proposition C.3). The argument does not extend to RC-eBH, because the mean e-collection (4.6) normalizes by the random number of true nulls.

## 4.5 Simulations

We compare RU-eBH and RC-eBH with six procedures that, like them, control the FDR under arbitrary dependence. Four are deterministic: e-BH, its minimally adaptive version (MA-eBH) [Ignatiadis et al., 2024], closed e-BH [Xu et al., 2026b] and anchored e-BH [Ramdas, 2026]. Two are randomized: U-eBH [Xu and Ramdas, 2023] and J-eBH (Section 4.2).

There are $K = 100$ hypotheses, of which 10, 50 or 90 are false. For each hypothesis we observe a Gaussian statistic $Z_j$ with unit variance, mean 0 under the null and a common signal mean between 1 and 4 otherwise; the statistics are equicorrelated with correlation $\rho \in \{0, 0.5, 0.9\}$. All methods use the same e-values, the likelihood ratios $E_j = \exp(2Z_j - 2)$ of $N(2, 1)$ against $N(0, 1)$, at level $\alpha = 0.1$. We report power, the expected fraction of false nulls that are rejected, and the FDR, each averaged over 50,000 data sets. Figure 3 shows all settings. The standard errors are very small, at most 0.0023 for power and 0.0013 for the FDR, so they are not shown.

With 10 false nulls, RU-eBH and RC-eBH are the most powerful methods in every panel, and they nearly coincide. At signal mean 2.5 with independent statistics, for example, RU-eBH has power 0.32, compared with 0.22 for U-eBH and J-eBH and at most 0.06 for the deterministic procedures. Its FDR stays close to the bound $\alpha|T|/K = 0.09$ of Proposition 4.1 with independent statistics, whereas the FDR of U-eBH is between about 40% and 70% of this bound, which reflects the budget that U-eBH leaves unused. J-eBH has nearly the same FDR as RU-eBH but much lower power, so the gain of RU-eBH also comes from how it allocates the budget rather than from how much of the budget it spends.

With more false nulls, RU-eBH remains at least as powerful as U-eBH and J-eBH, but the deterministic procedures based on larger e-collections catch up. With 50 false nulls and independent statistics, closed and anchored e-BH are more powerful than RU-eBH at strong signals, and with 90 false nulls, closed e-BH is more powerful at moderate and strong signals unless the correlation is high. This is consistent with the theory: the optimality of RU-eBH is relative to $E^{\mathrm{eBH}}$, which normalizes by $K$, whereas closed e-BH is certified by the mean e-collection, which normalizes by $|T|$ and is much larger when few hypotheses are null. RC-eBH, which is certified by the same e-collection as closed e-BH, removes this gap. As part (ii) of Proposition 4.2 guarantees, it is more powerful than RU-eBH in every setting, and the gain grows with the number of false nulls; with 90 false nulls, independent statistics and signal mean 2, its power is 0.64, compared with 0.52 for RU-eBH and 0.55 for closed e-BH. It is the most powerful of all eight methods, except at the strongest signals with independent statistics, where several methods reject nearly all false nulls and it trails the best of them by less than 0.005. Its FDR stays below $\alpha$ but exceeds $\alpha|T|/K$ when many hypotheses are false, as the mean e-collection permits.

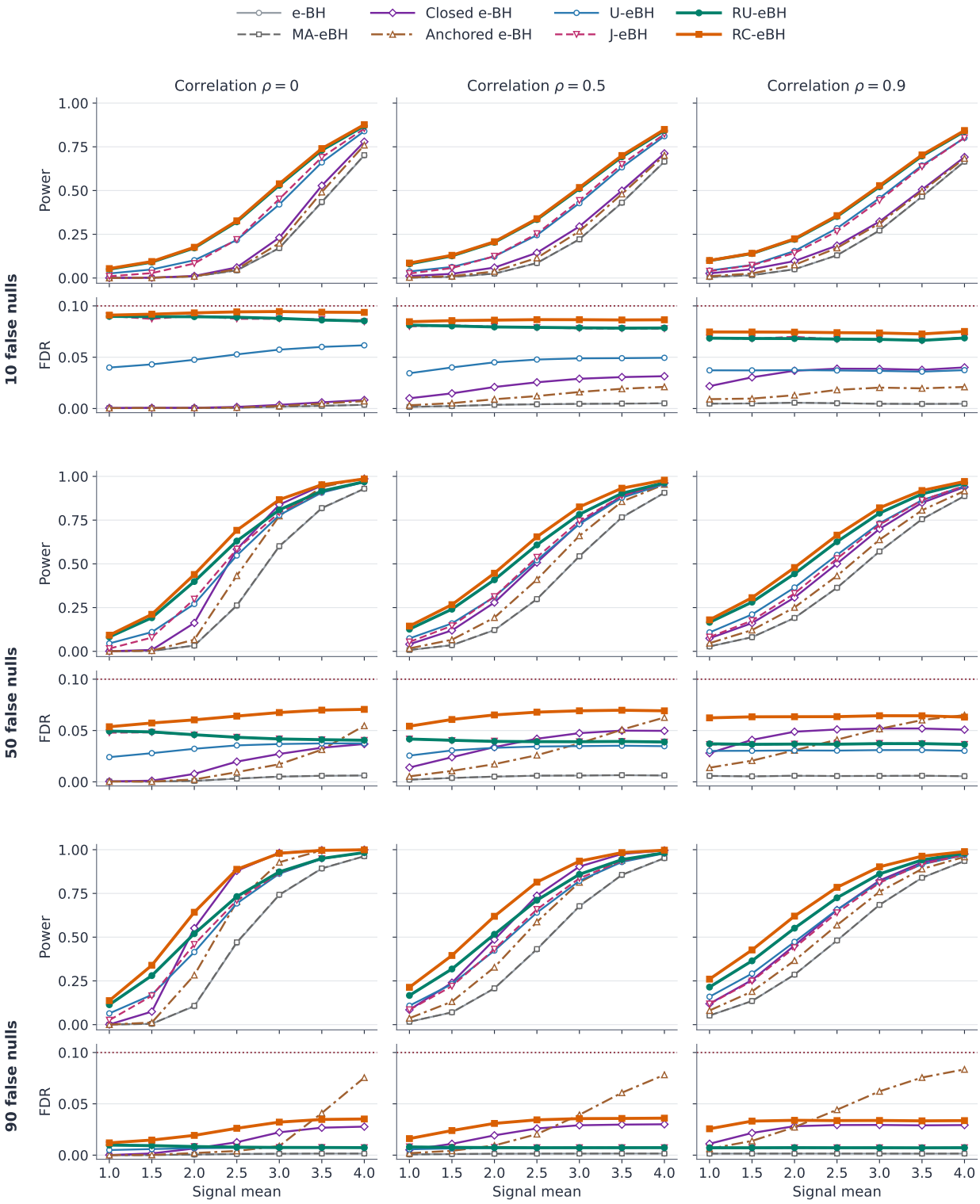



Figure 3: Power and FDR of RU-eBH, RC-eBH and six other procedures that are valid under arbitrary dependence, with 10, 50 and 90 false nulls among $K = 100$ hypotheses at $\alpha = 0.1$ and three levels of correlation (columns). Each point averages 50,000 data sets. The dotted lines mark $\alpha$.

# 5 Discussion

In both applications, recertification enlarges the menu of certified actions, and we selected its most liberal element – the least conservative controller in CRC, and the boosting rule that rejects the most hypotheses in multiple testing. Nothing in the framework requires this choice. By Theorem 2.1, every data-dependent selection from a certified menu is valid, so one may instead select the action that maximizes a utility, for instance a posterior expected utility under a prior. Such a selection is Bayes optimal among selections from the menu, and its validity does not depend on the prior. This differs from the constrained problems of Wald [1950] and Hodges and Lehmann [1952], which optimize a single rule subject to the risk constraints and, by Theorem 2.1, amount to optimizing the e-collection and the selection jointly. Similarly, our admissibility results compare procedures on every calibration sample or conditionally on the e-values; they do not address classical risk admissibility, which allows a rule to trade larger losses at some observations for smaller losses at others.

In summary, this paper formulates the e-closure principle for general decision problems and uses the resulting correspondence constructively. Recertification starts by measuring the budget that a valid procedure leaves unused, spends it through larger e-values, and inverts them into an improved procedure with the same guarantee. The source of this budget differs between our two applications. In conformal risk control, exchangeability makes the budget computable on every hypothetical augmented sample, and recertification yields pointwise improvements of CRC and an admissible limit. In multiple testing, the e-values themselves supply the budget once randomized procedures are treated as randomized actions, and recertifying U-eBH yields two closed-form procedures, RU-eBH and RC-eBH, each optimal among the procedures certified by its e-collection.

The two applications also differ in cost. RU-eBH and RC-eBH require only sorting the e-values and, for RC-eBH, a convex minorant, whereas repeated recertification of CRC trades computation for a uniform improvement, with diminishing returns, and can be expensive when the controller depends on individual scores. Exploiting the structure of the loss and the data to compute stronger certificates economically remains an important practical question, and whether RC-eBH is admissible among all FDR procedures is open. More broadly, any procedure whose validity proof leaves an identifiable budget unused is a candidate for recertification, and we expect further applications of this kind, especially those with more complex states, actions or loss structures.

# References


I. Andrews and J. Chen. Certified decisions. *arXiv preprint arXiv:2502.17830*, 2025.

A. N. Angelopoulos. Conformal risk control for non-monotonic losses. *arXiv preprint arXiv:2602.20151*, 2026.

A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

A. N. Angelopoulos, S. Bates, E. J. Candès, M. I. Jordan, and L. Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025.

T. Bai and Y. Jin. Optimized conformal selection: Powerful selective inference after conformity score optimization. *arXiv preprint arXiv:2411.17983*, 2024.

T. Bai and Y. Jin. Conformal selective prediction with general risk control. *arXiv preprint arXiv:2603.24704*, 2026.

S. Bates, A. Angelopoulos, L. Lei, J. Malik, and M. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):Article 43, 2021.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995.

J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, second edition, 1985.

G. Blanchard and E. Roquain. Two simple sufficient conditions for FDR control. *Electronic Journal of Statistics*, 2:963–992, 2008.

B. Chugg, T. Lardy, A. Ramdas, and P. D. Grünwald. On admissibility in post-hoc hypothesis testing. *International Journal of Approximate Reasoning*, 191:109634, 2026.

K. M. Cohen, S. Park, O. Simeone, and S. Shamai. Cross-validation conformal risk control. In *IEEE International Symposium on Information Theory*, 2024.

A. Farinhas, C. Zerva, D. Ulmer, and A. F. T. Martins. Non-exchangeable conformal risk control. In *International Conference on Learning Representations*, 2024.

W. Fithian and L. Lei. Conditional calibration for false discovery rate control under dependence. *The Annals of Statistics*, 50(6):3091–3118, 2022.

E. Gauthier, F. Bach, and M. I. Jordan. E-values expand the scope of conformal prediction. *arXiv preprint arXiv:2503.13050*, 2025.

J. J. Goeman. A uniform improvement of the Benjamini–Hochberg procedure via e-closure. *arXiv preprint arXiv:2606.01854*, 2026.

J. J. Goeman and A. Solari. Multiple testing for exploratory research. *Statistical Science*, 26(4):584–597, 2011.

J. J. Goeman, J. Hemerik, and A. Solari. Only closed testing procedures are admissible for controlling false discovery proportions. *The Annals of Statistics*, 49(2):1218–1238, 2021.

P. González-González. Embedded representations of popular image datasets (CIFAR10, CIFAR100, CIFAR100coarse, MNIST, FashionMNIST, SVHN). Zenodo dataset, version 1, https://doi.org/10.5281/zenodo.21131944, 2026.

P. D. Grünwald. The e-posterior. *Philosophical Transactions of the Royal Society A*, 381(2247):20220146, 2023.

P. 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.

J. L. Hodges, Jr. and E. L. Lehmann. The use of previous experience in reaching statistical decisions. *The Annals of Mathematical Statistics*, 23(3):396–407, 1952.

N. Ignatiadis, R. Wang, and A. Ramdas. E-values as unnormalized weights in multiple testing. *Biometrika*, 111(2):417–439, 2024.

Y. Jin and E. J. Candès. Model-free selective inference under covariate shift via weighted conformal p-values. *arXiv preprint arXiv:2307.09291*, 2023.

S. Kiyani, G. J. Pappas, A. Roth, and H. Hassani. Decision theoretic foundations for conformal prediction: Optimal uncertainty quantification for risk-averse agents. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 30943–30965, 2025.

N. W. Koning. Post-hoc $\alpha$ hypothesis testing and the post-hoc $p$-value. *arXiv preprint arXiv:2312.08040*, 2023.

N. W. Koning. Measuring evidence against exchangeability and group invariance with e-values. *The Annals of Statistics*, 54(4):1843–1869, 2026.

N. W. Koning and S. van Meer. Fuzzy prediction sets: Conformal prediction with e-values. *arXiv preprint arXiv:2509.13130*, 2026.

J. Lee and Z. Ren. Boosting e-BH via conditional calibration. *arXiv preprint arXiv:2404.17562*, 2024.

E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer, fourth edition, 2022.

J. Lekeufack, A. N. Angelopoulos, A. Bajcsy, M. I. Jordan, and J. Malik. Conformal decision theory: Safe autonomous decisions from imperfect predictions. In *IEEE International Conference on Robotics and Automation*, 2024.

R. Marcus, E. Peritz, and K. R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.

A. Ramdas. Anchored multiple testing: A transparent use of e-closure to improve FDR procedures. *arXiv preprint arXiv:2610.11279*, 2026.

A. Ramdas and R. Wang. Hypothesis testing with e-values. *Foundations and Trends in Statistics*, 1(1–2): 1–390, 2025.

A. Solari and J. J. Goeman. Minimally adaptive BH: A tiny but uniform improvement of the procedure of Benjamini and Hochberg. *Biometrical Journal*, 59(4):776–780, 2017.

E. Sonnemann. General solutions to multiple testing problems. *Biometrical Journal*, 50(5):641–656, 2008.

S. M. Srivastava. *A Course on Borel Sets*, volume 180 of *Graduate Texts in Mathematics*. Springer, 1998.

L. Sun and R. Wang. Admissibility and complete classes for false discovery rate control with e-values. *arXiv preprint arXiv:2607.14380*, 2026.

R. J. Tibshirani, R. F. Barber, and A. Ramdas. Conformal prediction through the lens of hypothesis testing: Universality, impossibility, and optimality. *arXiv preprint arXiv:2608.27310*, 2026.

V. Vovk. Conformal e-prediction. *Pattern Recognition*, 166:111674, 2025.

V. Vovk and R. Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49 (3):1736–1754, 2021.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer, second edition, 2022.

A. Wald. *Statistical Decision Functions*. John Wiley & Sons, 1950.

R. Wang and A. Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society: Series B*, 84(3):822–852, 2022.

Z. Xu and A. Ramdas. More powerful multiple testing under dependence via randomization. *arXiv preprint arXiv:2305.11126*, 2023.

Z. Xu, L. Fischer, and A. Ramdas. Improving online FDR procedures via online analogs of e-closure and compound e-values. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pages 7552–7561, 2026a.

Z. Xu, A. Solari, L. Fischer, R. de Heide, A. Ramdas, and J. Goeman. Bringing closure to false discovery rate control: A general principle for multiple testing. *arXiv preprint arXiv:2509.02517*, 2026b.

C. Yeh, N. Christianson, A. Wierman, and Y. Yue. Conformal risk training: End-to-end optimization of conformal risk control. In *Advances in Neural Information Processing Systems*, 2025.

# A Measurability of the certified menu

The proof of Theorem 2.1 fixes one state–loss pair before taking expectations, so it requires only that the rule and the losses be measurable. The certified menu is different: it intersects possibly uncountably many constraints, which in general need not produce a Borel graph, and selecting from it requires measurable selections. The following conditions are sufficient, though not necessary.

**Proposition A.1.** *Suppose that $\mathcal{X}$ is standard Borel, $\mathcal{S}$ and $\mathcal{T}$ are separable metric spaces, and $\mathcal{A}$ is a compact metric space. Suppose that the losses and the e-collection are finite, that $\alpha_t(s) > 0$, and that both $L_t(s, a; x)$ and*

$$f(x, s, t, a) = L_t(s, a; x) - \alpha_t(s) E_{s,t}(x)$$

*are Borel in $x$ for each fixed $(s, t, a)$ and continuous in $(s, t, a)$ for each fixed $x$. If the certified menu is nonempty for every $x$, then it has a Borel graph and nonempty compact values, and it admits a Borel selection. A utility $v(x, a)$ that is Borel in $x$ and continuous in $a$ has a Borel maximizing selection from the menu. Every loss envelope of the menu is Borel in $x$ and continuous in $(s, t)$.*

*Proof.* Choose a countable dense set $\{(s_j, t_j) : j \geq 1\}$ in $\mathcal{S} \times \mathcal{T}$. By continuity,

$$\mathcal{C}_\alpha(E; x) = \bigcap_{j \geq 1} \{a : f(x, s_j, t_j, a) \leq 0\}.$$

Each constraint is jointly Borel in $(x, a)$, so the graph is Borel, and its sections are closed, hence compact. The Arsenin–Kunugui theorem [Srivastava, 1998, Theorem 5.12.1] then gives a Borel selection. For a utility $v$, the sets $\{(x, a) : a \in \mathcal{C}_\alpha(E; x),\ v(x, a) > b\}$ have $\sigma$-compact sections, so their projections are Borel by the same theorem, and the maximum value is Borel in $x$. The maximizing graph is therefore Borel with nonempty compact sections, and a further application gives a maximizing selection. The envelopes are Borel by the same argument applied to each fixed-pair loss, and continuous in $(s, t)$ as maxima of jointly continuous functions over the compact set $\mathcal{C}_\alpha(E; x)$. □

# B Details for conformal risk control

## B.1 CRC through e-closure

We work in the formulation of Section 3.3, in which a state is an augmented multiset $S = (z_1, \ldots, z_{n+1})$ and $\mathcal{P}_S$ deletes a uniformly random point of $S$. For a fixed action $\lambda$, the losses $\ell(z_i, \lambda)$ act as scores of the points of $S$, and

$$E^\lambda_S(D) = \frac{(n+1)\,\ell(S \setminus D, \lambda)}{\sum_{i=1}^{n+1} \ell(z_i, \lambda)} \qquad (D \subseteq S),$$

with $0/0 = 0$, is the conformal e-value of the held-out point [Vovk, 2025, Koning, 2026]. Its expectation under $\mathcal{P}_S$ is at most one, so $E^\lambda$ is an e-collection for each fixed $\lambda$.

Theorem 2.1 certifies all actions with a single e-collection. Here we instead allow each action $\lambda$ to be certified by its own e-values $E^\lambda$, and consider, at an observed $D$, the actions that satisfy

$$\ell(z, \lambda) \leq \alpha E^\lambda_{D \cup \{z\}}(D) \qquad \text{for every } z \in \mathcal{Z}. \tag{B.1}$$

The constraint is trivial when $\ell(z, \lambda) = 0$, and otherwise it reads $\sum_{i=1}^n \ell(Z_i, \lambda) + \ell(z, \lambda) \leq (n+1)\alpha$. Hence (B.1) holds if and only if

$$\sum_{i=1}^n \ell(Z_i, \lambda) + U^*(\lambda) \leq (n+1)\alpha, \qquad U^*(\lambda) = \sup_{z \in \mathcal{Z}} \ell(z, \lambda),$$

which is the envelope CRC test (3.3) with the sharp envelope $U^*$, non-increasing and right-continuous as a supremum of such functions. A larger envelope $U$, or the constant $B$ of standard CRC, replaces the unobserved test loss $\ell(z,\lambda)$ in the denominator of $E^\lambda$ by an upper bound and gives a more stringent test. Envelope CRC and CRC therefore also select actions that satisfy (B.1).

In general, letting the e-values depend on the action is not valid: choosing the action after seeing the data then amounts to choosing among e-values, which inflates them. The conformal case avoids this due to a cancellation. Under $\mathcal{P}_S$ the augmented sample $S$ is fixed, and the data reveal only which of its points is held out. For a held-out point with positive loss, the constraint in (B.1) reads $\sum_{j=1}^{n+1}\ell(z_j,\lambda)\le(n+1)\alpha$, whichever point $z$ is held out. Whether an action passes is thus decided by the state alone, and the data cannot be used to choose among the e-values. Formally, let

$$\lambda^*(S)=\min\left\{\lambda\in\Lambda:\sum_{j=1}^{n+1}\ell(z_j,\lambda)\le(n+1)\alpha\right\}$$

be the smallest action that passes on $S$. Since $\lambda^*(S)$ is permutation invariant in $S$, the values $E_S^*(D)=E_S^{\lambda^*(S)}(D)$ form a single e-collection. If $\lambda$ satisfies (B.1) at $D$ and $\ell(z,\lambda)>0$, then $\lambda\ge\lambda^*(S)$ for $S=D\cup\{z\}$, and hence

$$\ell(z,\lambda)\le\ell(z,\lambda^*(S))\le\alpha E_S^*(D),$$

where the second inequality uses $\sum_j\ell(z_j,\lambda^*(S))\le(n+1)\alpha$. Every action that satisfies (B.1) therefore lies in the certified menu of $E^*$, and Theorem 2.1 gives the validity of CRC and envelope CRC.

The e-collection $E^*$ is obtained by first choosing an action through risk control on the augmented sample and then evaluating conformal e-values at that action. The e-values of Bai and Jin [2026] for selective prediction are built in the same way. There, a selection threshold is chosen by controlling the risk on the augmented sample, and the conformal e-value is evaluated at that threshold, with the unknown test loss handled by an infimum over its possible values, much as (B.1) is imposed for every possible test point.

**Exchangeability and i.i.d. sampling.** Finally, we explain why Lemma 3.1 requires validity under every finite exchangeable law. Validity under i.i.d. sampling extends to mixtures of i.i.d. laws, but finite exchangeable laws need not be such mixtures, and the two requirements already differ for $n=1$. Let $Z_1,Z_2\in\{0,1\}$, take the miscoverage loss with $\alpha=1/2$, and consider the controller that predicts the set $\{Z_1\}$ for $Z_2$. If $Z_1$ and $Z_2$ are i.i.d. Bernoulli$(p)$, it miscovers with probability $2p(1-p)\le1/2$, so it is valid under i.i.d. sampling. Under the exchangeable law that is uniform on $(0,1)$ and $(1,0)$, which is not a mixture of i.i.d. laws, it miscovers with probability one. In terms of Lemma 3.1, both held-out losses on the augmented sample $\{0,1\}$ equal one, so their total exceeds the budget $(n+1)\alpha=1$ and the unused budget $\Delta_A(\{0,1\})=-1$ is negative. Recertification spends $\Delta_A$ and therefore needs validity under exchangeability. CRC and envelope CRC have it, and so satisfy (3.5).

## B.2 Other allocations

Equal sharing is one way, but not the only way, to distribute the unused budget among the held-out positions. For each augmented multiset $S$ and risk $t$, let $w_{i,t}(S)\ge0$ be measurable weights with $\sum_i w_{i,t}(S)=1$ that are permutation equivariant: reordering the points of $S$ reorders the weights, and identical points receive identical weights. These weights give the budgets

$$b_{i,t}^A(S)=\ell_t(z_i,A(S_{-i}))+\Delta_{A,t}(S)\,w_{i,t}(S),\tag{B.2}$$

which sum to $(n+1)\alpha_t$ on every augmented sample. Write $b_t^A(S;z)$ for the common budget of the copies of $z$ in $S$. By Lemma 3.1, every measurable symmetric controller $C$ with

$$\ell_t(z,C(D))\le b_t^A(D\cup\{z\};z)\qquad\text{for all } D,\ z,\ t\tag{B.3}$$

is jointly valid, since its held-out losses on each $S$ sum to at most $(n+1)\alpha_t$. Adding the constraint $C \le A$ makes the improvement pointwise. Equal weights recover (3.12), and weights proportional to the held-out losses of $A$ are another choice. Restricting to allocations loses no generality, since every pointwise improvement satisfies (B.3) for some allocation.

**Proposition B.1** (Every pointwise improvement has an allocation)**.** *Suppose that every $\ell_t(z,\cdot)$ is non-increasing. If $A$ and $C$ are jointly valid symmetric controllers with $C \le A$, then $C$ satisfies (B.3) for some permutation-equivariant weights.*

*Proof.* Fix $S$ and $t$, and write $r^A_{i,t}$ and $r^C_{i,t}$ for the held-out losses of the two controllers. Since $C \le A$ and the losses are non-increasing, the extra losses $d_{i,t} = r^C_{i,t} - r^A_{i,t}$ are nonnegative, and they sum to $\Delta_{A,t} - \Delta_{C,t}$. If $\Delta_{A,t} > 0$, set

$$w_{i,t}(S) = \frac{d_{i,t} + \Delta_{C,t}/(n+1)}{\Delta_{A,t}}.$$

These weights are nonnegative and sum to one. They give each position its extra loss together with an equal share of the unused budget of $C$, so the budgets in (B.2) equal $r^C_{i,t} + \Delta_{C,t}/(n+1) \ge r^C_{i,t}$. If $\Delta_{A,t} = 0$, then $\sum_i d_{i,t} = -\Delta_{C,t} \le 0$ forces every $d_{i,t}$ to vanish, and any equivariant weights suffice. The weights are equivariant and measurable because both controllers are symmetric and measurable. □

By Proposition B.1 and Theorem 3.5, repeated equal sharing can also be interpreted as an implicit search over allocations. Its admissible limit $A^{(\infty)}$ satisfies (B.3) for some allocation of the unused budget of $A^{(0)}$, and no allocation yields a valid controller that improves on it.

## B.3 Proof of Theorem 3.5 and Proposition 3.6

We prove the second part of Proposition 3.6, of which Theorem 3.5 is the case of a single risk. Throughout, $\Lambda = [\lambda_-, \lambda_+]$, and every $\ell_t(z,\cdot)$ is non-increasing, right-continuous and convex. By Lemma 3.1, a controller $A$ is jointly valid if and only if $T_{A,t}(S) \le (n+1)\alpha_t$ for every $S$ and $t$.

**Lemma B.2** (One-step comparison)**.** *If $A$ and $C$ are jointly valid and $C \le A$, then $\mathcal{R}A \le \rho A + C/(n+1)$.*

*Proof.* Fix $D$, $z$ and $t$, and let $S = D \cup \{z\}$. Since $C \le A$ and the losses are non-increasing, each held-out loss of $C$ on $S$ is at least that of $A$, and by joint validity of $C$ these differences sum to $T_{C,t}(S) - T_{A,t}(S) \le \Delta_{A,t}(S)$. Each difference is therefore at most $\Delta_{A,t}(S)$, and for the held-out point $z$,

$$\ell_t(z, C(D)) - \ell_t(z, A(D)) \le \Delta_{A,t}(S).$$

The action $a = \rho A(D) + C(D)/(n+1)$ lies between $C(D)$ and $A(D)$, and convexity gives

$$\ell_t(z,a) - \ell_t(z,A(D)) \le \frac{\ell_t(z,C(D)) - \ell_t(z,A(D))}{n+1} \le \frac{\Delta_{A,t}(S)}{n+1}.$$

Since $z$ and $t$ are arbitrary, $a$ is feasible in (3.12), and hence $(\mathcal{R}A)(D) \le a$. □

*Proof of Proposition 3.6.* By the first part of the proposition, the iterates are jointly valid and $A^{(k+1)} \le A^{(k)}$. Hence $A^{(k)}(D)$ decreases to a limit $A^{(\infty)}(D) \in \Lambda$ for every $D$, and $A^{(\infty)}$ is measurable and symmetric as a pointwise limit. For fixed $S$ and $t$, the losses are non-increasing and right-continuous, so $\ell_t(z_i, A^{(k)}(S_{-i})) \uparrow \ell_t(z_i, A^{(\infty)}(S_{-i}))$, and $T_{A^{(\infty)},t}(S) = \lim_k T_{A^{(k)},t}(S) \le (n+1)\alpha_t$. Thus $A^{(\infty)}$ is jointly valid.

Let $C$ be jointly valid with $C \le A^{(\infty)}$. Then $C \le A^{(k)}$ for every $k$, and Lemma B.2 gives $A^{(k+1)} \le \rho A^{(k)} + C/(n+1)$. Letting $k \to \infty$ yields $A^{(\infty)} \le \rho A^{(\infty)} + C/(n+1)$, that is, $A^{(\infty)} \le C$. Hence $C = A^{(\infty)}$, and no jointly valid controller strictly dominates $A^{(\infty)}$. Taking $C = A^{(\infty)}$ in Lemma B.2 gives $0 \le A^{(k+1)} - A^{(\infty)} \le \rho\,(A^{(k)} - A^{(\infty)})$, and induction gives (3.9).

Finally, if $A$ is admissible, then $\mathcal{R}A$ is jointly valid with $\mathcal{R}A \le A$, so $\mathcal{R}A = A$ and $A$ is a fixed point. Conversely, if $\mathcal{R}A = A$ and $C \le A$ is jointly valid, Lemma B.2 gives $A \le \rho A + C/(n+1)$, so $A \le C$ and $C = A$. Hence $A$ is admissible. □

## B.4 Proof of Proposition 3.7

For a calibration multiset $D_0$ and an action $a \in \Lambda$, let $A^{D_0,a}$ denote the controller that agrees with $A$ except at $D_0$, where it takes the action $a$. It is symmetric, and measurable since it differs from $A$ at a single multiset.

*Proof of Proposition 3.7.* Consider changing the controller $A$ to $A^{D_0,a}$. The held-out position $i$ of an augmented multiset $S$ is affected by the change if and only if $S_{-i} = D_0$, that is, if and only if $S = D_0 \cup \{z\}$ and $z_i = z$ for some $z$. Such an $S$ contains $\nu_{D_0}(z) + 1$ copies of $z$, so

$$T_{A^{D_0,a},t}(D_0 \cup \{z\}) = T_{A,t}(D_0 \cup \{z\}) + (\nu_{D_0}(z) + 1)\{\ell_t(z, a) - \ell_t(z, A(D_0))\},$$

and every other total is unchanged. By Lemma 3.1, $A^{D_0,a}$ is jointly valid if and only if (3.13) holds at $D_0$ and $a$. If (3.13) holds at some $D_0$ and $a < A(D_0)$, then $A^{D_0,a}$ is jointly valid and strictly dominates $A$. Conversely, let $C \le A$ be a jointly valid symmetric controller with $C(D_0) < A(D_0)$ for some $D_0$, and put $a = C(D_0)$. Since $A^{D_0,a} \ge C$ pointwise and the losses are non-increasing, each held-out loss of $A^{D_0,a}$ is at most that of $C$. Hence $A^{D_0,a}$ is jointly valid, and (3.13) holds at $D_0$ and $a < A(D_0)$.

Since $\mathcal{Z}$ is finite in this case, there are finitely many calibration multisets. Their order is fixed in advance, so that the sweep defines a single controller rather than a data-dependent procedure. At each visit, the actions satisfying (3.13) form a closed upper interval of $\Lambda$ that contains the current entry, by the monotonicity and right continuity of the losses, so the smallest such action $a^*$ exists, and the update preserves joint validity. Two observations show that a visited entry cannot be lowered afterwards. First, right after the update, an action $a < a^*$ satisfies (3.13) at $D_0$ if and only if it did before the update, since the update lowers both sides of (3.13) by the same amount $(\nu_{D_0}(z) + 1)\{\ell_t(z, a^*) - \ell_t(z, A(D_0))\}$. By the choice of $a^*$, no such action exists. Second, a later update at another multiset can only increase held-out losses, so it shrinks every unused budget, and hence the right side of (3.13) at $D_0$, while the entry at $D_0$, and hence the left side, stays fixed. After one pass, the condition therefore fails at every multiset for every smaller action. The resulting controller is jointly valid and no larger than $A$, and it is admissible by the criterion. □

## B.5 The two-type example

In this subsection, we carefully solve the two-type example in Section 3.4.

We first write out the validity constraints, which describe the valid controllers as the points of $[0,1]^3$ that satisfy four linear inequalities. With the losses (3.11), $\ell(z,\lambda) = g(z)(1-\lambda)$ with $g(H) = 1$ and $g(L) = 1/2$, and the budget is $(n+1)\alpha = 3/2$. A symmetric controller $A$ is determined by its actions at $HH$, $HL$ and $LL$. We write

$$(h, u, v) = \big(1 - A(HH),\ 1 - A(HL),\ 1 - A(LL)\big) \in [0,1]^3,$$

so that larger coordinates correspond to less conservative actions. In $HHL$, for instance, deleting either copy of $H$ leaves $HL$ and deleting $L$ leaves $HH$, so the held-out losses are $u$, $u$ and $h/2$. Treating the four augmented multisets in this way, Lemma 3.1 shows that $A$ is valid if and only if

$$3h \le \tfrac{3}{2}\ \ (HHH), \qquad 2u + \tfrac{1}{2}h \le \tfrac{3}{2}\ \ (HHL), \qquad u + v \le \tfrac{3}{2}\ \ (HLL), \qquad \tfrac{3}{2}v \le \tfrac{3}{2}\ \ (LLL). \tag{B.4}$$

Let $G_D = \sum_{z \in D} g(z)$. For the baselines, we verify that standard CRC with $B = 1$ gives $1 - A_{\mathrm{CRC}}(D) = 1/(2G_D)$, and envelope CRC with the sharp envelope $U(\lambda) = 1 - \lambda$ gives $1 - A_{\mathrm{env}}(D) = 3/(2(G_D + 1))$. In particular, envelope CRC has $(h, u, v) = (1/2, 3/5, 3/4)$.

Starting from the controller of envelope CRC, the update (3.7) at $D$ replaces the current coordinate $w_D$ by the largest $w \le 1$ with $g(z)(w - w_D) \le \Delta_A(D \cup \{z\})/3$ for $z \in \{H, L\}$. The three coordinates evolve as follows:

- At $HH$, the candidate $H$ gives $w \le h$, since $\Delta_A(HHH) = 3/2 - 3h = 0$ at $h = 1/2$. Hence $h = 1/2$ throughout.
- At $HL$, with $h = 1/2$, the candidates $H$ and $L$ give $w \le u/3 + 5/12$ and $w \le u/3 + 1 - 2v/3$. The first bound is the smaller one whenever $v \le 7/8$.

- At $LL$, the candidates $H$ and $L$ give $w \le 2v/3 - u/3 + 1/2$ and $w \le 1$.

Hence, as long as $v_k \le 7/8$ and $2v_k/3 - u_k/3 + 1/2 \le 1$, the iterates satisfy

$$u_{k+1} = \tfrac{1}{3}u_k + \tfrac{5}{12}, \qquad v_{k+1} = \tfrac{2}{3}v_k - \tfrac{1}{3}u_k + \tfrac{1}{2},$$

with $u_0 = 3/5$ and $v_0 = 3/4$. By induction, the solutions are

$$u_k = \frac{5}{8} - \frac{1}{40}\left(\frac{1}{3}\right)^k, \qquad v_k = \frac{7}{8} - \frac{1}{10}\left(\frac{2}{3}\right)^k - \frac{1}{40}\left(\frac{1}{3}\right)^k.$$

They satisfy $v_k < 7/8$ for every $k$, so both conditions hold and the recurrences are the exact updates. In terms of actions,

$$A^{(k)}(HH) = \frac{1}{2}, \qquad A^{(k)}(HL) = \frac{3}{8} + \frac{1}{40}\left(\frac{1}{3}\right)^k, \qquad A^{(k)}(LL) = \frac{1}{8} + \frac{1}{10}\left(\frac{2}{3}\right)^k + \frac{1}{40}\left(\frac{1}{3}\right)^k,$$

which give the entries of Table 1 and converge geometrically at rate $\rho = 2/3$ to $A^{(\infty)} = (1/2, 3/8, 1/8)$.

In this case, a valid controller is admissible if and only if no other point satisfying (B.4) is coordinatewise at least as large. We observe the following.

- The limit has $(h, u, v) = (1/2, 5/8, 7/8)$. The constraints for $HHH$, $HHL$ and $HLL$ bind, and the one for $LLL$ is slack, since $\frac{3}{2} \cdot \frac{7}{8} = \frac{21}{16} < \frac{3}{2}$. If a feasible point $(h', u', v')$ is at least as large, then $HHH$ forces $h' = 1/2$, $HHL$ then forces $u' = 5/8$, and $HLL$ then forces $v' = 7/8$. Hence the limit is admissible.
- The controller $(1/2, 2/5, 1/10)$ has $(h, u, v) = (1/2, 3/5, 9/10)$ and satisfies (B.4), with $HHH$ and $HLL$ binding. A feasible point at least as large has $h' = 1/2$ by $HHH$, and $u' \ge 3/5$, $v' \ge 9/10$ and $u' + v' \le 3/2$ force $u' = 3/5$ and $v' = 9/10$. Hence this controller is also admissible.
- The two controllers are incomparable, and their pointwise minimum has $(h, u, v) = (1/2, 5/8, 9/10)$, which violates the constraint for $HLL$, since $5/8 + 9/10 = 61/40 > 3/2$.

Theorem 3.5 requires $\Lambda$ to be an interval, even for affine losses. On the grid $\{1/10, 1/4, 2/5, 1/2, 1\}$, the envelope controller $(1/2, 2/5, 1/4)$ is a fixed point of the restricted update, since its feasible sets in (3.7) at $HH$, $HL$ and $LL$, namely $[1/2, 1/2]$, $[23/60, 2/5]$ and $[1/5, 1/4]$, contain no smaller grid value. Yet the valid grid controller $(1/2, 2/5, 1/10)$ strictly dominates it. The step that fails is the interpolation in Lemma B.2: at $LL$, the action $\rho A(D) + C(D)/(n+1) = 1/5$ is feasible but not on the grid. Nevertheless, the sweep of Proposition 3.7 needs no interpolation and reaches $(1/2, 2/5, 1/10)$ in one pass.

## B.6 A two-risk example with a continuous observation space

We give one more example, with two risks and a continuous observation space, so that a controller is a function rather than a finite table. Recertification converges to the largest valid controller that dominates the joint envelope baseline, and the improvement is strict at every interior point.

Take $n = 1$ for simplicity, so that a single calibration observation $X \in [0, 1]$ is used to choose an action for a test observation $Y \in [0, 1]$, and let $\Lambda = [0, 1]$. Throughout this example we describe actions by $a = 1 - \lambda$, so that larger values are less conservative, and write $A(x) \in [0, 1]$ for the value chosen at the calibration observation $x$. The losses $ya$ and $(1 - y)a$ must each have risk at most $1/4$,

$$\mathbb{E}[YA(X)] \le \tfrac{1}{4}, \qquad \mathbb{E}[(1 - Y)A(X)] \le \tfrac{1}{4},$$

under every exchangeable law of $(X, Y)$. Both losses are affine in $\lambda$, with sharp envelope $a$. By Lemma 3.1, applied to each risk, $A$ is jointly valid if and only if

$$yA(x) + xA(y) \le \tfrac{1}{2} \quad \text{and} \quad (1 - y)A(x) + (1 - x)A(y) \le \tfrac{1}{2} \qquad \text{for all } x, y \in [0, 1]. \tag{B.5}$$

Envelope CRC for the two risks gives $A(x) \le 1/(2(1+x))$ and $A(x) \le 1/(2(2-x))$, and the joint envelope baseline $A_0$ takes the more conservative of the two. We compare it with $A_*$, where

$$A_0(x) = \frac{1}{2(1+\max(x,1-x))}, \qquad A_*(x) = \frac{1+\min(x,1-x)}{4}.$$

If a jointly valid $A$ dominates $A_0$, then (B.5) at $x = y = 1$ and at $x = y = 0$ forces $A(1) = A(0) = 1/4$, the values of $A_0$ there, and the constraints with $y = 1$ and $y = 0$ then give $A(x) \le \min\{(2-x)/4, (1+x)/4\} = A_*(x)$. Conversely, these two bounds give

$$yA_*(x) + xA_*(y) \le \frac{1-(1-x)(1-y)}{2} \le \frac{1}{2}, \qquad (1-y)A_*(x) + (1-x)A_*(y) \le \frac{1-xy}{2} \le \frac{1}{2},$$

so $A_*$ is jointly valid. Hence $A_*$ is the largest valid controller that dominates $A_0$, and in particular it is admissible. The improvement is strict at every interior point.

In terms of $a$, the update (3.12) at $x$ replaces $A(x)$ by the smallest of the thresholds

$$\frac{A(x)}{2} + \frac{1/2 - xA(y)}{2y} \quad (0 < y \le 1), \qquad \frac{A(x)}{2} + \frac{1/2 - (1-x)A(y)}{2(1-y)} \quad (0 \le y < 1),$$

which come from the first and the second risk. Suppose that $A \le A_*$ and $A(0) = A(1) = 1/4$, as for $A_0$. Then the tightest constraints come from the extreme test points, $y = 1$ for the first risk and $y = 0$ for the second. Indeed, for $0 < y \le 1$ and $c \in [0,1]$,

$$\frac{1}{2} - cA(y) - y\left(\frac{1}{2} - \frac{c}{4}\right) = \frac{1-y}{2} - c\left(A(y) - \frac{y}{4}\right) \ge \frac{1-y}{2} - c\left(A_*(y) - \frac{y}{4}\right) \ge 0,$$

since $A_*(y) - y/4 = \min\{1/4, (1-y)/2\} \le (1-y)/2$. With $c = x$, the first family of thresholds is therefore minimized at $y = 1$, with value $A(x)/2 + (2-x)/8$. Similarly, the second family is minimized at $y = 0$, with value $A(x)/2 + (1+x)/8$. The smaller of the two values is $\{A(x) + A_*(x)\}/2$, so

$$A^{(k+1)} = \tfrac{1}{2}\big(A^{(k)} + A_*\big).$$

The updated controller again satisfies $A \le A_*$ and $A(0) = A(1) = 1/4$, and by induction

$$A^{(k)}(x) = A_*(x) - 2^{-k}\{A_*(x) - A_0(x)\}.$$

The iterates thus increase to $A_*$, and the rate (3.9) of Proposition 3.6, with $\rho = 1/2$, holds with equality. Each iterate is a convex combination of $A_0$ and $A_*$, so its validity can also be read off from the linear constraints (B.5).

## B.7 Implementation details for the real-data study

All computations use the cached classifier outputs, and no model is fitted on the calibration or test images. Throughout, $A_k(D, x)$ denotes the depth assigned to a test image $x$ after $k$ rounds of recertification, starting from envelope CRC $A_0$; larger depths are less conservative.

**Error types.** For an observation $z = (x, y)$, let its type $c(z)$ be 0 if $y = f(x)$, 1 if $y \in g(x) \setminus \{f(x)\}$, and 2 if $y \notin g(x)$. By (3.14), the loss depends on $z$ only through its type,

$$\ell_0(a) = 0, \qquad \ell_1(a) = \tfrac{1}{2}(a-1)_+, \qquad \ell_2(a) = \tfrac{1}{2}a, \qquad a \in [0,2].$$

The pool contains 7,300, 969 and 1,731 images of the three types.

**Baselines.**

- *Common depth.* With $h_j$ calibration observations of type $j$, CRC and envelope CRC select the largest $a \in [0, 2]$ with

$$\frac{h_1(a-1)_+ + h_2 a}{2} + U(a) \le (n+1)\alpha, \qquad U(a) = 1 \text{ (CRC)}, \quad U(a) = a/2 \text{ (envelope CRC)},$$

  and the root, $a = 0$, if no depth qualifies. The left side is piecewise linear with a kink at $a = 1$, so the largest depth is found in closed form on $[0, 1]$ and $[1, 2]$. It is generally fractional, so these baselines already randomize between adjacent levels.

- *Confidence-adaptive scores.* Write $u_i = 1 - p_g(X_i)$ and $v_i = 1 - p_f(X_i)$, so that $u_i \le v_i$, and, under the threshold $t$, image $i$ moves from the root to its superclass when $t > u_i$ and to its leaf when $t > v_i$. The calibration loss at $t$ is

$$\frac{1}{2}\sum_{i=1}^{n} \left[\mathbb{1}\{c(Z_i) = 2\}\,\mathbb{1}\{u_i < t\} + \mathbb{1}\{c(Z_i) \ne 0\}\,\mathbb{1}\{v_i < t\}\right],$$

  a step function that increases by $1/2$ at the leaf breakpoint of every observation of nonzero type and at the superclass breakpoint of every observation of type 2. CRC selects the largest $t$ at which at most $b_0 = 2((n+1)\alpha - 1)$ of these jumps have occurred. These baselines return integer depths.

**The update.** Fix a calibration sample $D$ and a test image $x$; in the common-depth family, the depth does not depend on $x$. For an augmented sample $S = (z_1, \ldots, z_{n+1})$, let $T_k(S) = \sum_{i=1}^{n+1} \ell(z_i, A_k(S_{-i}, x_i))$ be the total held-out loss of $A_k$. Since the loss depends on the label only through its type, all labels of the same type give the same constraint in the update of Section 3.5. For $c \in \{1, 2\}$, let $S_c$ augment $D$ with $(x, y_c)$, where $y_c$ is any label of type $c$ for $x$, and let

$$s_{k,c}(D, x) = \alpha - \frac{T_k(S_c)}{n+1}$$

be the equal share of its unused budget. With $a = A_k(D, x)$, the next depth is the largest $a' \in [a, 2]$ with $\ell_c(a') \le \ell_c(a) + s_{k,c}(D, x)$ for $c = 1, 2$, that is,

$$A_{k+1}(D, x) = \min\left\{2,\ a + 2s_{k,2}(D, x),\ 1 + (a-1)_+ + 2s_{k,1}(D, x)\right\}. \tag{B.6}$$

For common depth, every iterate depends on $D$ only through the count vector $h = (h_0, h_1, h_2)$, so $A_k$ is a table with $\binom{n+2}{2} = 1{,}326$ entries at $n = 50$. For an augmented count vector $q$ with total $n + 1$,

$$T_k(q) = \sum_{j:\, q_j > 0} q_j\, \ell_j\big(A_k(q - e_j)\big),$$

where $e_j$ is the $j$th unit vector, and the slacks for the entry $h$ come from $q = h + e_1$ and $q = h + e_2$. Each round updates the whole table at once. The table does not depend on the data, so it is computed once for each target, and each split reads off the entry for its calibration counts.

For confidence-adaptive scores, there is no common table, and for each test image we compute a table of all the depths that the recursion requires. Unrolling (B.6), $A_k(D, x)$ depends on depths $A_{k-1}(S_{-i}, x_i)$, in which a calibration point is held out and the test image carries a hypothetical type, and these depend in turn on further depths of the same form. Every depth that arises is therefore indexed by a held-out point, among the calibration points and the test image, and a type for each remaining point. Calibration points of type 0 never need to be held out, since their loss is zero, and they keep their type; they enter only through the sample size. If $r - 1$ calibration points have nonzero types, call them and the test image the $r$ *active* points. The table then holds one depth for each held-out active point and each assignment of type 1 or 2 to the other active points, giving $r2^{r-1}$ entries in total. Each round updates all entries from the preceding table, in time linear in its size, and the reported depth is the entry in which the test image is held out and the calibration points carry their observed types. Consequently, the computation can be more demanding than in the common-depth case. Two exact reductions shrink the table without changing any depth.

- *Capping breakpoints.* Let $b = \lfloor b_0 \rfloor$ be the number of whole jumps within the budget. Every active point has a leaf breakpoint, whatever its type. If there are at least $b+2$ active points, let $\tau$ be the $(b+2)$th smallest of their leaf breakpoints. After any one point is held out, at least $b+1$ jumps at or below $\tau$ remain, so every CRC threshold in the table is at most $\tau$. Replacing each breakpoint above $\tau$ by $\tau$ therefore leaves all baseline depths unchanged, and hence all later ones.
- *Grouping.* Active points with identical breakpoint pairs, after capping, are interchangeable, so a group of $g$ such points is described by its number of type-2 points rather than by $2^g$ assignments. With group sizes $g_1, \dots, g_J$, the table has $\sum_j g_j \prod_{h \neq j}(g_h + 1)$ entries.

A larger nominal level allows more jumps, which raises $\tau$ and leaves more distinct breakpoints. This explains the longer runtimes at larger levels.

We record rounds $0, 1, 3, 10, 20, 50, 100, 200$ and $500$, and the common-depth table is updated for all 500 rounds. For an adaptive test image, the total slack $(n+1)\alpha - T_k(S_c)$ bounds how much its loss can grow in all later rounds. Once this slack allows a depth increase of at most $10^{-8}$, we stop updating the image and carry its depth forward, which preserves validity by the first observation at the end of Section 3.5.

**Evaluation.** Each of the 300 splits draws 50 calibration images and 16 test images at random without replacement from the pool of 10,000 images, and all methods, families and targets use the same splits. All metrics are computed exactly over the randomization of the prediction. With $\pi(a) = ((1-a)_+, 1-|a-1|, (a-1)_+)$ the probabilities of the root, superclass and leaf, a test image of type $c$ predicted at depth $a$ has risk $\ell_c(a)$, expected node size $100\pi_0(a) + 5\pi_1(a) + \pi_2(a)$ and expected correct specificity

$$\frac{\log 20}{\log 100}\,\pi_1(a)\,\mathbb{1}\{c \neq 2\} + \pi_2(a)\,\mathbb{1}\{c = 0\}.$$

Each metric is averaged over the 16 test images of a split and then over the splits, and standard errors are computed across splits. In Figure 2, let $d_s$ be the gain over envelope CRC in split $s$, and let $\bar{d}$ and se be the mean and standard error of $d_1, \dots, d_M$ with $M = 300$. The bands are $\bar{d} \pm t_{M-1,0.975}\,\mathrm{se}$, pointwise intervals that reflect variation across splits of the fixed pool.

# C Details for randomized e-BH

Throughout, $e$ is a fixed nonnegative vector, $b_{[r]} = (\alpha/K)e_{[r]}$ is the budget of the hypothesis of rank $r$, and $c_{[r]}(p) = \sum_{k \geq r} p_k/k$ is its charge under a distribution $p$ of the number of rejections, as in Section 4.2. We also write $a_r = r\, b_{[r]} + \sum_{j>r} b_{[j]}$, so that $q_r^* = \min\{1, a_r\}$.

## C.1 The certified menu and U-eBH

We first check that e-BH is the largest certified set. By Section 4.1, the sets certified by $E^{\mathrm{eBH}}$ are the self-consistent ones. Let $k^*$ be the e-BH count, the largest $k$ with $k\, e_{[k]} \geq K/\alpha$. The e-BH set is self-consistent, because each of its members has an e-value of at least $e_{[k^*]} \geq K/(\alpha k^*)$. Moreover, recall that e-BH rejects every hypothesis $j$ with $e_j \geq K/(\alpha k^*)$. If $R$ is self-consistent and $|R| = k$, then its $k$ members have e-values of at least $K/(\alpha k)$, so $k\, e_{[k]} \geq K/\alpha$ and $k \leq k^*$. Every member of $R$ then has $e_j \geq K/(\alpha k) \geq K/(\alpha k^*)$ and is rejected by e-BH.

We also verify explicitly that every boosting rule that selects from the certified menu (4.4) controls the FDR. If $p(e)$ is measurable in $e$ and lies in the menu for every $e$, then for $P \in \mathcal{P}_T$, since the number of rejections is drawn independently of the data,

$$\mathbb{E}\ \mathrm{FDP}_T(R_\kappa(E)) = \mathbb{E}_P \bar{L}(T, p(E); E) = \mathbb{E}_P \sum_{j \in T} c_{[r(j)]}(p(E)) \leq \frac{\alpha}{K} \sum_{j \in T} \mathbb{E}_P E_j \leq \frac{\alpha |T|}{K}.$$

This is Theorem 2.1 for the action space of distributions.

As described in the main text, the rule given by U-eBH leaves a nonnegative unused budget. Fix $r$, and let $k_r \geq r$ maximize $k\, e_{[k]}$ over $k \geq r$. By (4.3), the tails of $p^0$ are equal at all ranks $r' \in \{r, \dots, k_r\}$, since the maximum of $k\, e_{[k]}$ over $k \geq r'$ is still attained at $k_r$. Since $\{\kappa_0 \geq k\} = \{U \leq \mathbb{P}(\kappa_0 \geq k \mid E = e)\}$, the events $\{\kappa_0 \geq r\}$ and $\{\kappa_0 \geq k_r\}$ coincide. Whenever U-eBH rejects the hypothesis of rank $r$, it rejects at least $k_r$ hypotheses, so that hypothesis receives a share of at most $1/k_r$. Hence

$$c_{[r]}(p^0) = \mathbb{E}\left[\frac{\mathbb{1}\{\kappa_0 \geq r\}}{\kappa_0} \,\middle|\, E = e\right] \leq \frac{\mathbb{P}(\kappa_0 \geq r \mid E = e)}{k_r} \leq \frac{\alpha}{K}\, e_{[k_r]} \leq b_{[r]}.$$

U-eBH therefore selects from the menu, and the difference $b_{[r]} - c_{[r]}(p^0)$ is the budget that it leaves unused for the hypothesis of rank $r$, which is positive whenever $e_{[r]} > e_{[k_r]}$.

Recertifying U-eBH with $E^{\mathrm{eBH}}$ is exactly an instance of Proposition 2.2. The state is the set $T$ of true nulls, and the loss is the averaged false discovery proportion. Since U-eBH selects the single distribution $p^0(e)$, its canonical envelope is its own loss, $g_T = \bar{L}(T, p^0(E); E) = \sum_{j \in T} c_{[r(j)]}(p^0(E))$. Take the conditioning statistic $V_T = E$. Given $E$, the loss $g_T$ is known, because the law of $U$ is known, so we may take $H_T = g_T$. Take the budget $B_T = \alpha E_T^{\mathrm{eBH}} = (\alpha/K) \sum_{j \in T} E_j$. Summing the previous paragraph over $j \in T$ gives $H_T \leq B_T$, and $\mathbb{E}_P B_T \leq \alpha |T|/K \leq \alpha$ for $P \in \mathcal{P}_T$, so (2.8) holds. The allowance $B_T - H_T$ is the unused budget of the previous paragraph, and it is a function of $E$. The increment $W_T = B_T - H_T$ therefore satisfies (2.9) with equality, and the recertified e-collection $(g_T + W_T)/\alpha = B_T/\alpha$ is $E_T^{\mathrm{eBH}}$. Its certified menu is (4.4), which contains $p^0$ and every other distribution whose charges stay within the budgets.

## C.2 Proof of Proposition 4.1

Section 4.3 shows that every distribution in the menu (4.4) satisfies the tail bounds (4.5). The following lemma shows that a single distribution in the menu attains all of them. Put $e_{[K+1]} = 0$ and $a_{K+1} = 0$ for notational convenience.

**Lemma C.1.** *Let $\widetilde{p}_k = (\alpha k/K)(e_{[k]} - e_{[k+1]})$ for $k = 1, \dots, K$, define $p_k^* = \min\{\widetilde{p}_k, 1 - \sum_{l>k} p_l^*\}$ for $k = K, K-1, \dots, 1$, and let $p_0^* = 1 - \sum_{k \geq 1} p_k^*$. Then $p^*$ is a distribution in the menu (4.4) with tails $\sum_{k \geq r} p_k^* = q_r^*$ for $r = 1, \dots, K$. If $a_1 \leq 1$, then $p_k^* = \widetilde{p}_k$ for $k \geq 1$, and every constraint in (4.4) holds with equality.*

*Proof.* We check that $\widetilde{p}_k$ is nonnegative, $\sum_{k \geq r} \widetilde{p}_k/k = (\alpha/K) \sum_{k \geq r} (e_{[k]} - e_{[k+1]}) = b_{[r]}$, and $\widetilde{p}_r = a_r - a_{r+1}$. By construction, $0 \leq p_k^* \leq \widetilde{p}_k$ for $k \geq 1$ and $\sum_{k \geq 1} p_k^* \leq 1$, so $p^*$ is a distribution whose charges are at most the budgets $b_{[r]}$, and it lies in the menu. Downward induction in $r$ gives

$$\sum_{k \geq r} p_k^* = \min\Big\{1, \sum_{k > r} p_k^* + \widetilde{p}_r\Big\} = \min\{1, \min\{1, a_{r+1}\} + \widetilde{p}_r\} = \min\{1, a_r\} = q_r^*.$$

If $a_1 \leq 1$, then $\widetilde{p}_k \leq 1 - a_{k+1}$ for every $k$, so no truncation occurs, $p_k^* = \widetilde{p}_k$ for $k \geq 1$, and every charge equals its budget. □

*Proof of Proposition 4.1.* By Lemma C.1, the bounds $q_r^*$ are the tails of the distribution $p^*$, so they are non-increasing in $r$. Hence $\{\kappa^* \geq r\} = \{U \leq q_r^*\}$, and $\kappa^*$ has the distribution $p^*$ given $E = e$.

(i) The tails $q_r^*(e)$ are measurable in $e$, so $p^*(e)$ is a measurable selection from the menu (4.4), and the validity of boosting rules in Appendix C.1 gives $\mathbb{E}\ \mathrm{FDP}_T(R_{\kappa^*}(E)) \leq \alpha |T|/K$.

(ii) If $k \geq r$, then

$$r\, e_{[r]} + \sum_{j > r} e_{[j]} \geq r\, e_{[k]} + \sum_{j=r+1}^{k} e_{[j]} \geq k\, e_{[k]},$$

so $q_r^*$ is at least the U-eBH tail (4.3). With the same $U$, it follows that $\{\kappa_0 \geq r\} \subseteq \{\kappa^* \geq r\}$ for every $r$, so $\kappa^* \geq \kappa_0$ and $R_{\kappa_0}(e) \subseteq R_{\kappa^*}(e)$. Clearly, U-eBH in turn contains e-BH.

(iii) Every boosting rule that selects from (4.4) has tails at most $q_r^*$ by (4.5), and $\kappa^*$ attains them. □

**Corollary C.2.** *Let $\widetilde{E}_j = (K/\alpha)\,\mathbb{1}\{j \in R_{\kappa^*}\}/(|R_{\kappa^*}| \vee 1)$. Then $\mathbb{E}[\widetilde{E}_j \mid E] \le E_j$, so $\widetilde{E}_j$ is an e-value whenever $E_j$ is, and e-BH applied to $\widetilde{E}$ rejects exactly $R_{\kappa^*}(E)$.*

*Proof.* Given $E = e$, $\mathbb{E}[\widetilde{E}_j \mid E = e] = (K/\alpha)\, c_{[r(j)]}(p^*) \le (K/\alpha)\, b_{[r(j)]} = e_j$. Every member of $R_{\kappa^*}$ has $\widetilde{E}_j = K/(\alpha|R_{\kappa^*}|)$ and every other hypothesis has $\widetilde{E}_j = 0$, so the largest $k$ with $k\,\widetilde{E}_{[k]} \ge K/\alpha$ is $|R_{\kappa^*}|$, and e-BH applied to $\widetilde{E}$ rejects exactly $R_{\kappa^*}$. □

## C.3 Randomization over arbitrary rejection sets

The main text considers randomized rules that reject one of the nested sets $R_k(e)$. More generally, a randomized rule may choose, given $E = e$, any distribution $Q$ over the subsets $R \subseteq \{1, \dots, K\}$. Each rejected null hypothesis contributes $1/|R|$ to the false discovery proportion, so in state $T$,

$$\mathbb{E}_Q\,\mathrm{FDP}_T(R) = \sum_{j \in T} c_j(Q), \qquad c_j(Q) = \mathbb{E}_Q\left[\frac{\mathbb{1}\{j \in R\}}{|R| \vee 1}\right],$$

where $c_j(Q)$ is the expected share charged to hypothesis $j$. As in Section 4.1, both sides of $\mathbb{E}_Q\,\mathrm{FDP}_T(R) \le \alpha E_T^{\mathrm{eBH}}(e)$ are sums over $T$, so $Q$ is certified by $E^{\mathrm{eBH}}$ if and only if $c_j(Q) \le b_j = (\alpha/K)e_j$ for every $j$.

J-eBH is such a rule whose rejection sets are not nested, and it is certified by $E^{\mathrm{eBH}}$, so that in particular it controls the FDR. To bound the charge of hypothesis $j$, fix $E = e$ and the uniforms $U_i$ for $i \ne j$, and let $f(u)$ be the number of rejections when $U_j = u$. Increasing $u$ decreases the boosted value $e_j/u$ and leaves the others unchanged, so $f$ is non-increasing. Since e-BH rejects every hypothesis whose boosted value is at least $K/(\alpha f(u))$, hypothesis $j$ is rejected exactly when $u \le (\alpha/K)e_j f(u)$. If this condition holds at $u$, it holds at every $u' < u$, because $u' < u \le (\alpha/K)e_j f(u) \le (\alpha/K)e_j f(u')$. The values of $u$ at which hypothesis $j$ is rejected therefore form an interval with left endpoint 0 and some right endpoint $u^*$. If $u^* > 0$, then $f(u) \ge f^-(u^*) \ge 1$ on this interval, where $f^-(u^*)$ is the left limit of $f$ at $u^*$, and letting $u \uparrow u^*$ in the rejection condition gives $u^* \le (\alpha/K)e_j f^-(u^*)$. Hence

$$\mathbb{E}\left[\frac{\mathbb{1}\{j \in R\}}{|R| \vee 1} \,\middle|\, U_i,\, i \ne j\right] = \int_0^{u^*} \frac{du}{f(u)} \le \frac{u^*}{f^-(u^*)} \le \frac{\alpha}{K} e_j = b_j,$$

which also holds trivially if $u^* = 0$, and averaging over the other uniforms gives $c_j(Q) \le b_j$.

**Proposition C.3.** *Every distribution $Q$ over rejection sets that is certified by $E^{\mathrm{eBH}}$ satisfies $\mathbb{E}_Q|R| \le \mathbb{E}[\kappa^* \mid E = e]$. The stochastic comparison of Proposition 4.1(iii) does not extend to this class: such a $Q$ may satisfy $Q(|R| \ge r) > q_r^*$ for some $r$.*

*Proof.* Let $m$ be the number of ranks $r$ with $a_r \ge 1$. Since $a_r$ is non-increasing in $r$, $q_r^* = 1$ for $r \le m$ and $q_r^* = a_r$ for $r > m$. Summing the tails,

$$\mathbb{E}[\kappa^* \mid E = e] = \sum_{r=1}^{K} q_r^* = m + \sum_{r > m} a_r = m + \sum_{j > m} (2j - m - 1)\, b_{[j]},$$

because for $j > m$, the budget $b_{[j]}$ enters $a_j$ with coefficient $j$ and $a_r$ with coefficient one for each $m < r < j$. Give the hypothesis $j$ the weight $y_j = 0$ if $r(j) \le m$ and $y_j = 2r(j) - m - 1$ otherwise, so that $\mathbb{E}[\kappa^* \mid E = e] = m + \sum_j y_j b_j$. The weights, ordered by rank, are $m$ zeros followed by $m+1, m+3, m+5, \dots$, so the weights in a set $R$ of size $k \ge 1$ sum to at least 0 if $k \le m$ and at least $\sum_{i=1}^{k-m}(m + 2i - 1) = k(k-m)$ otherwise. In both cases $\sum_{j \in R} y_j \ge k(k-m)$, which leads to

$$|R| \le m\ \mathbb{1}\{R \ne \emptyset\} + \sum_j y_j \frac{\mathbb{1}\{j \in R\}}{|R| \vee 1}.$$

This also holds for $R = \emptyset$. Taking expectations under $Q$ and using $c_j(Q) \le b_j$,

$$\mathbb{E}_Q|R| \le m\, Q(R \neq \emptyset) + \sum_j y_j\, c_j(Q) \le m + \sum_j y_j b_j = \mathbb{E}[\kappa^* \mid E = e].$$

The constant $m$ and the weights $y_j$ thus act as a dual certificate for the linear program that maximizes $\mathbb{E}_Q|R|$ subject to the budget constraints.

For the second claim, take $K = 5$, $\alpha = 0.05$ and $e = (100, 1, 1, 1, 1)$, so that $b_1 = 1$ and $b_j = 0.01$ for $j \ge 2$. Let $Q$ always reject $H_1$ and reject each other hypothesis independently with probability 0.02. Then $c_1(Q) \le 1 = b_1$, and for $j \ge 2$, every rejection of $H_j$ comes with that of $H_1$, so $c_j(Q) \le 0.02/2 = b_j$. Hence $Q$ is certified, yet $Q(|R| \ge 2) = 1 - 0.98^4 \approx 0.078$, whereas $q_2^* = 0.05$. Its expected number of rejections, 1.08, is below $\mathbb{E}[\kappa^* \mid E = e] = 1.2$, as the first claim requires. □

## C.4 Proof of Proposition 4.2

Throughout, $e$ is fixed, $S = S^\sharp$, and $c_r = S_r - S_{r+1}$ for $r = 1, \dots, K$, with $c_{K+1} = 0$. We show below that these are the charges of the rule. A state is identified with the set of ranks of its hypotheses, so that $\{s, \dots, K\}$ denotes the $K - s + 1$ hypotheses with the smallest e-values. Write $\Phi_s = \alpha \sum_{j \ge s} e_{[j]}$ for $s = 1, \dots, K+1$, so that $\alpha e_{[r]} = \Phi_r - \Phi_{r+1}$ and $\Phi_s = (K - s + 1)\alpha m_s$. The sequence $\Phi$ is convex, because its decrements $\alpha e_{[s]}$ are non-increasing. We call $s$ a vertex if $S_s = \beta_s$. The points 1 and $K+1$ are vertices, and $S$ is linear between consecutive vertices. Since $S \le \beta$ and $\beta_s \le \alpha m_s$, we have $(K - s + 1)S_s \le \Phi_s$ for every $s$, with equality at every vertex $s \ge 2$. We first prove three lemmas that are central to the proof of Proposition 4.2.

**Lemma C.4.** *The sequence $S$ is nonnegative, non-increasing and convex, and the charges $c_r$ are nonnegative and non-increasing in $r$. The numbers $p_k^\sharp = k(c_k - c_{k+1})$ for $k = 1, \dots, K$ and $p_0^\sharp = 1 - S_1$ form a distribution with charges $c_{[r]}(p^\sharp) = c_r$ and tails $q_r^\sharp$, and $\kappa^\sharp$ has the distribution $p^\sharp$ given $E = e$. Moreover, $q_r^\sharp$ is a continuous function of $e$.*

*Proof.* The zero sequence is a convex minorant of $\beta$, so $S \ge 0$, and a nonnegative convex sequence that vanishes at $K+1$ is non-increasing. Hence $c_r \ge 0$, and $c_r$ is non-increasing in $r$ by convexity, so $p_k^\sharp \ge 0$ for $k \ge 1$. Since $\sum_{k \ge 1} p_k^\sharp = \sum_{k \ge 1} c_k = S_1 \le \beta_1 \le 1$, also $p_0^\sharp \ge 0$, and the numbers sum to one. Telescoping gives $\sum_{k \ge r} p_k^\sharp / k = \sum_{k \ge r}(c_k - c_{k+1}) = c_r$, so $p^\sharp$ has the charges $c_r$, and by the accounting identity of Section 4.3 its tails are $r\, c_r + \sum_{j > r} c_j = rS_r - (r-1)S_{r+1} = q_r^\sharp$. Tails of a distribution are non-increasing, so $\{\kappa^\sharp \ge r\} = \{U \le q_r^\sharp\}$, and $\kappa^\sharp$ has the distribution $p^\sharp$. Finally, $\beta$ is continuous in $e$, and so is its greatest convex minorant. □

**Lemma C.5.** *If a distribution $p$ of the number of rejections satisfies $S_s(p) \le \alpha m_s$ for $s = 1, \dots, K$, then $S(p) \le S$. In particular, this holds for every monotone rule certified by $E^{\overline{\mathrm{eBH}}}$.*

*Proof.* Since also $S_1(p) = \sum_{k \ge 1} p_k \le 1$, we have $S(p) \le \beta$. The charges $c_{[r]}(p)$ are non-increasing in $r$, so $S(p)$ is convex with $S_{K+1}(p) = 0$, and it is therefore a convex minorant of $\beta$, which lies below the greatest one, $S$. The constraints $S_s(p) \le \alpha m_s$ are the constraints (4.7) for the states $\{s, \dots, K\}$. □

**Lemma C.6.** *For $1 \le r < r' \le K$,*

$$(K - r)(c_r - c_{r'}) \le \alpha(e_{[r]} - e_{[r']}). \tag{C.1}$$

*Proof.* Let $w$ be the smallest vertex greater than $r$, and write $\gamma = c_r$, the slope magnitude of the linear piece of $S$ that ends at $w$. If $r' < w$, then $c_{r'} = \gamma$, and (C.1) holds because $e_{[r]} \ge e_{[r']}$. Otherwise, let $v$ be the largest vertex with $v \le r'$, so that $2 \le w \le v$, $c_{r'} = c_v =: \gamma'$ and $e_{[r']} \le e_{[v]}$. We bound $e_{[r]}$ from below and

$e_{[v]}$ from above. Since $\Phi$ is convex and $r < w$, its decrement at $r$ is at least its average decrement over $[r, w]$. With $\Phi_r \geq (K-r+1)S_r$, $\Phi_w = (K-w+1)S_w$ and $S_r = S_w + (w-r)\gamma$, this gives

$$\alpha e_{[r]} \geq \frac{\Phi_r - \Phi_w}{w-r} \geq \frac{(K-r+1)S_r - (K-w+1)S_w}{w-r} = (K-r+1)\gamma + S_w.$$

With $\Phi_v = (K-v+1)S_v$ and $\Phi_{v+1} \geq (K-v)S_{v+1}$,

$$\alpha e_{[v]} = \Phi_v - \Phi_{v+1} \leq (K-v+1)S_v - (K-v)S_{v+1} = S_v + (K-v)\gamma'.$$

Since the slopes are non-increasing, $S_w - S_v = \sum_{i=w}^{v-1} c_i \geq (v-w)\gamma'$. Combining the three bounds,

$$\alpha(e_{[r]} - e_{[r']}) - (K-r)(\gamma - \gamma') \geq (K-r+1)\gamma + (v-w)\gamma' - (K-v)\gamma' - (K-r)(\gamma-\gamma') = \gamma + (2v-w-r)\gamma' \geq 0,$$

since $v \geq w > r$. □

*Proof of Proposition 4.2.* (i) Fix $t$ and a state $T$ with $|T| = t$, let $U_t = \{K-t+1, \ldots, K\}$, and let $F(T) = \sum_{r \in T}(t\, c_r - \alpha e_{[r]})$, so that, by Lemma C.4, the constraint (4.7) for $T$ reads $F(T) \leq 0$. If $T \neq U_t$, choose $r \in T \setminus U_t$ and $r' \in U_t \setminus T$. Then $r \leq K - t < r'$, so $t \leq K - r$, and replacing $r$ by $r'$ in $T$ changes $F$ by

$$\alpha(e_{[r]} - e_{[r']}) - t(c_r - c_{r'}) \geq \alpha(e_{[r]} - e_{[r']}) - (K-r)(c_r - c_{r'}) \geq 0,$$

by $c_r \geq c_{r'}$ and Lemma C.6. Repeating such exchanges reaches $U_t$, so

$$F(T) \leq F(U_t) = t\,(S_{K-t+1} - \alpha m_{K-t+1}) \leq 0.$$

Every constraint in (4.7) therefore holds. By Lemma C.4, $p^\sharp(e)$ is a measurable selection from the certified menu of $E^{\overline{\mathrm{eBH}}}$, and Theorem 2.1 gives $\mathbb{E}\,\mathrm{FDP}_T(R_{\kappa^\sharp}(E)) \leq \alpha\, \mathbb{E}_P E_T^{\overline{\mathrm{eBH}}} \leq \alpha$ for every nonempty $T$ and $P \in \mathcal{P}_T$.

(ii) Fix $r$, and let $v < w$ be consecutive vertices with $v \leq r < w$, so that $c_r = c_v$ and $S_r = S_v - (r-v)c_v$. Then

$$q_r^\sharp = rS_r - (r-1)S_{r+1} = S_v + (v-1)c_v.$$

If $v = 1$, then $q_r^\sharp = S_1 = \min\{1, \alpha m_1\} = q_1^* \geq q_r^*$. If $v = K$, then $r = K$, $c_K = S_K = \alpha e_{[K]}$ and $q_K^\sharp = K\alpha e_{[K]} \geq q_K^*$. Otherwise $2 \leq v \leq K-1$, so $S_v = \alpha m_v$ and $c_v = S_v - S_{v+1} \geq \alpha(m_v - m_{v+1})$. With $d = K - v + 1$ and $\bar{e} = m_{v+1}$, we have $m_v - m_{v+1} = (e_{[v]} - \bar{e})/d$ and $(d-1)\bar{e} = \sum_{j>v} e_{[j]}$, and a direct computation gives

$$q_r^\sharp \geq \alpha m_v + (v-1)\,\alpha\,\frac{e_{[v]} - \bar{e}}{d} = \frac{\alpha}{K}\Big(v\, e_{[v]} + \sum_{j>v} e_{[j]}\Big) + \frac{\alpha\, v(v-1)(e_{[v]} - \bar{e})}{dK} \geq q_v^* \geq q_r^*,$$

because $e_{[v]} \geq \bar{e}$ and $q^*$ is non-increasing. In all cases $q_r^\sharp \geq q_r^*$, so with the same $U$, $\{\kappa^* \geq r\} \subseteq \{\kappa^\sharp \geq r\}$ for every $r$, and $R_{\kappa^*}(e) \subseteq R_{\kappa^\sharp}(e)$.

(iii) Consider the linear program that maximizes $\sum_{k=1}^K k\, p_k$ over $p_1, \ldots, p_K \geq 0$ with $\sum_k p_k \leq 1$, subject only to the constraints for the states $\{s, \ldots, K\}$,

$$S_s(p) = \sum_{k=1}^{K} p_k \frac{(k-s+1)_+}{k} \leq \alpha m_s, \qquad s = 1, \ldots, K.$$

By the accounting identity, its objective is $\sum_r (2r-1)c_{[r]}(p) = S_1(p) + 2\sum_{s \geq 2} S_s(p)$, which increases with every $S_s(p)$. Every feasible $p$ satisfies $S(p) \leq S$ by Lemma C.5, and $p^\sharp$ is feasible with $S(p^\sharp) = S$, so the

value of the program is $\mathbb{E}[\kappa^\sharp \mid E = e]$. By linear programming duality, there are $\lambda_1, \dots, \lambda_K \geq 0$ and $\mu \geq 0$ with $\sum_s \lambda_s \alpha m_s + \mu = \mathbb{E}[\kappa^\sharp \mid E = e]$ and

$$\sum_{s=1}^{K} \lambda_s \frac{(k-s+1)_+}{k} + \mu \geq k, \qquad k = 1, \dots, K.$$

A set $R$ of size $k \geq 1$ contains at least $(k-s+1)_+$ of the $K-s+1$ hypotheses of rank at least $s$, so

$$|R| \leq \sum_{s=1}^{K} \lambda_s \frac{|R \cap \{s, \dots, K\}|}{|R| \vee 1} + \mu \, \mathbb{1}\{R \neq \varnothing\},$$

which also holds for $R = \varnothing$. Let $Q$ be a distribution of rejection sets certified by $E^{\overline{\text{eBH}}}$. Its averaged false discovery proportion in the state $\{s, \dots, K\}$ is at most $\alpha m_s$, so taking expectations under $Q$ gives

$$\mathbb{E}_Q|R| \leq \sum_s \lambda_s \, \mathbb{E}_Q\Big[\frac{|R \cap \{s, \dots, K\}|}{|R| \vee 1}\Big] + \mu \, Q(R \neq \varnothing) \leq \sum_s \lambda_s \alpha m_s + \mu = \mathbb{E}[\kappa^\sharp \mid E = e].$$

The rejection set of closed e-BH is certified by $E^{\overline{\text{eBH}}}$, which proves the last assertion. □

## C.5 The minimally adaptive improvement

We show that the minimally adaptive principle yields a valid rule that rejects at least as much as RU-eBH and sometimes strictly more, so that RU-eBH is not admissible. For a level $\gamma > 0$, write $\kappa^*_\gamma$ for the rule $\kappa^*$ computed at level $\gamma$ instead of $\alpha$, with the same $U$, so that $\kappa^* = \kappa^*_\alpha$. Let $K \geq 2$ and $\alpha' = K\alpha/(K-1)$, and define

$$R^\dagger = \begin{cases} R_{\kappa^*_{\alpha'}}(E), & \kappa^*_\alpha > 0, \\ \emptyset, & \kappa^*_\alpha = 0. \end{cases}$$

**Proposition C.7.** *The rule $R^\dagger$ has the following properties.*

*(i) Its FDR is at most $\alpha$ in every state $T$ and under every law in $\mathcal{P}_T$.*

*(ii) It contains the RU-eBH rejection set for every $E$ and $U$, and under some law it makes strictly more true discoveries with positive probability.*

*(iii) It is certified by $E^{\overline{\text{eBH}}}$, so RC-eBH rejects at least as many hypotheses as $R^\dagger$ in expectation given $E = e$.*

*Proof.* Let $B^\dagger_T = E^{\text{eBH}}_T$ for the full set $T = \{1, \dots, K\}$ and $B^\dagger_T = (K-1)^{-1} \sum_{j \in T} E_j$ for every other nonempty $T$. This is an e-collection, because a state other than the full set contains at most $K-1$ true nulls, and it exceeds $E^{\text{eBH}}$ on every proper nonempty state.

(i) Given $E = e$, $R^\dagger$ is a monotone rule, and we bound its averaged loss in each state. In the full state, every rejection is false, so the averaged loss is $\mathbb{P}(\kappa^*_\alpha > 0 \mid E = e) = q^*_1 \leq (\alpha/K) \sum_j e_j = \alpha B^\dagger_T$. In any other nonempty state $T$, the set $R^\dagger$ is either $R_{\kappa^*_{\alpha'}}$ or empty, so its averaged loss is at most that of $R_{\kappa^*_{\alpha'}}$. By the menu (4.4) at level $\alpha'$, the latter is at most $(\alpha'/K) \sum_{j \in T} e_j = \alpha B^\dagger_T$. Hence $R^\dagger$ selects from the certified menu of $B^\dagger$, measurably in $e$, and Theorem 2.1 gives an FDR of at most $\alpha \, \mathbb{E}_P B^\dagger_T \leq \alpha$.

(ii) The tails $q^*_r$ are non-decreasing in the level, so $\kappa^*_{\alpha'} \geq \kappa^*_\alpha$, and $R^\dagger \supseteq R_{\kappa^*_\alpha}(E)$ for every $E$ and $U$. For the strict improvement, let $e^\circ = (1/\alpha, 1/(2\alpha), \dots, 1/(2\alpha))$. At level $\alpha$, $q^*_1 = (K+1)/(2K)$ and $q^*_r = 1/2$ for $r \geq 2$, whereas at level $\alpha'$, every $q^*_r$ is at least $K/(2(K-1)) > (K+1)/(2K)$. For $1/2 < U \leq (K+1)/(2K)$, RU-eBH therefore rejects only $H_1$, whereas $R^\dagger$ rejects all $K$ hypotheses. Under the law with $T = \{1\}$ in which $E = e^\circ$ with probability $\alpha$ and $E = 0$ otherwise, $R^\dagger$ makes $K-1$ more true discoveries than RU-eBH with probability $\alpha/(2K)$.

(iii) Since a nonempty state $T \neq \{1, \dots, K\}$ has $|T| \leq K-1$, we have $B^\dagger \leq E^{\overline{\text{eBH}}}$, with equality on the full set. The certified menu of $B^\dagger$ is therefore contained in that of $E^{\overline{\text{eBH}}}$, and Proposition 4.2(iii) applies. □

## C.6 Conditional calibration as recertification

This subsection spells out the correspondence summarized in Remark 4.3. We restate the e-BH version of conditional calibration of Lee and Ren [2024] in our terms, without modifying it. Let $e_1(X), \dots, e_K(X)$ be finite e-values computed from data $X$, let $R_0 = \mathrm{eBH}_\alpha(e(X))$ and $r = |R_0|$, and for each $j$ and $c \geq 1$ let

$$t_j(X) = \frac{K}{\alpha\,|R_0 \cup \{j\}|}, \qquad h_j(c, X) = t_j(X)\,\mathbb{1}\{c\,e_j(X) \geq t_j(X)\}.$$

For $j \in R_0$, $t_j = K/(\alpha r)$ is the self-consistency threshold that $e_j$ already passes, and for $j \notin R_0$, $t_j = K/(\alpha(r+1))$ is the value that $e_j$ would need for $R_0 \cup \{j\}$ to be self-consistent. The quantity $h_j(c, X)$ inflates $e_j$ by the factor $c$ and then records only whether the threshold is reached; it is non-decreasing in $c$.

At $c = 1$, these are the pruned e-values of Remark 4.3,

$$h_j(1, X) = \frac{K\,\mathbb{1}\{j \in R_0\}}{\alpha(r \vee 1)} \leq e_j(X).$$

Indeed, every member of $R_0$ passes its threshold, whereas no $j \notin R_0$ can satisfy $e_j \geq K/(\alpha(r+1))$, since all $r+1$ members of $R_0 \cup \{j\}$ would then reach this value and e-BH would reject at least $r+1$ hypotheses. Applying e-BH to the pruned e-values therefore rejects exactly $R_0$. They retain all the evidence that e-BH uses, and the difference $e_j - h_j(1, X) \geq 0$ is the evidence that it discards.

Conditional calibration spends this discarded evidence. Suppose that under $H_j$ a statistic $S_j$ makes the conditional law of $X$ given $S_j$ known and common to all distributions in $H_j$. Then

$$\Phi_j(c; S_j) = \mathbb{E}[h_j(c, X) - e_j(X) \mid S_j]$$

can be computed, for instance by sampling from this conditional law. It is non-decreasing in $c$ and nonpositive at $c = 1$, so the inflation can be raised until $\Phi_j$ reaches zero. For an $S_j$-measurable $c_j \geq 1$ with $\Phi_j(c_j; S_j) \leq 0$, the recalibrated value $e'_j = h_j(c_j, X)$ satisfies

$$\mathbb{E}[e'_j \mid S_j] = \mathbb{E}[e_j \mid S_j] + \Phi_j(c_j; S_j) \leq \mathbb{E}[e_j \mid S_j],$$

because $c_j$ is fixed once $S_j$ is known, and it is therefore an e-value under $H_j$. For $j \in R_0$, $c_j e_j \geq e_j \geq t_j$, so $e'_j = K/(\alpha r)$, and e-BH applied to $e'$ rejects at least $R_0$. A hypothesis outside $R_0$ is added whenever its inflated e-value reaches $K/(\alpha(r+1))$.

This is exactly Proposition 2.2 in the language of Section 2.3. The conditioning statistic is $S_j$, the baseline is the pruned e-value $h_j(1, X)$ with conditional mean $H = \mathbb{E}[h_j(1, X) \mid S_j]$, and the budget is $B = \mathbb{E}[e_j \mid S_j]$, whose mean is at most one under $H_j$, although it may exceed one for some values of $S_j$. The increment $W = h_j(c_j, X) - h_j(1, X)$ is nonnegative, and since $B - H = \mathbb{E}[e_j - h_j(1, X) \mid S_j]$, the condition $\Phi_j(c_j; S_j) \leq 0$ is exactly (2.9). The recalibrated e-values then certify rejection sets through the e-collection (4.1) evaluated at $e'$, whose largest certified set is $\mathrm{eBH}_\alpha(e')$. Conditional calibration thus spends unused budget across the values of the data that share the same $S_j$, whereas RU-eBH spends it across the values of the external seed.

Finally, Lee and Ren [2024] take the largest admissible factor $c_j^* = \sup\{c \geq 1 : \Phi_j(c; S_j) \leq 0\}$, which need not be attained, because $h_j(c, X)$ jumps in $c$. If $\Phi_j(c_j^*; S_j) > 0$, the value $t_j(X)\,\mathbb{1}\{c_j^* e_j(X) > t_j(X)\}$, with strict inequality, is the increasing limit of $h_j(c, X)$ as $c \uparrow c_j^*$, so monotone convergence preserves the bound on its conditional mean, and since $c_j^* > 1$, it still equals $K/(\alpha r)$ on $R_0$.