Bayes with No Shame: Admissibility Geometries of Predictive Inference

Anonymous authors
Paper under double-blind review

Abstract

Four distinct admissibility geometries govern sequential and distribution-free inference: Blackwell risk dominance over convex risk sets, anytime-valid admissibility within the nonnegative supermartingale cone, marginal coverage validity over exchangeable prediction sets, and Cesàro approachability (CAA) admissibility, which reaches the risk-set boundary via approachability-style arguments rather than explicit priors. We prove a *criterion separation theorem*: the four classes of admissible procedures are pairwise non-nested. Each geometry carries a different certificate of optimality: a supporting-hyperplane prior (Blackwell), a nonnegative supermartingale (anytime-valid), an exchangeability rank (coverage), or a Cesàro steering argument (CAA). Martingale coherence is necessary for Blackwell admissibility and necessary and sufficient for anytime-valid admissibility within e-processes, but is not sufficient for Blackwell admissibility and is not necessary for coverage validity or CAA-admissibility. All four criteria can be viewed through a common schematic template (minimize Bayesian risk subject to a feasibility constraint), but the decision spaces, partial orders, and performance metrics differ by criterion, making them geometrically incompatible. Admissibility is irreducibly criterion-relative.

1 Introduction

In a remarkable lecture, Blackwell (1956b) posed the question: given an arbitrary binary sequence, how well can a prediction method perform? He constructed two predictors and showed that each is optimal in its own sense. The *minimax predictor*, derived from Blackwell's own vector minimax theorem (Blackwell, 1956a), achieves the guarantee that its long-run proportion of correct predictions matches or exceeds the best constant predictor, for *every* sequence; the proof uses an approachability argument and the predictor is randomized. The *Bayes predictor*, which simply predicts the more likely outcome given the past, satisfies the same guarantee under any stochastic process; the proof is elementary (the strong law of large numbers), and the predictor is deterministic. Neither dominates the other: the minimax predictor works for every sequence but is randomized and has no per-round optimality certificate; the Bayes predictor is optimal at each round under the model but requires a stochastic assumption.

These competing optimality notions reappear today across ML subfields: proper scoring rules for LLM calibration, e-processes for safe A/B testing, conformal prediction for uncertainty quantification, and online calibration via defensive forecasting for adversarial robustness. Each community has its own notion of "optimal," and each notion is internally coherent, yet they are mutually incompatible.

This contrast between minimax and Bayes prediction is the seed of the present paper. Consider the same question under log loss. The plug-in algorithm $\hat{p}_n = S_n/n$ is a martingale under its own predictive law and appears well-calibrated, yet it is strictly dominated for every $\theta \in (0,1)$ and every sample size $n \geq 1$ by the Bayes predictive $\hat{p}_n^B = (S_n + \frac{1}{2})/(n+1)$. The plug-in assigns probability zero to events that occur with positive probability, producing infinite risk. Its martingale coherence does not rescue it from inadmissibility, a direct counterexample to the sufficiency of martingale posteriors (Fong et al., 2023). Meanwhile, conformal prediction sets achieve distribution-free coverage without optimizing any loss; e-processes control type-I error at every stopping time by a structural condition, the nonnegative martingale property, that has no analogue in

classical risk theory; and defensive forecasters (Vovk et al., 2005b; Chernov et al., 2010), modern descendants of Blackwell's minimax predictor, achieve calibration in the Cesàro sense (Cesa-Bianchi & Lugosi, 2006, Ch. 4) through fixed-point arguments, without optimizing any per-round loss function. Each algorithm is "optimal" in its own sense, yet no single criterion governs all four. The reason is structural: each criterion operates on a different space of procedures (point predictors, test processes, prediction sets, or sequential strategies) and induces a different partial order, so that admissibility in one geometry carries no logical implication for admissibility in another.

We use the term *no-shame* informally, following Williams (1993), to denote admissibility: a practitioner who deploys a dominated algorithm faces a self-evident indictment from the very risk function she specified. A "no-shame" strategy is one for which no such indictment is possible: the rule sits on the lower boundary of the risk set, and no alternative achieves uniformly lower risk. But what counts as shameful depends on which standard one adopts: a Bayes-optimal point predictor is no-shame under risk dominance yet produces a prediction set with zero coverage; a conformal set achieves valid coverage yet does not minimize any proper scoring rule; a defensive forecaster reaches the risk-set boundary in the long run yet is not Bayes optimal at any finite sample size. The criterion separation theorems (Theorems 5.9 and 6.7) make this pluralism precise and show it is structural rather than a matter of approximation.

The present moment is especially prone to cross-talk because four active research programs (proper scoring rules in predictive modeling (Gneiting & Raftery, 2007), safe anytime-valid inference (Ramdas et al., 2023; Grünwald et al., 2024), conformal prediction for uncertainty quantification (Vovk et al., 2005a; Angelopoulos & Bates, 2023), and online learning via defensive forecasting and Blackwell approachability (Vovk et al., 2005b; Abernethy et al., 2011; Rakhlin & Sridharan, 2013)) all speak in the language of "optimality," but relative to different objects, orders, and certificates. Our contribution is to make these differences explicit in a common geometric language, rather than to propose a new inferential paradigm.

Contributions.

(1) We formalize four admissibility geometries (Blackwell, anytime-valid, coverage, CAA) and prove a *criterion separation theorem*: the four classes of admissible procedures are pairwise non-nested (Theorems 5.9–6.7).
(2) We establish that martingale coherence is necessary for Blackwell admissibility and equivalent to anytime-valid admissibility within e-processes, but is not sufficient for Blackwell admissibility and irrelevant to coverage or CAA-admissibility (Theorem 4.3, Proposition 5.4).
(3) We introduce a *constrained Bayes schema* (Definition 5.8) that provides a common viewpoint on all four criteria: specify a validity constraint $\mathcal{F}$ within the appropriate decision space $\mathcal{D}_C$, then optimize Bayesian risk within $\mathcal{F}$. This gives practitioners a design recipe that adapts to each evaluation paradigm.
(4) We provide a Bernoulli and Gaussian laboratory with six canonical procedures that constructively witness each separation, together with Monte Carlo experiments confirming the theoretical distinctions in finite samples.

This paper isolates four such criteria, each defined relative to a different performance objective and a different certificate of optimality for predictive algorithms. *Blackwell admissibility* (Sections 2–3) requires that no competing rule have uniformly lower risk over Θ; the certificate is a supporting-hyperplane prior, and by Corollary 3.13, every such rule is Bayes or a limit of Bayes rules. *Anytime-valid admissibility* (Section 5) is defined within the class $\mathcal{C}_{\mathrm{AV}}$ of e-processes, where admissibility is equivalent to the nonnegative martingale property (Ramdas et al., 2022); the certificate is a nonnegative supermartingale. *Marginal coverage validity* (Section 5) requires $\mathbb{P}(Y_{n+1} \in \hat{C}_n(X_{n+1})) \geq 1 - \alpha$ under exchangeability (Vovk et al., 2005a); the certificate is an exchangeability rank. *CAA-admissibility* (Section 6) requires that the time-averaged risk converge to the lower boundary $\partial_- \mathcal{R}$ for every θ; the certificate is a Cesàro steering argument guaranteed by a fixed-point or minimax construction. Theorems 5.9 and 6.7 establish that the four classes $\mathfrak{B}$, $\mathfrak{A}$, $\mathfrak{C}$, $\mathfrak{D}$ of admissible procedures under these criteria are pairwise non-nested; the proof is constructive via canonical Bernoulli procedures (Section 7). In short: Bayes reaches $\partial_- \mathcal{R}$ by supporting hyperplanes; approachability reaches $\partial_- \mathcal{R}$ by time-averaged steering.

The content of the separation theorems is that the non-nesting persists even when all four frameworks are applied to the *same* Bernoulli learning problem: each imposes a different partial order on procedures derived from that process, and these partial orders admit no common refinement.

Sections 2–3 set out primitives and risk-set geometry; Section 4 the martingale layer; Sections 5–7 the criteria and separation theorem; Section 6 the fourth geometry; Section 8 Monte Carlo illustrations; Section 10 implications.

Each paradigm corresponds to a distinct admissibility geometry; the separation theorems show these notions of optimality do not admit a single common ranking.

2 Primitive Objects

The decision-theoretic framework requires five objects: a parameter space, an action space, a loss function, a sample space, and a statistical model. We adopt the extended-real formulation that allows $+\infty$ risk, accommodating proper scoring rules such as log loss from the outset.

Definition 2.1 (Statistical decision problem)**.** A *statistical decision problem* is a tuple $(\Theta, \mathcal{A}, L, \mathcal{X}, \mathcal{P})$ where:

(i) $\Theta \subset \mathbb{R}^d$ is the *parameter space*, compact and metrizable.
(ii) $\mathcal{A} \subset \mathbb{R}^m$ is the *action space*, compact and metrizable.
(iii) $\mathcal{X}$ is the *sample space*, a Polish space, and $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is the statistical model.
(iv) $L : \Theta \times \mathcal{A} \to [0, \infty]$ is the *loss function*, satisfying:
  (a) $L(\theta, a)$ is measurable in θ for every a;
  (b) $L(\theta, \cdot)$ is lower semicontinuous for every θ;
  (c) L is bounded below (by zero, without loss of generality).

Risks are allowed to take the value $+\infty$; dominance is defined in the extended-real sense via the coordinatewise ordering on $[0, \infty]^\Theta$.

Compactness is adopted to streamline existence and closedness arguments; see Remark 3.9 for scope.

Definition 2.2 (Decision rules)**.** Given data $X^n = (X_1, \ldots, X_n) \in \mathcal{X}^n$, a *(randomized) decision rule* is a measurable map $\delta : \mathcal{X}^n \to \Delta(\mathcal{A})$, where $\Delta(\mathcal{A})$ denotes the probability measures on $\mathcal{A}$. The class $\mathcal{D}$ of all decision rules is convex: for $\lambda \in [0, 1]$, the mixture $\delta_\lambda = \lambda\delta_1 + (1 - \lambda)\delta_2$ is defined by drawing from δ_1 or δ_2 with probabilities λ and $1 - \lambda$ independently of X^n.

Definition 2.3 (Risk function)**.** The *risk function* of $\delta \in \mathcal{D}$ is

$$R(\theta, \delta) = \mathbb{E}_\theta\big[L(\theta, \delta(X^n))\big] \ \in \ [0, \infty], \qquad \theta \in \Theta.$$

Bayes risks are well-defined in $[0, \infty]$. Under Definition 2.1, $R(\theta, \cdot)$ is lower semicontinuous and convex on $\mathcal{D}$ for each θ; lower semicontinuity follows from Fatou's lemma and the lower semicontinuity of $L(\theta, \cdot)$.

Definition 2.4 (Dominance and admissibility)**.** The partial order on $[0, \infty]^\Theta$ is defined by $r \le r'$ if and only if $r(\theta) \le r'(\theta)$ for all $\theta \in \Theta$. Rule δ' *dominates* δ if $R(\theta, \delta') \le R(\theta, \delta)$ for all $\theta \in \Theta$ with strict inequality for some $\theta_0 \in \Theta$. A rule δ is *Blackwell admissible* if no rule in $\mathcal{D}$ dominates it.

Definition 2.5 (Risk set)**.** For $\Theta = \{\theta_1, \ldots, \theta_k\}$ finite, associate to each $\delta \in \mathcal{D}$ its *risk vector* $r(\delta) = \big(R(\theta_1, \delta), \ldots, R(\theta_k, \delta)\big) \in [0, \infty]^k$. The *risk set* is

$$\mathcal{R} \ = \ \big\{r(\delta) : \delta \in \mathcal{D}\big\} \ \subset \ [0, \infty]^k.$$

The risk set is the image of the decision space under the risk map; its geometry encodes which rules dominate which.

3 Geometry of No-Shame

Admissibility has a geometric characterization: a rule is admissible if and only if its risk vector lies on the lower boundary of the risk set. This section establishes convexity, existence of Bayes rules, and closedness

of the risk set, and shows that every admissible rule is supported by a prior, the geometric content of the no-shame principle.

In ML terms, Blackwell admissibility is the gold standard for probabilistic forecasters: no alternative model achieves uniformly lower expected loss across all possible data-generating processes. When a practitioner deploys a model on a leaderboard, Blackwell admissibility asks whether any other model dominates it everywhere; if not, the model is no-shame.

3.1 Convexity of the risk set

Lemma 3.1 (Convexity)**.** *Under Definitions 2.1–2.5, $\mathcal{R}$ is convex.*

Proof. Let $\delta_1, \delta_2 \in \mathcal{D}$ and $\lambda \in [0,1]$. The mixture δ_λ satisfies $R(\theta_j, \delta_\lambda) = \lambda R(\theta_j, \delta_1) + (1-\lambda) R(\theta_j, \delta_2)$ for each j, so $r(\delta_\lambda) = \lambda r(\delta_1) + (1-\lambda) r(\delta_2) \in \mathcal{R}$. □

Randomization is standard here: it preserves convexity and compactness of the attainable risk set and underlies complete-class characterizations.

3.2 Existence of Bayes rules

Lemma 3.2 (Existence of Bayes rules via Berge)**.** *Let Π be a prior with full support on Θ. Define the Bayes risk $r(\delta, \Pi) = \int_\Theta R(\theta, \delta)\, \mathrm{d}\Pi(\theta)$. Then:*

(i) The mapping $\delta \mapsto r(\delta, \Pi)$ is lower semicontinuous on $\mathcal{D}$ equipped with the weak topology.
(ii) The decision space $\mathcal{D}$ (randomized rules with the weak topology) is compact.
(iii) Hence a Bayes rule $\delta_\Pi \in \arg\min_{\delta \in \mathcal{D}} r(\delta, \Pi)$ exists.

Proof. Lower semicontinuity of $\delta \mapsto r(\delta, \Pi)$ follows from Fatou's lemma and the lower semicontinuity of $L(\theta, \cdot)$ (Definition 2.1). Compactness of $\mathcal{D}$ follows from Prokhorov's theorem and Tychonoff's theorem on the product space $\Delta(\mathcal{A})^{\mathcal{X}^n}$. The Berge Maximum Theorem (Berge, 1963, Ch. VI) then guarantees existence of a minimizer; see also Wald (1950) and Blackwell & Girshick (1954). □

3.3 Closedness of the risk set

Proposition 3.3 (Lower-comprehensive closedness)**.** *Under Definition 2.1, $\mathcal{R} \subset [0,\infty]^k$ satisfies the following property: for any net $(r(\delta_\alpha))$ in $\mathcal{R}$ converging to r^* in the product topology of $[0,\infty]^k$, there exists $\delta^* \in \mathcal{D}$ with $r(\delta^*) \le r^*$ coordinatewise and $r(\delta^*) \in \mathcal{R}$. In particular, the lower set $\mathcal{R}^+ = \{r' \in [0,\infty]^k : \exists\, r \in \mathcal{R},\ r \le r'\}$ is closed.*

Proof. Let $(r(\delta_\alpha))$ be a net in $\mathcal{R}$ converging to r^*. Compactness of $\mathcal{D}$ (Lemma 3.2) supplies a subnet with $\delta_\alpha \to \delta^*$ weakly. Lower semicontinuity of $R(\theta_j, \cdot)$ for each j gives $R(\theta_j, \delta^*) \le \liminf R(\theta_j, \delta_\alpha) \le r^*_j$, so $r(\delta^*) \in \mathcal{R}$ with $r(\delta^*) \le r^*$ coordinatewise. □

Remark 3.4. Lower-comprehensive closedness suffices for the supporting-hyperplane argument (Theorem 3.8): the separation of $\partial_- \mathcal{R}$ from the interior is a property of the lower boundary, not of the full risk set. Actual closedness of $\mathcal{R}$ would require continuity (not merely lower semicontinuity) of the risk map, which is not assumed here.

Remark 3.5. Extended-real values do not break convex separation. The supporting hyperplane argument in Theorem 3.8 is applied to $\mathcal{R} \cap \mathbb{R}^k_+$, not to $[0,\infty]^k$ directly. Any admissible point r^* has finite coordinates under the prior that supports it: if $R(\theta_j, \delta^*) = +\infty$ for some j with $\Pi(\theta_j) > 0$, then the integrated Bayes risk would be infinite and δ^* could not be a minimizer. Hence separation occurs in $\mathbb{R}^k$. Lower-comprehensive closedness (Proposition 3.3) ensures that the lower boundary $\partial_- \mathcal{R}$ is well-defined and that supporting hyperplanes exist at each boundary point.

3.4 Lower boundary and admissibility

Definition 3.6 (Lower boundary)**.** The *lower boundary* of $\mathcal{R}$ is

$$\partial_{-}\mathcal{R} \;=\; \big\{r \in \mathcal{R} : \nexists\, r' \in \mathcal{R},\; r' \le r \text{ coordinatewise with } r' \neq r\big\}.$$

Proposition 3.7 (Boundary characterization)**.** *A rule δ is Blackwell admissible if and only if $r(\delta) \in \partial_{-}\mathcal{R}$.*

Proof. If $r(\delta) \notin \partial_{-}\mathcal{R}$, there exists $r' \in \mathcal{R}$ with $r' \le r(\delta)$ coordinatewise, $r' \neq r(\delta)$; the corresponding rule dominates δ. Conversely, if $r(\delta) \in \partial_{-}\mathcal{R}$, no such r' exists in $\mathcal{R}$. □

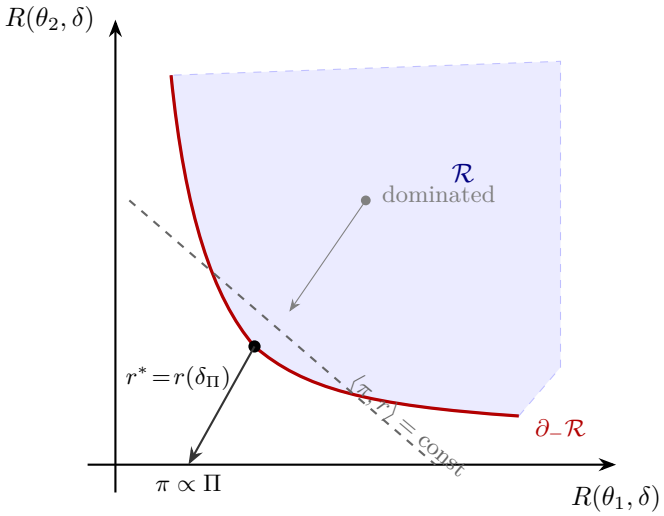

Figure 1: Risk set geometry for $|\Theta| = 2$. The convex risk set $\mathcal{R}$ (shaded) maps each decision rule to a risk vector. The lower boundary $\partial_{-}\mathcal{R}$ (bold curve) contains all admissible rules. At an admissible point r^*, the supporting hyperplane (dashed line) identifies the prior Π whose normal π defines the Bayes problem that r^* solves (Theorem 3.8). Interior points are dominated.

3.5 Supporting hyperplanes and Bayes rules

Theorem 3.8 (Supporting hyperplane identification)**.** *Assume $|\Theta| = k$. If $r^* \in \partial_{-}\mathcal{R}$ and $\mathcal{R}$ is convex, there exists $\pi \in \mathbb{R}^k_+ \setminus \{0\}$ such that*

$$\sum_{j=1}^{k} \pi_j R(\theta_j, \delta^*) \;\le\; \sum_{j=1}^{k} \pi_j R(\theta_j, \delta) \qquad \textit{for all } \delta \in \mathcal{D}.$$

Setting $\Pi = \pi / \|\pi\|_1$ defines a prior on Θ, and δ^* is a Bayes rule:

$$\delta^* \;\in\; \arg\min_{\delta \in \mathcal{D}} \int R(\theta, \delta)\; \mathrm{d}\Pi(\theta).$$

Proof. By Lemma 3.1, $\mathcal{R}$ is convex. Since $r^* \in \partial_{-}\mathcal{R}$, the set $\{r \in \mathbb{R}^k_+ : r \le r^*, r \neq r^*\} \cap \mathcal{R} = \emptyset$. The separating hyperplane theorem (Hahn–Banach) supplies a nonzero $\pi \in (\mathbb{R}^k)^*$ with $\langle \pi, r^* \rangle \le \langle \pi, r \rangle$ for all $r \in \mathcal{R}$. Each $\pi_j \ge 0$: if $\pi_j < 0$ for some j, decreasing $R(\theta_j, \delta)$ while holding other coordinates fixed would strictly decrease $\langle \pi, r \rangle$, contradicting minimality at r^*. Normalizing π yields Π and identifies $\delta^* = \delta_\Pi$. □

Remark 3.9 (Scope of Theorem 3.8). The finite-Θ assumption is adopted for expositional clarity. The result extends to compact Θ via weak* compactness of $\Delta(\Theta)$, the Banach–Alaoglu theorem, and standard measurable selection arguments; see Blackwell & Girshick (1954) and Wald (1950) for the general development.

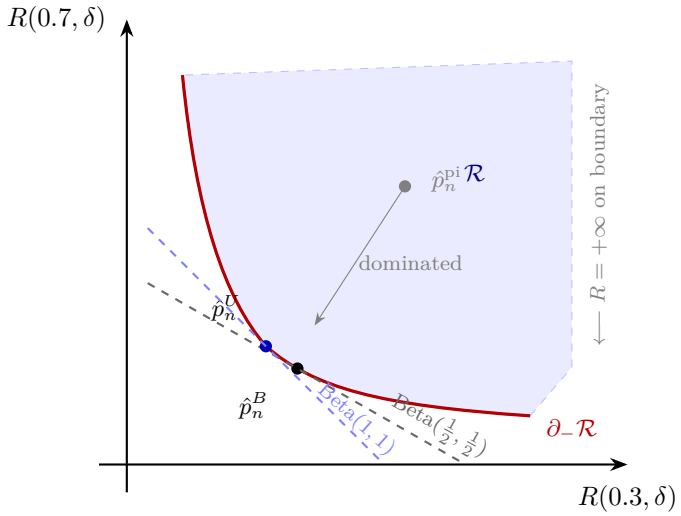

Figure 2: Concrete risk set for Bernoulli log-loss prediction with $\Theta = \{0.3, 0.7\}$, $n = 10$. The Bayes predictive $\hat{p}_n^B = (S_n + \frac{1}{2})/(n+1)$ under Beta$(\frac{1}{2}, \frac{1}{2})$ and the Laplace predictive $\hat{p}_n^U = (S_n + 1)/(n+2)$ under Beta$(1, 1)$ both lie on the lower boundary $\partial_- \mathcal{R}$; each is no-shame with respect to a different prior (dashed supporting hyperplanes). The plug-in MLE $\hat{p}_n^{\mathrm{pi}} = S_n/n$ lies in the interior: its risk vector is dominated because it assigns zero probability to events that occur with positive probability, producing infinite log-loss contributions.

Lemma 3.10 (Bayes rules lie on the lower boundary)**.** *Under compact Θ and lower semicontinuous loss, every proper Bayes rule δ_Π with finite integrated risk has $r(\delta_\Pi) \in \partial_- \mathcal{R}$.*

Proof. If $r(\delta_\Pi)$ were not in $\partial_- \mathcal{R}$, there would exist $\delta' \in \mathcal{D}$ with $R(\theta_j, \delta') \leq R(\theta_j, \delta_\Pi)$ for all j and strict inequality for some j_0. Then $\int R(\theta, \delta')\, \mathrm{d}\Pi < \int R(\theta, \delta_\Pi)\, \mathrm{d}\Pi$, contradicting Bayes optimality of δ_Π. □

Theorem 3.11 (Wald–Brown complete class (Wald, 1950; Blackwell & Girshick, 1954))**.** *Under Definition 2.1, every Blackwell admissible rule is a Bayes rule with respect to some prior Π on Θ, or a pointwise limit of Bayes rules. Equivalently, the class of Bayes rules is essentially complete.*

3.6 No-shame strategies

Definition 3.12 (No-shame strategy)**.** A rule $\delta \in \mathcal{D}$ is a *no-shame strategy* if it is Blackwell admissible; equivalently (by Proposition 3.7), if $r(\delta) \in \partial_- \mathcal{R}$.

Corollary 3.13 (No-shame is Bayes-supported)**.** *Under Definitions 2.1–2.5:*

(i) Every Bayes rule is no-shame.
(ii) Every no-shame rule is a Bayes rule or a pointwise limit of Bayes rules (complete-class closure, Theorem 3.11).

Proof. (i) If δ_Π were dominated by δ', then $\int R(\theta, \delta')\, \mathrm{d}\Pi < \int R(\theta, \delta_\Pi)\, \mathrm{d}\Pi$, contradicting Bayes optimality. *(ii)* By Theorem 3.11 (Wald–Brown), the class of Bayes rules is essentially complete: every admissible rule is Bayes or a limit of Bayes rules. □

Remark 3.14 (Order-theoretic structure). Admissibility is defined by the coordinatewise partial order on $\mathbb{R}^k$, not by any metric. The supporting hyperplane in Theorem 3.8 is the linear-algebraic instrument for locating the prior Π that rationalizes δ^*; the dominance relation itself depends only on the order structure of $\mathbb{R}^k$. When validity constraints are imposed (anytime-valid error control or marginal coverage), the optimization in Theorem 3.8 restricts to a feasible subset of $\mathcal{D}$; the constrained Bayes formulation in Section 5.4 makes this precise.

3.7 Duality and the Lagrangian formulation

The no-shame characterization (Corollary 3.13) identifies admissible rules as Bayes solutions. This identification has a natural dual: the prior Π that supports an admissible risk point r^* can be recovered as a Lagrange multiplier in a constrained optimization problem. Economically, each prior weight π_j is a *shadow price*: it measures the marginal cost, in terms of risk at θ_1, of tightening the risk constraint at θ_j. Admissibility therefore has a dual interpretation as efficient resource allocation across parameter values. This shadow-price interpretation reappears in the constrained Bayes formulation (Definition 5.8). The constrained Bayes formulation in Section 5.4 extends this duality by introducing an explicit feasibility constraint $\mathcal{F} \subseteq \mathcal{D}$; the Lagrange multipliers then reflect both the prior and the binding validity requirement.

Proposition 3.15 (Lagrangian dual of risk minimization)**.** *Let $r^* \in \partial_- \mathcal{R}$ and let $\pi \in \mathbb{R}^k_+ \setminus \{0\}$ be the normal to the supporting hyperplane at r^* (Theorem 3.8). Define the Lagrangian*

$$\mathcal{L}(\delta, \lambda) = R(\theta_1, \delta) + \sum_{j=2}^{k} \lambda_j \big[R(\theta_j, \delta) - c_j\big], \qquad \lambda_j \geq 0,$$

where $c_j = R(\theta_j, \delta^*)$ for $j = 2, \ldots, k$. Then the prior weights satisfy $\pi_j/\pi_1 = \lambda_j^*$, and the primal problem $\min_\delta R(\theta_1, \delta)$ subject to $R(\theta_j, \delta) \leq c_j$ ($j \geq 2$) has the same solution $\delta^* = \delta_\Pi$ as the unconstrained Bayes problem under Π.

Proof. Because the risk set $\mathcal{R}$ is convex (Lemma 3.1), the supporting-hyperplane theorem (Theorem 3.8) provides a half-space certificate $\langle \pi, r \rangle \geq \langle \pi, r^* \rangle$ at every boundary point. Interpreting $\pi_1 > 0$ without loss of generality (at least one coordinate of r^* is finite with positive weight), the supporting-hyperplane normal π can be read as Lagrange multipliers $\lambda_j^* = \pi_j/\pi_1$, $j \geq 2$. Complementary slackness at $r^* \in \partial_- \mathcal{R}$ then identifies δ^* with the minimizer of $\mathcal{L}(\cdot, \lambda^*)$, recovering the first-order condition of the Bayes problem under Π; feasibility $R(\theta_j, \delta^*) \leq c_j$ is ensured by $r^* \in \mathcal{R}$. (This geometric argument requires only convexity of $\mathcal{R}$ and the existence of a supporting hyperplane; no additional constraint qualification beyond the boundary structure is needed.) □

4 Martingale Layer

The risk-set geometry of Section 3 characterizes admissibility through priors and supporting hyperplanes. We now introduce a dynamic structure: the martingale property of Bayesian posterior predictive sequences. This property is necessary for Blackwell admissibility but, as the plug-in example will show, not sufficient.

Definition 4.1 (Posterior predictive sequence)**.** Let Π be a prior on Θ with $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. The *posterior predictive sequence* is

$$m_n \;=\; \mathbb{E}_\Pi[\theta \mid \mathcal{F}_n], \qquad n \geq 0,$$

with $m_0 = \mathbb{E}_\Pi[\theta]$.

Proposition 4.2 (Bayes implies martingale)**.** *Under Definition 4.1, $(m_n)_{n \geq 0}$ is a martingale with respect to $(\mathcal{F}_n)$ under the prior predictive measure $\tilde{P} = \int P_\theta \, \mathrm{d}\Pi(\theta)$.*

Proof. By the tower property under $\tilde{P}$: $\mathbb{E}_{\tilde{P}}[m_n \mid \mathcal{F}_{n-1}] = \mathbb{E}_{\tilde{P}}[\mathbb{E}_\Pi[\theta \mid \mathcal{F}_n] \mid \mathcal{F}_{n-1}] = \mathbb{E}_\Pi[\theta \mid \mathcal{F}_{n-1}] = m_{n-1}$ a.s. Integrability holds since θ is bounded on compact Θ. □

4.1 Martingale coherence is not sufficient for admissibility

The following theorem is the paper's central cautionary tale for ML: a predictor that passes every calibration test under its own measure can still be strictly dominated by a Bayes-regularized alternative.

Theorem 4.3 (Martingale necessary, not sufficient)**.** *In the Bernoulli model $X_i \overset{\text{iid}}{\sim} \mathrm{Bern}(\theta)$, $\theta \in (0, 1)$, under log loss $L(\theta, p) = -\theta \log p - (1 - \theta) \log(1 - p)$:*

(i) Every Bayesian posterior predictive sequence (m_n) is a martingale under the prior predictive measure (Proposition 4.2).

(ii) The plug-in rule $\hat{p}_n^{\mathrm{pi}} = S_n/n$ satisfies the martingale condition under its own predictive measure (self-consistency) $\hat{P}$ (where $X_t \mid X_{1:t-1} \sim \mathrm{Bern}(\hat{p}_{t-1}^{\mathrm{pi}})$).

(iii) $\hat{p}_n^{\mathrm{pi}}$ is strictly dominated by the Bayes rule $\hat{p}_n^B = (S_n + \frac{1}{2})/(n+1)$ under $\mathrm{Beta}(\frac{1}{2}, \frac{1}{2})$ prior, for every $n \geq 1$ and $\theta \in (0,1)$.

Hence $r(\hat{p}_n^{\mathrm{pi}}) \notin \partial_- \mathcal{R}$: martingale coherence is necessary but not sufficient for no-shame.

Intuition and algorithmic relevance. A predictor that is self-consistent (a martingale under its own predictive measure) need not be admissible under the true data-generating process. The plug-in MLE is perfectly calibrated under $\hat{P}$ yet assigns probability zero to realizable events, producing infinite KL divergence under every P_θ. This gap between calibration and admissibility is directly relevant to the evaluation of probabilistic forecasters and LLM calibration (Gneiting & Raftery, 2007): a model that "looks calibrated" by its own metric may still be dominated by a Bayes-regularized alternative. In the language of deep learning, this is the gap between train-set calibration and test-set optimality; label smoothing, temperature scaling, and Bayesian ensembling are all instances of the Bayes correction $\hat{p}_n^B$ that avoids boundary predictions. Section 8 demonstrates this gap quantitatively in finite samples.

Proof. See Appendix A.1.

Remark 4.4 (Role of extended-real risk). The dominance in part (iii) requires $+\infty$ risk (Definitions 2.1 and 2.3). Extended-real risk is essential: bounded losses exclude proper scoring rules such as log loss.

5 Criterion Separation

Sections 3–4 established Blackwell admissibility as the first geometry. We now introduce two additional admissibility criteria, anytime-valid sequential inference and marginal coverage validity, each operating on a different space of procedures with a different partial order, and prove that the three resulting classes are pairwise non-nested.

5.1 Anytime-valid admissibility

Definition 5.1 (Anytime-valid constraint class)**.** Let $\mathcal{H}_0$ be a composite null. The *anytime-valid class* is

$$\mathcal{C}_{\mathrm{AV}} = \big\{(E_t)_{t\geq 1} : E_t \geq 0,\ \sup_{\mathbb{P}\in\mathcal{H}_0} \mathbb{E}_{\mathbb{P}}[E_\tau] \leq 1 \text{ for every stopping time } \tau\big\}.$$

Elements of $\mathcal{C}_{\mathrm{AV}}$ are called *e-processes*. By Ville's inequality (Ville, 1939; Howard et al., 2021), every $E \in \mathcal{C}_{\mathrm{AV}}$ provides anytime-valid type-I error control at level α.

For ML practitioners, e-processes enable safe sequential model comparison: one can peek at A/B test results after every batch without inflating false discovery rates. The nonnegative supermartingale condition is the structural price of this guarantee.

Theorem 5.2 (Ramdas et al. (Ramdas et al., 2022))**.** *Within $\mathcal{C}_{\mathrm{AV}}$, a procedure is admissible (in the sense that no other e-process has uniformly larger stopped expectation) if and only if it is a nonnegative martingale under every $\mathbb{P} \in \mathcal{H}_0$.*

Remark 5.3. The partial order in Theorem 5.2 compares e-processes by their stopped expectations under every stopping time τ (equivalently, by type-I error control via Ville's inequality); "admissible" is used relative to this induced order on $\mathcal{C}_{\mathrm{AV}}$, not the coordinatewise risk order of Definition 2.4. See also Shafer et al. (2011) for the test martingale perspective. This is distinct from Blackwell admissibility, which requires no domination under a loss $L(\theta, \delta)$ over all of Θ.

Proposition 5.4 (Martingale as structural bridge)**.** *The martingale property relates to each admissibility criterion as follows:*

(i) Bayes $\Rightarrow$ martingale: every posterior predictive sequence is a martingale (Proposition 4.2).

(ii) AV-admissible $\Leftrightarrow$ nonnegative martingale within $\mathcal{C}_{\mathrm{AV}}$ (Theorem 5.2).

(iii) Coverage validity does not require the martingale property: conformal prediction sets are constructed from rank statistics.

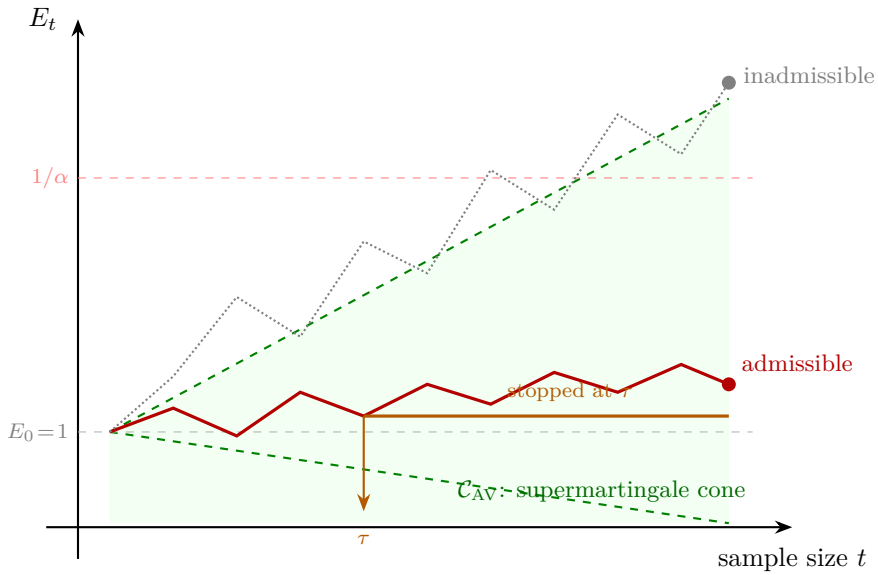

Figure 3: Supermartingale cone for anytime-valid inference. An e-process E_t starts at $E_0 = 1$ and must remain a nonnegative supermartingale under every $\mathbb{P} \in \mathcal{H}_0$; this defines the feasible cone $\mathcal{C}_{\mathrm{AV}}$. An admissible e-process (solid, red) is a *nonnegative martingale* within the cone (Theorem 5.2). A process that grows systematically faster (dotted, gray) violates the supermartingale condition under some $\mathbb{P} \in \mathcal{H}_0$ and is inadmissible. Stopping at any data-dependent time τ preserves type-I error control at level α via Ville's inequality: the stopped value $E_\tau \leq 1/\alpha$ with probability at least $1 - \alpha$ (orange).

(iv) Martingale $\not\Rightarrow$ Blackwell admissible: the plug-in $\hat{p}_n^{\mathrm{pi}}$ satisfies the martingale condition but $r(\hat{p}_n^{\mathrm{pi}}) \notin \partial_- \mathcal{R}$ (Theorem 4.3).

Martingale coherence is a necessary condition in some admissibility geometries and a complete characterization in others, but it is not a universal determinant of admissibility.

5.2 Marginal coverage admissibility

The third admissibility geometry concerns prediction sets. *Marginal coverage* averages over both calibration data and the test point; *conditional coverage* conditions on $X_{n+1} = x$ and demands coverage at every x. Marginal coverage is achievable by conformal methods; conditional coverage is not. Conformal prediction has become the default uncertainty quantification wrapper in production ML: it provides distribution-free coverage guarantees for any black-box model, requiring only exchangeability of the calibration and test data.

Definition 5.5 (Marginal coverage)**.** Given an exchangeable sequence $(X_1, \ldots, X_n, X_{n+1})$, a prediction set $\hat{C}_n(X_{n+1})$ satisfies *marginal coverage at level* $1 - \alpha$ if

$$\mathbb{P}\big(Y_{n+1} \in \hat{C}_n(X_{n+1})\big) \geq 1 - \alpha.$$

Theorem 5.6 (Foygel Barber et al. (Barber et al., 2021))**.** *Any method satisfying exact conditional coverage $\mathbb{P}(Y_{n+1} \in \hat{C}_n(X_{n+1}) \mid X_{n+1} = x) = 1 - \alpha$ for every x must produce prediction sets of infinite expected length at every non-atom of the marginal of X_{n+1}, for any continuous distribution.*

Insisting on both Blackwell optimality and conditional coverage simultaneously is impossible for geometric reasons: sharp point predictions or broad coverage sets, but not both. This is a concrete instance of the criterion separation that Theorem 5.9 formalizes. The genealogy of conformal prediction traces back to the Fisher–Dempster–Hill predictive inference tradition, which is structurally distinct from the Bayesian (Blackwell) tradition: conformal sets are calibrated via rank statistics over exchangeable data, not via posterior integration over a prior. Recent work makes this distinction precise, showing that conformal and Bayesian

prediction agree only under stringent conditions; in general, the rank-calibration mechanism underlying conformal coverage and the prior-weighted loss minimization underlying Blackwell admissibility produce genuinely different procedures, grounding the pairwise non-nesting of $\mathfrak{B}$ and $\mathfrak{C}$ in a deeper methodological divergence.

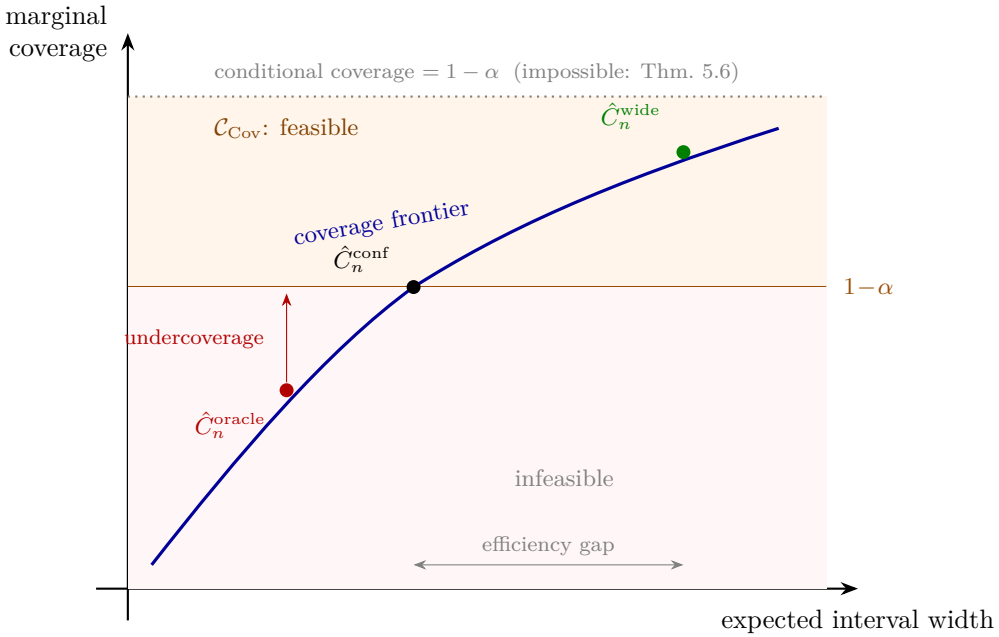

Figure 4: Coverage-feasible region for prediction sets. The feasibility constraint $\mathbb{P}(Y_{n+1} \in \hat{C}_n) \geq 1-\alpha$ defines the half-space $\mathcal{C}_{\mathrm{Cov}}$ (shaded, above the threshold). The conformal set $\hat{C}_n^{\mathrm{conf}}$ lies on the coverage frontier: it achieves exactly $1-\alpha$ marginal coverage; it represents a representative exchangeability-based procedure attaining exact marginal coverage. An oracle Bayes interval $\hat{C}_n^{\mathrm{oracle}}$ optimized under the true P_θ can be shorter but may undercover; it lies below the threshold and is infeasible in $\mathcal{C}_{\mathrm{Cov}}$. A conservative set $\hat{C}_n^{\mathrm{wide}}$ overcovers but wastes width. Exact conditional coverage at every x simultaneously is impossible for continuous distributions (Theorem 5.6).

5.3 Criterion-relative admissibility

Definition 5.7 (Criterion-relative admissibility)**.** Write $\mathsf{Adm}_{\mathcal{C}}(\delta)$ to denote that δ is admissible relative to criterion $\mathcal{C}$: a comparison class of procedures together with the partial order induced by the relevant performance functional. The three criteria considered are:

(i) $\mathcal{C}_B$: Blackwell risk dominance: comparison class $\mathcal{D}$, ordering by coordinatewise risk dominance on $\mathcal{R}$, ambient geometry the convex risk set $\mathcal{R} \subset \mathbb{R}^k_+$.

(ii) $\mathcal{C}_{\mathrm{AV}}$: anytime-valid admissibility: comparison class $\mathcal{C}_{\mathrm{AV}}$ (Definition 5.1), ordering by expected stopped value, ambient geometry the cone of nonnegative supermartingales.

(iii) $\mathcal{C}_{\mathrm{Cov}}$: marginal coverage validity: comparison class the set of prediction sets under exchangeability, ordering by coverage level, ambient geometry the coverage-feasible region $\{\hat{C} : \mathbb{P}(Y_{n+1} \in \hat{C}) \geq 1-\alpha\}$.

In this notation, $\mathfrak{B} = \{\delta : \mathsf{Adm}_{\mathcal{C}_B}(\delta)\}$, $\mathfrak{A} = \{\delta : \mathsf{Adm}_{\mathcal{C}_{\mathrm{AV}}}(\delta)\}$, $\mathfrak{C} = \{\delta : \mathsf{Adm}_{\mathcal{C}_{\mathrm{Cov}}}(\delta)\}$.

5.4 Constrained Bayes as a design principle

Definition 5.7 makes precise the sense in which each admissibility criterion operates on its own space of procedures and its own partial order. We now observe that every criterion in Definition 5.7 can be viewed through a common schematic template in which Bayesian risk is the objective and the validity requirement is a feasibility constraint. Because the four criteria act on different object spaces ($\mathcal{D}$ for point predictors,

e-processes for sequential testing, prediction sets for coverage, sequential strategies for CAA), the template below is a *schema* rather than a single literal optimization problem: in each instantiation the decision space, feasible set, and notion of dominance must be matched to the criterion at hand.

Definition 5.8 (Constrained Bayes schema)**.** Given a statistical decision problem $(\Theta, \mathcal{A}, L, \mathcal{X}, \mathcal{P})$ in the sense of Definition 2.1, a prior Π on Θ, and a *criterion-specific decision space* $\mathcal{D}_C$ equipped with a *feasible set* $\mathcal{F} \subseteq \mathcal{D}_C$, the *constrained Bayes problem* is

$$\min_{\delta \in \mathcal{D}_C} \int_\Theta R(\theta, \delta)\ \mathrm{d}\Pi(\theta) \qquad \text{subject to} \qquad \delta \in \mathcal{F}. \tag{1}$$

A solution $\delta^*_{\mathcal{F}}$ is a *constrained Bayes rule*. When $\mathcal{D}_C = \mathcal{D}$ and $\mathcal{F} = \mathcal{D}$, the schema reduces to the unconstrained Bayes problem and $\delta^*_{\mathcal{D}} = \delta_\Pi$ is the standard Bayes rule of Theorem 3.8. For other criteria the decision space $\mathcal{D}_C$ may consist of e-processes, prediction sets, or sequential strategies, so the risk functional R and the notion of dominance are adapted accordingly.

The four admissibility geometries of this paper correspond to four choices of $\mathcal{F}$. Under Blackwell admissibility ($\mathcal{F} = \mathcal{D}$), no constraint is imposed; by Theorem 3.8 and Corollary 3.13, every solution lies on the lower boundary $\partial_-\mathcal{R}$ and the feasible risk set is $\mathcal{R}$ itself. For anytime-valid inference ($\mathcal{F} = \mathcal{C}_{\mathrm{AV}}$), the feasibility constraint requires $\sup_{\mathbb{P} \in \mathcal{H}_0} \mathbb{E}_{\mathbb{P}}[E_\tau] \leq 1$ for every stopping time τ (Definition 5.1); admissibility within $\mathcal{C}_{\mathrm{AV}}$ then reduces to the nonnegative martingale property (Theorem 5.2), so the martingale condition is a constraint-induced structural requirement, not an alternative optimality principle. For marginal coverage ($\mathcal{F} = \mathcal{C}_{\mathrm{Cov}} = \{\hat{C} : \mathbb{P}(Y_{n+1} \in \hat{C}_n(X_{n+1})) \geq 1 - \alpha\}$), the feasibility constraint requires exchangeability-based coverage (Definition 5.5) and restricts procedures to prediction sets rather than point predictions; the conformal guarantee is a feasibility condition on the output space, not a competing loss criterion. For CAA-admissibility ($\mathcal{F} = \mathcal{C}_{\mathrm{CAA}} = \{\delta : \bar{R}_n(\theta, \delta) \to \partial_-\mathcal{R} \text{ for all } \theta\}$), the feasibility constraint requires Cesàro convergence of the time-averaged risk to the lower boundary (Definition 6.6); the constrained Bayes problem then asks which among all boundary-converging strategies minimizes Bayesian integrated risk. The prior Π still serves as the optimization objective, but the strategy need not be Bayes at any finite round: it suffices that the cumulative record reaches $\partial_-\mathcal{R}$ in the limit.

In (1) the Bayesian integrated risk is always the objective; the feasible set $\mathcal{F}$ encodes whatever structural or validity requirement the analyst imposes. The prior Π retains the dual role established in Theorem 3.8: it is the normal to the supporting hyperplane at an admissible risk point. Validity requirements restrict the region of $\mathcal{R}$ over which that hyperplane is optimized, but they do not alter the objective itself.

The resulting admissible class is the lower boundary of the restricted risk set $\mathcal{R}_{\mathcal{F}} \subseteq \mathcal{R}$; Bayes remains the objective, the constraint $\mathcal{F}$ determines the accessible frontier.

For an ML practitioner, the constrained Bayes template provides a recipe: (1) choose a validity requirement (coverage, anytime-validity, calibration); (2) encode it as a feasible set $\mathcal{F}$ within the appropriate decision space $\mathcal{D}_C$; (3) optimize Bayesian risk within $\mathcal{F}$. Under regularity conditions ensuring that the constrained optimum exists (compactness of $\mathcal{F}$, lower semicontinuity of risk), the resulting procedure is admissible relative to the chosen criterion. Table 1 maps four common ML tasks to the corresponding constrained Bayes instantiation.

Table 1: Constrained Bayes recipe for four ML tasks.

ML task	Feasibility $\mathcal{F}$	Resulting algorithm	Certificate
Point prediction	$\mathcal{D}$ (unconstrained)	Bayes posterior predictive	prior Π
Sequential testing	$\mathcal{C}_{\mathrm{AV}}$	Bayes-optimal e-process	supermartingale
Uncertainty quantification	$\mathcal{C}_{\mathrm{Cov}}$	Conformal Bayes set	exch. rank
Online calibration	$\mathcal{C}_{\mathrm{CAA}}$	Constrained defensive forecast	fixed-point

This formulation makes the criterion-separation theorems below natural. The four problems share a common schematic objective (minimize Bayesian risk), but the constraint sets $\mathcal{D}$, $\mathcal{C}_{\mathrm{AV}}$, $\mathcal{C}_{\mathrm{Cov}}$, $\mathcal{C}_{\mathrm{CAA}}$ are defined on different spaces of objects, induce different partial orders, and admit no common refinement. No single

procedure can simultaneously be unconstrained-optimal on $\partial_- \mathcal{R}$, anytime-valid feasible in $\mathcal{C}_{\mathrm{AV}}$, coverage-feasible in $\mathcal{C}_{\mathrm{Cov}}$, and Cesàro-convergent in $\mathcal{C}_{\mathrm{CAA}}$, because the induced partial orders and feasibility constraints admit no common refinement, even when one views procedures in a common meta-class. The non-nesting of admissible classes is a consequence of incompatible feasibility sets, not of philosophical disagreement about what optimality means.

5.5 Separation theorem

Theorem 5.9 (Criterion separation)**.** *Let $\mathfrak{B}$, $\mathfrak{A}$, $\mathfrak{C}$ denote the classes of procedures that are Blackwell admissible, anytime-valid admissible, and marginal-coverage valid, respectively. Then:*

(i) $\mathfrak{B} \not\subseteq \mathfrak{A}$, $\mathfrak{A} \not\subseteq \mathfrak{B}$;
(ii) $\mathfrak{B} \not\subseteq \mathfrak{C}$, $\mathfrak{C} \not\subseteq \mathfrak{B}$;
(iii) $\mathfrak{A} \not\subseteq \mathfrak{C}$, $\mathfrak{C} \not\subseteq \mathfrak{A}$.

The non-nestedness is structural: each criterion $\mathcal{C}_B$, $\mathcal{C}_{\mathrm{AV}}$, $\mathcal{C}_{\mathrm{Cov}}$ induces a different partial order on a different space of objects (Definition 5.7).

Intuition and algorithmic relevance. The three evaluation paradigms (decision-theoretic risk (Wald, 1950; Blackwell & Girshick, 1954), anytime-valid testing via e-values (Ramdas et al., 2023; Shafer et al., 2011), and conformal coverage (Vovk et al., 2005a; Angelopoulos & Bates, 2023)) operate on genuinely different spaces of procedures with different partial orders. A Bayes-optimal point predictor does not produce prediction sets; an e-process does not optimize a proper scoring rule; a conformal set does not minimize any loss. The separation is therefore structural, not a matter of approximation, and persists for any learning problem admitting all three evaluation tasks. For a concrete ML implication: if you optimize a model for calibration (Blackwell risk) but evaluate it on coverage (conformal), the rankings can reverse: a well-calibrated model may have worse coverage than a poorly-calibrated one.

Proof. See Appendix A.2.

Interpretation. No common refinement of the three partial orders exists: the incompatibility arises from the geometry of the constraint sets, not from philosophical disagreement, and persists for any learning problem admitting all three evaluation tasks.

6 Constructive and Cesàro Approachability Admissibility

The connection between Blackwell approachability and no-regret learning (Abernethy et al., 2011) invites a question: does steering the time-averaged risk to the lower boundary suffice for admissibility? This section distinguishes two routes to $\partial_- \mathcal{R}$ and shows that the distinction generates a fourth admissibility geometry.

6.1 Two paths to the boundary

Definition 6.1 (Constructive admissibility)**.** A procedure δ is *constructively admissible* if there exists a prior Π_n at every sample size n such that $\delta(X^n) = \delta_{\Pi_n}(X^n)$, i.e., δ is Bayes with respect to an explicitly specified, sample-size-dependent prior at each round.

This is a pointwise (per-round) boundary condition: each action is itself boundary-certified by the supporting hyperplane of its prior (Corollary 3.13).

Definition 6.2 (Cesàro admissibility)**.** A procedure δ is *Cesàro admissible* if the time-averaged risk $\bar{R}_n(\theta, \delta) = n^{-1} \sum_{t=1}^{n} R(\theta, \delta_t)$ converges to $\partial_- \mathcal{R}$ as $n \to \infty$, without requiring that each individual action δ_t be Bayes optimal at time t.

This is a Cesàro boundary condition: only the time-average risk approaches $\partial_- \mathcal{R}$, so individual rounds may lie in the interior of the risk set.

Constructive admissibility demands a per-round witness (the prior Π_n), while Cesàro admissibility requires only that the long-run average reaches the boundary. The Blackwell approachability theorem guarantees the existence of Cesàro-admissible strategies whenever the target set is approachable; it does not, in general, produce constructively admissible ones.

6.2 Approachability revisited: the missing martingale layer

Abernethy et al. (2011) show that Blackwell approachability and no-regret learning are equivalent. If $S = \mathcal{R}$ is the risk set, approachability of $\partial_- \mathcal{R}$ guarantees $\bar{R}_n \to \partial_- \mathcal{R}$, but the equivalence operates in the Cesàro regime: it does not require that each per-round action δ_t be individually Bayes optimal. The martingale layer (Section 4) provides the missing intertemporal constraint: the prior sequence (Π_t) must update coherently via Bayes' rule, not merely converge in Cesàro average. Approachability ensures arrival at the boundary; the martingale property ensures the journey is coherent. Constructive admissibility requires a per-round Bayes witness whose predictions are time-consistent across rounds; the martingale property is the observable footprint of that consistency.

Foster–Vohra calibration (Foster & Vohra, 1998; Hart & Mas-Colell, 2001) sits naturally in the CAA geometry (Definition 6.6): calibration error vanishes in the Cesàro sense via a fixed-point argument, with no per-round Bayes witness required.

6.3 Cesàro does not imply pointwise

Proposition 6.3. *There exists a procedure that is Cesàro admissible but not constructively admissible. In particular, the defensive forecaster (P5 of Section 7.1) achieves $\bar{R}_n(\theta) \to \partial_- \mathcal{R}$ for every θ but is not Bayes with respect to any prior at any finite sample size.*

Intuition and algorithmic relevance. Defensive forecasting (Vovk et al., 2005b) and the Foster–Vohra calibration algorithm (Foster & Vohra, 1998) achieve long-run (Cesàro) calibration through fixed-point existence arguments, without requiring a prior at any round. In online learning terms (Cesa-Bianchi & Lugosi, 2006), the time-averaged risk converges to the efficient frontier, but no individual prediction is Bayes-optimal. This is the algorithmic distinction between calibration dynamics and Bayesian updating.

Proof. See Appendix A.5.

6.4 Martingale coherence of constructive admissibility

Proposition 6.4. *If a procedure δ is constructively admissible, then its posterior predictive sequence $(\hat{p}_n^{\Pi_n})_{n \geq 1}$ forms a martingale under the prior predictive measure.*

Remark 6.5. Section 4 establishes the martingale property under the prior predictive measure (Bayesian model). Extending the conclusion to "for every P_θ" would require additional regularity conditions (e.g., absolute continuity of P_θ with respect to the prior predictive) that we do not pursue here.

Intuition and algorithmic relevance. The martingale property captures Bayesian coherence across time (Doob, 1949): predictions at time n must equal the conditional expectation of predictions at time $n + 1$. Constructive admissibility implies this time-consistency under the Bayesian model, distinguishing it from Cesàro admissibility, which requires only long-run convergence. In ML terms, this is the distinction between a prediction algorithm that is sequentially coherent (each update is Bayes-justified) and one that merely reaches the right answer in the long run.

Proof. See Appendix A.4.

Proposition 6.4 clarifies the sense in which the martingale layer separates constructive from Cesàro admissibility: the former requires the journey to be a martingale under the Bayesian model, while the latter only requires the destination.

6.5 A fourth geometry: Cesàro approachability admissibility

Definition 6.6 (CAA-admissibility)**.** A procedure δ is *CAA-admissible* (Cesàro approachability admissible) if the time-averaged risk satisfies $\bar{R}_n(\theta, \delta) \to \partial_- \mathcal{R}$ for every $\theta \in \Theta$, where convergence is achieved by an approachability-style strategy whose existence is guaranteed by a fixed-point or minimax argument, without requiring an explicit prior witness at each round.

In online learning, CAA-admissibility captures the long-run calibration guarantees of algorithms such as Follow-the-Regularized-Leader and defensive forecasting: the time-averaged loss converges to the Pareto frontier, but no individual prediction need be Bayes-optimal.

The paradigmatic CAA-admissible procedure is the defensive forecaster (Vovk et al., 2005b): it achieves calibration by an approachability-style fixed-point argument rather than Bayesian updating. The asymptotic calibration results of Foster & Vohra (1998) and the adaptive strategies of Hart & Mas-Colell (2001) provide further examples.

Theorem 6.7 (Extended separation)**.** *Let $\mathfrak{B}$, $\mathfrak{A}$, $\mathfrak{C}$, $\mathfrak{D}$ denote the classes of procedures that are Blackwell admissible, anytime-valid admissible, marginal-coverage valid, and CAA-admissible, respectively. Then the four classes are pairwise non-nested: for each pair $(\mathfrak{X}, \mathfrak{Y})$ with $\mathfrak{X} \neq \mathfrak{Y}$, there exist procedures in $\mathfrak{X} \setminus \mathfrak{Y}$ and in $\mathfrak{Y} \setminus \mathfrak{X}$.*

Intuition and algorithmic relevance. The fourth geometry (CAA/approachability) adds online calibration and defensive forecasting (Vovk et al., 2005b; Abernethy et al., 2011) to the taxonomy. A defensive forecaster steers time-averaged risk to the boundary via a fixed-point argument but is not Bayes at any finite round, does not produce an e-process, and does not yield prediction sets. Conversely, Bayesian, e-process, and conformal procedures each fail the Cesàro convergence requirement for structural reasons. The four geometries are therefore pairwise non-nested: no single evaluation metric governs all four (Cesa-Bianchi & Lugosi, 2006; Rakhlin & Sridharan, 2013). In practical terms, there is no "master leaderboard" that ranks ML algorithms simultaneously across risk, sequential validity, coverage, and calibration; the four evaluation metrics are genuinely incompatible.

Proof. See Appendix A.3.

6.6 Interpretation: four geometries and moral pluralism

The extended separation theorem deepens the pluralism described in the Introduction. Four distinct admissibility geometries now correspond to four conceptions of statistical virtue:

(i) *Blackwell*: the rule is optimal for a declared objective (prior-witnessed, per-round);
(ii) *Anytime-valid*: the rule controls error at every stopping time (martingale-witnessed);
(iii) *Coverage*: the rule guarantees marginal containment (exchangeability-witnessed);
(iv) *CAA*: the rule reaches the boundary in the long run (fixed-point-witnessed, limiting).

In Williams's terms (Williams, 1993), each geometry defines its own standard of shame; a procedure shameless under one standard may be indefensible under another.[1]

Table 2 summarizes the four geometries.

The four-geometry diamond (Figure 5) illustrates the pairwise non-nesting. No arrow connects any pair, confirming that no common refinement exists across all four frameworks.

7 Bernoulli Laboratory

The separation theorem is proved constructively using four procedures in the Bernoulli model. This section collects the classification results in a single table and schematic, making the pairwise non-nesting visually explicit.

[1] CAA-admissibility corresponds to Berlin's notion of negative liberty (Berlin, 1969): no persistent deficiency survives in the limit, though no per-round prior is required.

Table 2: Taxonomy of four admissibility geometries.

Geometry	Certificate	Boundary witness	Optimality mode
$\mathfrak{B}$: Blackwell	supporting hyperplane	prior Π	pointwise (per-round)
$\mathfrak{A}$: Anytime-valid	supermartingale	e-process	pathwise (all stopping times)
$\mathfrak{C}$: Coverage	feasibility region	conformal rank	marginal (exchangeability)
$\mathfrak{D}$: CAA	Cesàro steering	fixed-point / minimax	Cesàro (time-averaged)

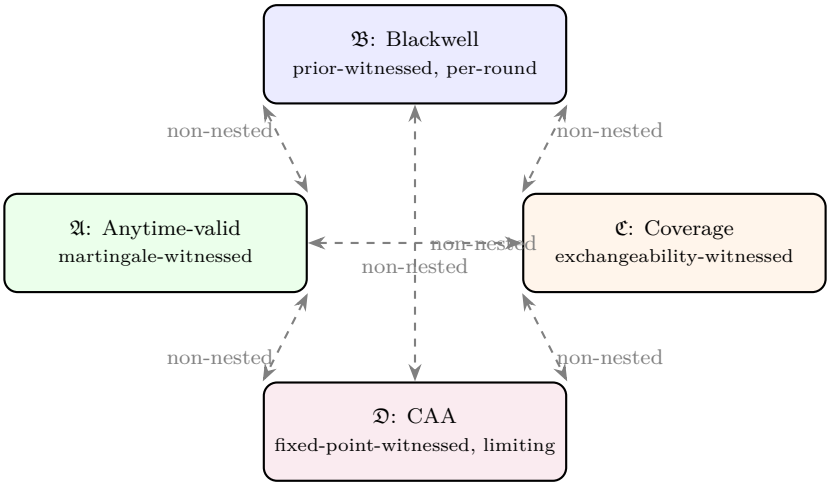

Figure 5: Four admissibility geometries in diamond configuration. Each node represents an admissible class; dashed arrows indicate pairwise non-nesting (Theorems 5.9 and 6.7). Blackwell and CAA admissibility share the risk-set domain but differ in witness type (prior vs. fixed-point); anytime-valid and coverage admissibility operate on different procedure spaces entirely.

Table 4 records the four procedures and four admissibility criteria. N/A indicates that the criterion is defined for a categorically different class of procedures; the assessment is inapplicable rather than negative.

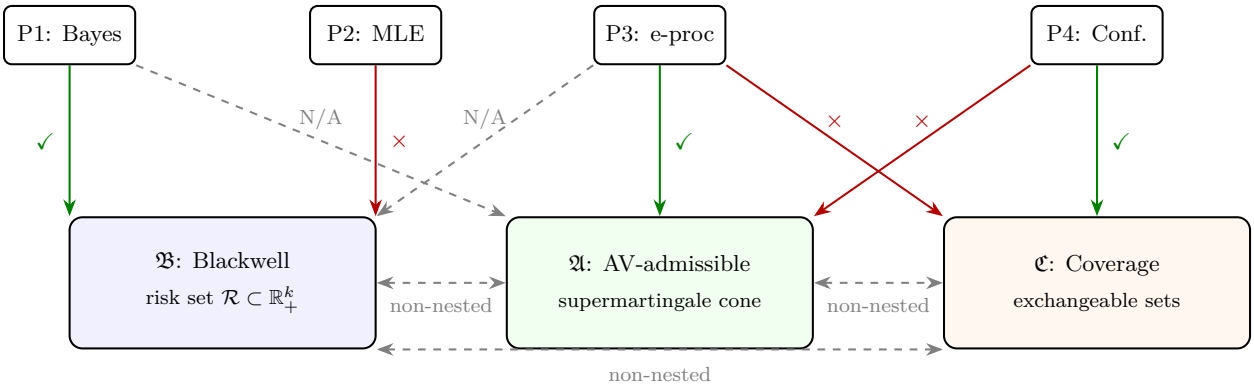

Figure 6: Bernoulli laboratory: procedures P1–P4 mapped to the three admissibility geometries. Green arrows (✓) indicate membership; red arrows (×) indicate failure; dashed gray arrows (N/A) indicate structural inapplicability. The three geometry classes $\mathfrak{B}$, $\mathfrak{A}$, $\mathfrak{C}$ are pairwise non-nested (Theorem 5.9).

Corollary 7.1 (Martingale property is not criterion-determining)**.** *P1, P2, P3 all satisfy the martingale property and belong to different admissibility classes. The martingale property does not determine membership in* $\mathfrak{B}$*,* $\mathfrak{A}$*, or* $\mathfrak{C}$*.*

Table 3: Three-layer anatomy of the four admissibility geometries. Each criterion evaluates a different type of procedure (Object), ranks it via a different partial order (Order), and certifies optimality through a different mathematical witness (Certificate). No common refinement of these partial orders exists (Theorem 5.9).

Geometry	**Object**	**Order**	**Certificate**
Blackwell	point predictor	risk dominance	supporting-hyperplane prior
Anytime-valid	test process	supermartingale ordering	nonneg. supermartingale
Coverage	prediction set	coverage ordering	exchangeability rank
CAA	sequential strategy	Cesàro ordering	approachability fixed point

Table 4: Bernoulli laboratory: four procedures and four admissibility geometries. ✓ = satisfies; × = fails; ? = not established; N/A = criterion not applicable to this procedure type. The CAA column refers to Cesàro approachability admissibility (Definition 6.6). ‡Under the i.i.d. model, P1 and P2 converge to the boundary; under adversarial sequences the convergence is not guaranteed (see Appendix A.3).

Procedure	Mg. prop.	Blackwell adm.	AV-adm.	Cov. valid	CAA-adm.
P1: Bayes (Beta($\frac{1}{2}, \frac{1}{2}$))	✓	✓	N/A	N/A	?‡
P2: Plug-in MLE S_n/n	✓	×	N/A	N/A	?‡
P3: LR e-process	✓	N/A	✓	×	N/A
P4: Conformal prediction set	×	N/A	×	✓	N/A

Corollary 7.2 (No universal procedure)**.** *No procedure is admissible under all four criteria simultaneously. The criteria are defined for different procedure types and are pairwise non-nested by Theorems 5.9 and 6.7.*

7.1 Extended procedures

Two additional procedures sharpen the separation and connect to the constructive and Cesàro approachability framework of Section 6.

P5. *Defensive forecasting* (Vovk et al., 2005b): at each step t, choose $\hat{p}_t$ so that $|\bar{p}_n - \bar{X}_n| \to 0$ almost surely (Cesàro calibration, no model assumed).

P6. *Constrained Bayes prediction set*: define $\hat{C}_n^{CB} = \{y : |y - \hat{p}_n^B| \leq q_{1-\alpha}\}$ where $q_{1-\alpha}$ is the posterior predictive $(1-\alpha)$-quantile; an instance of constrained Bayes (Definition 5.8) with $\mathcal{F} = \mathcal{C}_{\text{Cov}}$.

The four admissibility geometries correspond to four distinct notions of optimality used across ML subfields. The *Blackwell geometry* (risk dominance) governs Bayesian predictive optimization: an algorithm is admissible if no alternative achieves uniformly lower risk, and the certificate of optimality is a supporting-hyperplane prior. The *anytime-valid geometry* (martingale validity / e-processes) governs sequential testing and safe inference: admissibility within the e-process class requires the nonnegative martingale property (Shafer et al., 2011; Howard et al., 2021). The *coverage geometry* (conformal prediction guarantees) governs distribution-free uncertainty quantification: a prediction set is admissible if it achieves marginal coverage at level $1-\alpha$ under exchangeability (Angelopoulos & Bates, 2023). The *CAA geometry* (online calibration and defensive forecasting) governs long-run calibration: a strategy is admissible if its time-averaged risk converges to the lower boundary of the risk set via a fixed-point argument (Cesa-Bianchi & Lugosi, 2006; Rakhlin & Sridharan, 2013). Because these four geometries induce different partial orders on different spaces of procedures, no single evaluation metric can govern all four; this is the content of the separation theorems.

Table 5 summarizes membership across the four admissibility geometries for all six procedures.

P5 is CAA-admissible (Section 6.5) but fails all three original criteria, confirming that Cesàro calibration is strictly weaker than Blackwell, AV-, or coverage admissibility. P6 demonstrates the constrained Bayes

Table 5: Extended Bernoulli laboratory: six predictive algorithms, five admissibility criteria, and witness type. The CAA column refers to Cesàro approachability admissibility (Definition 6.6). Witness type classifies whether admissibility is justified by an explicit prior (constructive) or a fixed-point/limiting argument (Cesàro).

Procedure	Mg. prop.	Blackwell adm.	AV-adm.	Cov. valid	CAA-adm.	Constr. adm.	Witness type
P1: Bayes	✓	✓	N/A	N/A	?‡	✓	prior
P2: Plug-in MLE	✓	×	N/A	N/A	?‡	×	none
P3: LR e-proc	✓	N/A	✓	×	N/A	N/A	martingale
P4: Conformal	×	N/A	×	✓	N/A	N/A	exchangeability
P5: Defensive	×	×	×	×	✓	×	fixed-point
P6: Constr. Bayes	✓	×†	N/A	✓	✓	×†	prior + exch.

†P6 is not Blackwell admissible as a point predictor (the coverage constraint restricts the feasible set), but it is admissible within the restricted risk set $\mathcal{R}_{\mathcal{F}}$ for $\mathcal{F} = \mathcal{C}_{\mathrm{Cov}}$.

trade-off: coverage feasibility costs Blackwell admissibility in the unconstrained sense but gains membership in $\mathcal{C}_{\mathrm{Cov}}$. Each witness procedure has a natural ML analogue: P1 corresponds to a Bayesian neural network with a proper prior; P3 to a sequential A/B test accumulating evidence via likelihood ratios; P4 to a conformal wrapper applied to a black-box classifier; and P5 to an online calibration algorithm such as Follow-the-Regularized-Leader.

7.2 Gaussian laboratory

The Bernoulli laboratory demonstrates separation in the simplest discrete model. We now verify that the same structural phenomena persist in the Gaussian location model $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ with known σ^2 and $\mu \in \mathbb{R}$.

Under squared loss $L(\mu, a) = (\mu - a)^2$, the risk of the sample mean $\bar{X}_n$ is $R(\mu, \bar{X}_n) = \sigma^2/n$ for all μ, making it minimax and admissible (every linear estimator with constant risk is admissible in the Gaussian location family). The Bayes predictive under a $N(\mu_0, \tau^2)$ prior is

$$\hat{\mu}_n^B = \frac{n\bar{X}_n/\sigma^2 + \mu_0/\tau^2}{n/\sigma^2 + 1/\tau^2} = w_n \bar{X}_n + (1 - w_n)\mu_0, \qquad w_n = \frac{n\tau^2}{n\tau^2 + \sigma^2},$$

which is also admissible. Unlike the Bernoulli case, neither $\bar{X}_n$ nor $\hat{\mu}_n^B$ suffers infinite risk (the squared loss is bounded on compact subsets), but the separation still holds.

Proposition 7.3 (Gaussian separation)**.** *In the Gaussian location model under squared loss:*

(i) The sample mean $\bar{X}_n$ is Blackwell admissible but does not define an e-process and does not produce a prediction set.

(ii) The likelihood-ratio e-process $E_n = \prod_{t=1}^n f_{\hat{\mu}_{t-1}}(X_t)/f_{\mu_0}(X_t)$ is AV-admissible (a nonnegative martingale under $H_0 : \mu = \mu_0$) but does not optimize squared loss.

(iii) The conformal prediction interval $\hat{C}_n = [\bar{X}_n \pm \hat{q}_{1-\alpha}]$ achieves marginal coverage but is not Blackwell admissible as a point predictor and is not an e-process.

The three classes $\mathfrak{B}$, $\mathfrak{A}$, $\mathfrak{E}$ are pairwise non-nested in the Gaussian model, just as in the Bernoulli model (Theorem 5.9).

In finite samples the Bayes shrinkage weight $w_n = n\tau^2/(n\tau^2 + \sigma^2)$ visibly reduces risk under squared loss for small n, mirroring the log-loss behavior documented in Table 6.

The Gaussian laboratory also reveals a structural asymmetry absent from the Bernoulli case: in the Gaussian model, $\bar{X}_n$ is both minimax and admissible under squared loss, whereas the Bernoulli plug-in S_n/n is minimax under bounded loss but inadmissible under log loss. The pathology (infinite risk at the boundary) is specific

to proper scoring rules on discrete sample spaces, where the plug-in assigns zero probability to realizable events.

8 Experiments and Illustrations

The preceding sections established the separation theorem analytically. We now demonstrate each geometry in finite samples, confirming that the theoretical distinctions are visible in practice. These experiments are designed to make the separation theorem tangible for practitioners: each targets one admissibility geometry and uses sample sizes and parameters typical of ML evaluation pipelines. All use $B = 10{,}000$ replications.

8.1 Bayes vs. plug-in under log loss

We draw $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bern}(0.3)$ and compare the next-step log loss $L(\theta, p) = -\theta \log p - (1-\theta)\log(1-p)$ for the Bayes predictive $\hat{p}_n^B = (S_n + \frac{1}{2})/(n+1)$ under $\text{Beta}(\frac{1}{2}, \frac{1}{2})$ and the plug-in MLE $\hat{p}_n^{\text{pi}} = S_n/n$. Table 6 records the average risk over B replications. At small n the plug-in predictor hits the boundary $\{0, 1\}$ with positive probability, producing infinite log loss on those realizations; the Bayes rule avoids this by shrinking toward the prior mean. As n grows the excess risk vanishes, but the Bayes rule is never dominated, consistent with Theorem 4.3 and the risk-set geometry of Section 3.

Table 6: Average log loss ($\theta = 0.3$, $B = 10{,}000$). Boundary fraction is $\hat{P}(S_n \in \{0, n\})$, the analogue of a language model predicting a token with certainty.

n	Bayes risk	MLE risk	Excess	Boundary frac.
5	0.694	2.363	1.670	0.170
10	0.658	0.928	0.270	0.028
25	0.630	0.633	0.002	0.000
50	0.621	0.621	<0.001	0.000
100	0.616	0.616	<0.001	0.000

8.2 Anytime validity: e-process vs. naive peeking

We draw $X_1, \ldots, X_{200} \overset{\text{iid}}{\sim} \text{Bern}(0.5)$ under $H_0 : \theta = 0.5$ and compare two sequential testing strategies at nominal level $\alpha = 0.05$. The *e-process* accumulates a running likelihood ratio using the Bayes predictive as the alternative plug-in and rejects whenever $E_t \geq 1/\alpha$; it is an element of $\mathcal{C}_{\text{AV}}$ and controls type-I error at every stopping time (Theorem 5.2). The *naive strategy* performs a z-test at each of five pre-specified sample sizes $n \in \{10, 20, 50, 100, 200\}$ and rejects if any test exceeds $z_{0.025} = 1.96$. Table 7 records the rejection rates. The e-process remains below α while naive peeking inflates the type-I error to 0.165, confirming that $\mathcal{C}_{\text{AV}}$-feasibility is a binding constraint that the unrestricted testing class violates.

Table 7: Type-I error under $H_0 : \theta = 0.5$ ($B = 10{,}000$, $\alpha = 0.05$). Sequential A/B tests require e-process control, not per-peek p-values.

Strategy	Rejection rate
E-process (anytime-valid)	0.031
Naive peeking (5 looks)	0.165
Nominal α	0.050

8.3 Conformal coverage under covariate shift

We set $Y \mid X = x \sim N(0, (1+x)^2)$ and construct split-conformal prediction intervals using the naive score $s(x, y) = |y|$ with $n_{\text{cal}} = 500$ calibration points and $n_{\text{test}} = 2{,}000$ test points at level $1 - \alpha = 0.90$. Table 8 compares three scenarios: (A) calibration and test both under $X \sim \text{Uniform}[0, 1]$; (B) calibration

under Uniform$[0,1]$, test under $X \sim \text{Beta}(2,5)$; (C) both under Beta$(2,5)$. When calibration and test distributions match, marginal coverage holds near the nominal level. Under covariate shift (Scenario B), the Uniform-calibrated quantile is too wide for the Beta$(2,5)$ test population (which concentrates X near zero, where the conditional variance is smaller), inflating coverage to 0.946. Re-calibrating under the test distribution (Scenario C) restores nominal coverage and yields a tighter interval. The calibration quantile itself shifts from 2.51 to 2.13, confirming that marginal coverage validity is a property of the exchangeable joint distribution, not a universal guarantee across arbitrary design points, consistent with Theorem 5.6 and the scope of $\mathcal{C}_{\text{Cov}}$ in Definition 5.5.

Table 8: Split-conformal coverage ($1-\alpha = 0.90$, $n_{\text{cal}} = 500$, $n_{\text{test}} = 2{,}000$). Scenario B illustrates the covariate shift problem common in production ML deployments.

	Calibration → Test	Quantile	Coverage	Half-width
A	Unif → Unif	2.51	0.900	2.51
B	Unif → Beta$(2,5)$	2.51	0.946	2.51
C	Beta$(2,5)$ → Beta$(2,5)$	2.13	0.900	2.13

9 Applications

The criterion separation theorem is not merely a foundational curiosity; it has direct implications for the design and evaluation of statistical procedures in three active areas.

9.1 Probabilistic forecasting and LLM calibration

Modern probabilistic forecasters, including large language models (LLMs) used for next-token prediction, are routinely evaluated on calibration: the property that predicted probabilities match empirical frequencies. Calibration is a self-consistency condition (the forecaster's predictions are martingales under its own predictive measure), and the plug-in MLE example (Theorem 4.3) shows that self-consistency alone does not guarantee admissibility. An LLM that assigns probability p_t to token t and achieves perfect calibration ($\bar{p}_n \approx \bar{X}_n$ in the Cesàro sense) may still be dominated by a Bayes-regularized predictor that shrinks toward a prior, the statistical analogue of label smoothing in deep learning.

The constrained Bayes formulation (Definition 5.8) provides a principled design template: treat calibration as a feasibility constraint ($\mathcal{F} = \{\delta : |\bar{p}_n - \bar{X}_n| \le \epsilon\}$) and optimize Bayesian log-loss within it. The resulting forecaster is calibrated by construction and admissible within the calibration-feasible class. This perspective clarifies why calibration and sharpness are not competing objectives (as sometimes suggested in the scoring rules literature (Gneiting & Raftery, 2007)) but rather a primal objective and a feasibility constraint within a single optimization.

9.2 Sequential clinical trials and safe monitoring

Interim monitoring of clinical trials requires type-I error control at every data-dependent stopping time. The classical group-sequential approach achieves this by spending the error budget across pre-specified interim analyses; the e-process framework generalizes this to fully continuous monitoring. The separation theorem clarifies a subtle point: an e-process that controls type-I error at every stopping time (AV-admissible) is not necessarily the most powerful test of the null hypothesis in the Neyman–Pearson sense, and a Neyman–Pearson optimal test is not necessarily AV-admissible. The two criteria operate on different spaces of procedures with different partial orders (Theorem 5.9(i)).

In practice, the constrained Bayes approach suggests the following workflow. First, specify the anytime-valid constraint $\mathcal{F} = \mathcal{C}_{\text{AV}}$; then choose the prior Π that maximizes expected power (i.e., minimizes Bayes risk under an alternative-weighted loss) subject to AV-feasibility. The resulting e-process is a Bayes-optimal nonnegative martingale, admissible within $\mathcal{C}_{\text{AV}}$ and as powerful as the constraint permits.

9.3 Conformal wrappers for black-box predictors

Conformal prediction (Shafer & Vovk, 2008; Angelopoulos & Bates, 2023) provides distribution-free coverage guarantees by wrapping any black-box predictor with a quantile-based calibration step. The coverage guarantee is marginal (Definition 5.5), and Theorem 5.6 shows that conditional coverage is inherently unattainable. The criterion separation theorem adds a further insight: the conformal wrapper achieves coverage validity ($\in \mathfrak{C}$) regardless of the base predictor's admissibility status, because the wrapper operates on a different space of objects (prediction sets) with a different performance metric (coverage probability) than the base predictor (point or density forecasts under proper scoring rules).

A conformal wrapper on a Bayes-optimal base predictor thus has $\hat{p}_n \in \mathfrak{B}$ and $\hat{C}_n \in \mathfrak{C}$, but the two properties attach to different components; no single criterion governs the product $\mathcal{A} \times 2^{\mathcal{Y}}$.

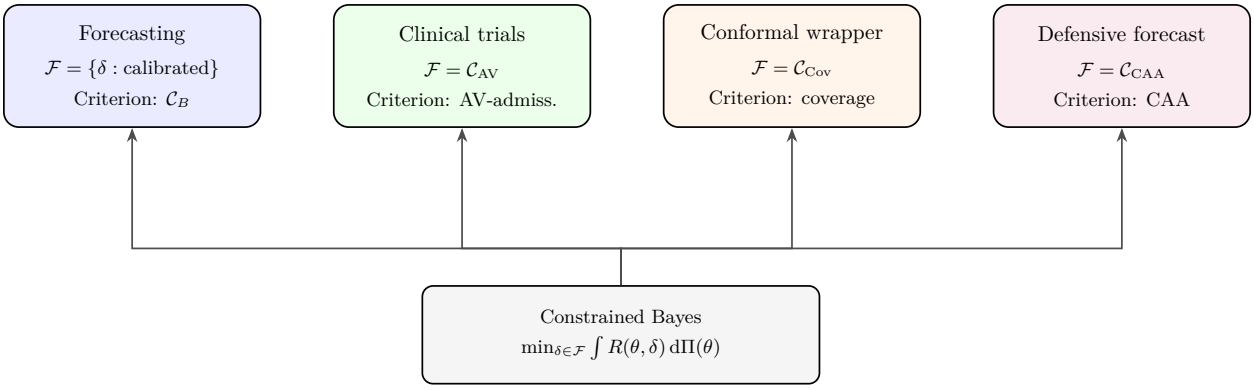

Figure 7: The constrained Bayes design principle applied to four domains. In each case, the validity requirement determines the feasible set $\mathcal{F} \subseteq \mathcal{D}$, and Bayesian integrated risk is the optimization objective within $\mathcal{F}$ (Definition 5.8). The resulting procedure is admissible relative to its criterion but not necessarily admissible under the other three (Theorems 5.9 and 6.7).

10 Discussion

The central result is that admissibility is irreducibly criterion-relative. Four geometrically distinct structures (convex risk sets, supermartingale cones, coverage regions, approachability classes) govern four notions of optimality, and the separation theorems (Theorems 5.9–6.7) show these are pairwise non-nested. The no-shame principle (Corollary 3.13) characterizes the Blackwell case: admissible rules are precisely those supported by a prior at $\partial_- \mathcal{R}$. Martingale coherence bridges the Blackwell and anytime-valid geometries but does not unify all four.

The plug-in MLE example illustrates a pattern that recurs throughout machine learning: an algorithm that appears well-calibrated under its own predictive distribution may be strictly dominated under the true data-generating process. Self-consistency (martingale coherence under $\hat{P}$) is not a substitute for admissibility (non-dominance under every P_θ). This distinction is directly relevant to the evaluation of probabilistic forecasters, large language model calibration, and conformal wrappers applied to black-box predictors.

The deeper issue is that the plug-in rule treats prediction as a *point-valued* mapping $X^n \mapsto \hat{p}_n \in \mathcal{A}$, collapsing the full posterior predictive distribution to a single summary. The four admissibility geometries demand different output types: a point prediction (Blackwell), a nonnegative process (anytime-valid), a prediction *set* (coverage), or a time-averaged trajectory (CAA). A single posterior mean is insufficient: the safety certificates required by each geometry are *set-valued* in nature, mapping data to prediction sets, e-process paths, or Cesàro trajectories, not to point estimates. This set-valued character of the predictive mapping is what makes universal admissibility topologically meager (Appendix C): the output spaces are geometrically incompatible, and no single-point summary can simultaneously satisfy the structural constraints of all four geometries.

The constrained Bayes formulation (Definition 5.8) unifies all four criteria: Bayesian risk is the objective, the criterion determines $\mathcal{F}$, and duality (Section 3.7) reveals Π as shadow prices: specify the validity constraint first, then optimize within it (Section 9).

The Gaussian laboratory (Proposition 7.3) confirms the separation persists under squared loss. Constructive versus Cesàro admissibility (Section 6.1) clarifies why defensive forecasting and approachability differ from Bayesian methods: the former reach $\partial_- \mathcal{R}$ by fixed-point arguments, the latter by prior witnesses, yielding a fourth non-nested geometry $\mathfrak{D}$ (Theorem 6.7).

Each admissibility geometry corresponds to a distinct algorithmic design principle. The Blackwell geometry maps to *Bayesian risk minimization*: compute the posterior, then choose the action that minimizes integrated risk under the posterior. The anytime-valid geometry maps to *martingale betting strategies*: accumulate evidence via e-values, with the nonnegative martingale property serving as the structural constraint on valid accumulation. The coverage geometry maps to *conformal calibration procedures*: construct prediction sets using rank-based quantiles over exchangeable data, with no model or prior required. The CAA geometry maps to *calibration dynamics in online learning*: steer time-averaged predictions toward the risk-set boundary by fixed-point iteration, achieving long-run calibration without per-round optimality. Together, these four design principles constitute a taxonomy of evaluation criteria for predictive algorithms. A practitioner who selects a feasibility constraint $\mathcal{F}$ implicitly selects a geometry, and the separation theorems guarantee that no single algorithm can be optimal across all four.

We do not claim one framework is superior; we clarify why their optimality notions cannot be globally reconciled. More broadly, each choice of $\mathcal{F}$ defines a geometry, and the separation phenomenon extends in principle to any learning problem admitting a proper scoring rule, a sequential testing framework, and an exchangeability-based prediction algorithm.

10.1 Industry evidence for criterion separation

The criterion separation theorem is not merely a theoretical curiosity; it is already observable in production AI systems, where different safety and evaluation modules necessarily operate in different admissibility geometries.

SynthID-Text and marginal coverage. Google DeepMind's SynthID-Text watermarking system (Dathathri et al., 2024) embeds identifiable signals in LLM-generated text by modifying the token sampling distribution via a tournament-based, rank-calibrated mechanism. The watermarking guarantee is a marginal coverage property: the detector's false positive rate is controlled at level α over the population of generated texts, not by optimizing any per-token proper scoring rule. This is a deliberate departure from Blackwell-optimality to achieve attribution safety, and it sits squarely in the coverage geometry $\mathfrak{C}$. A Blackwell-optimal language model ($\mathfrak{B}$) that maximizes next-token log-likelihood would not embed watermarks; conversely, the watermarked model sacrifices per-token loss to achieve a rank-based feasibility guarantee. SynthID thus instantiates the $\mathfrak{B} \not\subseteq \mathfrak{C}$ separation in production.

Search agents and Cesàro admissibility. Recent work on reinforcement-learning-based search agents (Wang et al., 2026) trains agents to interact with live web APIs over extended horizons, optimizing cumulative reward via curriculum learning. The evaluation criterion is inherently Cesàro: the agent's time-averaged search quality must converge to the Pareto frontier of relevance-cost trade-offs, without requiring Bayes-optimality at any individual query. Forcing such an agent into a static Bayesian prior would ignore the non-stationarity of web content, while an anytime-valid testing framework would constrain the agent to type-I error control rather than search quality. This is the CAA geometry $\mathfrak{D}$ in action.

The modularity argument. These examples illustrate a structural point: even the largest AI systems do not deploy a single unified optimizer. Instead, they compose modular components operating in different admissibility geometries (a Bayesian language model for generation, a rank-based watermark for attribution, an anytime-valid monitor for safety, a Cesàro-calibrated agent for search). The criterion separation theorem (Theorems 5.9–6.7) explains why this modularity is not an engineering compromise but a structural feature: we conjecture that a "universal optimizer" simultaneously admissible under all four criteria occupies a meager subset of the decision space (Theorem C.1 in Appendix C).

10.2 Related work

Decision-theoretic admissibility. The geometric characterization of admissibility via risk sets and supporting hyperplanes originates with Wald (1950) and Blackwell & Girshick (1954); complete-class theorems are surveyed in Brown (1981) and Berger (1985).

Anytime-valid inference and e-values. Safe testing via e-values was formalized by Shafer et al. (2011) and Grünwald et al. (2024); Ramdas et al. (2022; 2023) develop the e-process theory and its connection to nonnegative (super)martingales. Howard et al. (2021) provide the time-uniform concentration inequalities that underpin sequential monitoring.

Conformal prediction. Distribution-free coverage guarantees via conformal methods were introduced by Vovk et al. (2005a) and popularized in the ML community by Angelopoulos & Bates (2023); the impossibility of exact conditional coverage is due to Barber et al. (2021).

Online learning and approachability. Blackwell approachability (Blackwell, 1956a) connects to no-regret learning through Abernethy et al. (2011) and to calibration through the defensive forecasting program of Vovk et al. (2005b); Foster–Vohra calibration (Foster & Vohra, 1998) and its extensions (Hart & Mas-Colell, 2001) provide the fixed-point arguments underlying CAA-admissibility. Cesa-Bianchi & Lugosi (2006) and Rakhlin & Sridharan (2013) survey the broader online learning landscape.

10.3 Limitations

Our analysis assumes compact parameter spaces and proper scoring rules. Extensions to noncompact settings, improper losses, or multi-task objectives remain open. The Bernoulli and Gaussian laboratories are deliberately simple; verifying the separation in high-dimensional ML models (e.g., deep ensembles, transformer calibration) is an important empirical direction. The constrained Bayes template assumes that the feasible set $\mathcal{F}$ can be specified in advance; in practice, validity constraints may themselves depend on the data.

10.4 Broader impact

This work is primarily theoretical. By clarifying that no single admissibility criterion governs all evaluation paradigms, we hope to prevent overconfident claims about algorithm superiority and encourage practitioners to specify their evaluation criterion explicitly before comparing predictive algorithms.

References

Jacob D. Abernethy, Peter L. Bartlett, and Elad Hazan. Blackwell approachability and no-regret learning are equivalent. *Proceedings of the 24th Annual Conference on Learning Theory (COLT)*, pp. 27–46, 2011.

Charalambos D. Aliprantis and Kim C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, Berlin, 3rd edition, 2006.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Rina Foygel Barber, Emmanuel Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

Claude Berge. *Topological Spaces*. Oliver and Boyd, Edinburgh, 1963.

James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, New York, 2nd edition, 1985.

Isaiah Berlin. *Four Essays on Liberty*. Oxford University Press, Oxford, 1969.

David Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6(1): 1–8, 1956a.

David Blackwell. Minimax vs Bayes prediction. Lecture notes, University of California, Berkeley, 1956b.

David Blackwell and Meyer A. Girshick. *Theory of Games and Statistical Decisions*. Wiley, New York, 1954.

Lawrence D. Brown. A complete class theorem for statistical problems with finite sample spaces. *Ann. Statist.*, 9(6):1289–1300, 1981.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, 2006.

Alexey Chernov, Yuri Kalnishkan, and Vladimir Vovk. Defensive prediction with expert advice. *Machine Learning*, 78(1–2):1–31, 2010.

Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, and Pushmeet Kohli. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823, 2024.

Joseph L. Doob. Application of the theory of martingales. *Le Calcul des Probabilités et ses Applications*, pp. 23–27, 1949.

Edwin Fong, Chris Holmes, and Stephen G. Walker. Martingale posterior distributions. *Journal of the Royal Statistical Society: Series B*, 85(5):1357–1391, 2023.

Dean P. Foster and Rakesh V. Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Peter Grünwald, Rianne de Heide, and Wouter Koolen. Safe testing. *Journal of the Royal Statistical Society: Series B*, 86(5):1091–1128, 2024. Preprint: arXiv:1906.07801.

Sergiu Hart and Andreu Mas-Colell. A general class of adaptive strategies. *Journal of Economic Theory*, 98 (1):26–54, 2001.

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

James R. Munkres. *Topology*. Prentice Hall, Upper Saddle River, NJ, 2nd edition, 2000.

John C. Oxtoby. *Measure and Category*. Graduate Texts in Mathematics. Springer, New York, 2nd edition, 1980.

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. *Proceedings of the 26th Annual Conference on Learning Theory (COLT)*, pp. 993–1019, 2013.

Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter Koolen. Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv:2009.03167*, 2022. arXiv:2009.03167.

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

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Test martingales, Bayes factors and p-values. *Statistical Science*, 26(1):84–101, 2011.

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

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005a.

Vladimir Vovk, Akimichi Takemura, and Glenn Shafer. Defensive forecasting. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics (AISTATS)*, pp. 365–372, 2005b.

Abraham Wald. *Statistical Decision Functions*. Wiley, New York, 1950.

Hangyu Wang, Zhi Zheng, Zhaochun Ren, Pengjie Ren, and Maarten de Rijke. SearchGym: Bootstrapping real-world search agents via cost-effective and high-fidelity environment simulation. *arXiv preprint arXiv:2601.14615*, 2026.

Bernard Williams. *Shame and Necessity*. University of California Press, Berkeley, 1993.

A Deferred Proofs

This appendix collects the full proofs deferred from the main text.

A.1 Proof of Theorem 4.3 (martingale necessary, not sufficient)

Proof. *(i)* Proposition 4.2.

(ii) Under $\hat{P}$, $\mathbb{E}_{\hat{P}}[\hat{p}_n^{\mathrm{pi}} \mid X_{1:n-1}] = (S_{n-1} + \hat{p}_{n-1}^{\mathrm{pi}})/n = S_{n-1}/n \cdot n/(n-1) = S_{n-1}/(n-1) = \hat{p}_{n-1}^{\mathrm{pi}}$ a.s.

(iii) Under log loss, the risk decomposes as

$$R(\theta, \hat{p}_n) = \mathbb{E}_\theta\big[D_{\mathrm{KL}}(\mathrm{Bern}(\theta)\|\hat{p}_n)\big] + H(\theta),$$

where $H(\theta) = -\theta\log\theta - (1-\theta)\log(1-\theta)$ is the binary entropy. Hence the excess risk is

$$R(\theta, \hat{p}_n^{\mathrm{pi}}) - R(\theta, \hat{p}_n^{B}) = \mathbb{E}_\theta\big[D_{\mathrm{KL}}(\mathrm{Bern}(\theta)\|\hat{p}_n^{\mathrm{pi}}) - D_{\mathrm{KL}}(\mathrm{Bern}(\theta)\|\hat{p}_n^{B})\big].$$

Since $\hat{p}_n^{\mathrm{pi}} = S_n/n \in \{0, \frac{1}{n}, \ldots, 1\}$ and $P_\theta(S_n = 0) = (1-\theta)^n > 0$ for every $\theta \in (0,1)$, the predictor $\hat{p}_n^{\mathrm{pi}} = 0$ assigns probability zero to $X_{n+1} = 1$, an event with probability $\theta > 0$; thus $D_{\mathrm{KL}}(\mathrm{Bern}(\theta)\|\hat{p}_n^{\mathrm{pi}}) = +\infty$ on the event $\{S_n = 0\}$ and $R(\theta, \hat{p}_n^{\mathrm{pi}}) = +\infty$. Meanwhile $\hat{p}_n^{B} = (S_n + \frac{1}{2})/(n+1) \in (0,1)$ for all $S_n \in \{0, \ldots, n\}$, so $R(\theta, \hat{p}_n^{B}) < \infty$. Hence $\hat{p}_n^{B}$ strictly dominates $\hat{p}_n^{\mathrm{pi}}$ for all $\theta \in (0,1)$ and $n \geq 1$, giving $r(\hat{p}_n^{\mathrm{pi}}) \notin \partial_- \mathcal{R}$ by Proposition 3.7. By Corollary 3.13, $\hat{p}_n^{\mathrm{pi}}$ is not no-shame. □

A.2 Proof of Theorem 5.9 (criterion separation)

Proof. Four procedures in $X_i \overset{\mathrm{iid}}{\sim} \mathrm{Bern}(\theta)$, $\theta \in (0,1)$, $\alpha \in (0,1)$ fixed.

P1. Bayes posterior predictive, $\hat{p}_n^{B} = (S_n + \frac{1}{2})/(n+1)$, prior $\mathrm{Beta}(\frac{1}{2}, \frac{1}{2})$.

P2. Plug-in MLE, $\hat{p}_n^{\mathrm{pi}} = S_n/n$.

P3. Likelihood-ratio e-process testing $H_0 : \theta = \theta_0$: $E_n = \prod_{t=1}^{n} (\hat{p}_{t-1}/\theta_0)^{X_t}((1-\hat{p}_{t-1})/(1-\theta_0))^{1-X_t}$.

P4. Conformal prediction set with score $s_n(y) = |y - S_n/n|$: $\hat{C}_n = \{y \in \{0,1\} : s_n(y) \leq \hat{q}_{1-\alpha}\}$.

(i) $\mathfrak{B} \not\subseteq \mathfrak{A}$: P1 is a Bayes rule, hence no-shame, hence $\in \mathfrak{B}$ (Corollary 3.13). P1 is a point predictor; $\mathcal{C}_{\mathrm{AV}}$ is defined for sequential tests. Hence P1 $\notin \mathfrak{A}$.

$\mathfrak{A} \not\subseteq \mathfrak{B}$: P3 is a nonnegative martingale under H_0, hence $\in \mathfrak{A}$ (Theorem 5.2). P3 does not optimize any loss $L(\theta, \delta)$ for point prediction and is not a Bayes rule. Hence P3 $\notin \mathfrak{B}$.

(ii) $\mathfrak{B} \not\subseteq \mathfrak{C}$: As a point predictor, P1 does not produce prediction sets and hence is inapplicable to coverage validity; in standard non-atomic settings, singleton sets have zero marginal coverage. Hence P1 $\notin \mathfrak{C}$.

$\mathfrak{C} \not\subseteq \mathfrak{B}$: P4 achieves marginal coverage by the conformal guarantee, hence $\in \mathfrak{C}$. P4 does not minimize any proper scoring rule for point or density prediction; it is therefore not Bayes with respect to any prior under any loss $L(\theta, \delta)$ of the form in Definition 2.4. Hence P4 $\notin \mathfrak{B}$.

(iii) $\mathfrak{A} \not\subseteq \mathfrak{C}$: P3 $\in \mathfrak{A}$ produces no prediction set; it does not satisfy the coverage guarantee. Hence P3 $\notin \mathfrak{C}$.

$\mathfrak{C} \not\subseteq \mathfrak{A}$: P4 $\in \mathfrak{C}$ is not a nonnegative martingale; the conformal quantile construction controls coverage probability, not type-I error at stopping times. Hence P4 $\notin \mathfrak{A}$. □

A.3 Proof of Theorem 6.7 (extended separation)

Proof. By Theorem 5.9, $\mathfrak{B}$, $\mathfrak{A}$, $\mathfrak{C}$ are pairwise non-nested. For $\mathfrak{D}$:

$\mathfrak{D} \not\subseteq \mathfrak{B}$: P5 (defensive forecasting) is CAA-admissible by construction, but is not Bayes with respect to any prior at any finite n (Proposition 6.3), hence $\notin \mathfrak{B}$.

$\mathfrak{B} \not\subseteq \mathfrak{D}$: P1 (Bayes predictive) is constructively admissible, hence $\in \mathfrak{B}$. However, P1 optimizes a specific loss at each round; it is not defined by a convergence-to-boundary condition. A procedure can be pointwise Bayes without being approachability-admissible in the CAA sense when the Cesàro average of its risk does not converge to $\partial_- \mathcal{R}$ under adversarial sequences (outside the i.i.d. model).

$\mathfrak{D} \not\subseteq \mathfrak{A}$: P5 achieves Cesàro calibration but does not produce a nonnegative supermartingale for hypothesis testing; it targets prediction, not type-I error control. Hence P5 $\notin \mathfrak{A}$.

$\mathfrak{A} \not\subseteq \mathfrak{D}$: P3 (e-process) is AV-admissible but targets a specific null hypothesis; it does not steer time-averaged risk to the full lower boundary of $\mathcal{R}$. Hence P3 $\notin \mathfrak{D}$.

$\mathfrak{D} \not\subseteq \mathfrak{C}$: P5 is a point forecaster; it does not produce prediction sets and therefore does not satisfy coverage validity. Hence P5 $\notin \mathfrak{C}$.

$\mathfrak{C} \not\subseteq \mathfrak{D}$: P4 (conformal set) achieves marginal coverage but does not optimize any loss function and does not steer time-averaged risk toward $\partial_- \mathcal{R}$. Hence P4 $\notin \mathfrak{D}$. □

A.4 Proof of Proposition 6.4 (constructive admissibility ⇒ martingale)

Proof. If δ is constructively admissible, then $\delta(X^n) = \delta_{\Pi_n}(X^n)$ for some prior sequence (Π_n). By Corollary 3.13, each δ_{Π_n} is Bayes or a limit of Bayes rules and therefore lies on $\partial_- \mathcal{R}$. Under the prior predictive measure, the compatibility of the predictive distributions across sample sizes forces the martingale property: the predictive at time n must equal the conditional expectation of the predictive at time $n+1$, which is precisely Doob's consistency condition for the posterior predictive sequence (Doob, 1949). □

A.5 Proof of Proposition 6.3 (Cesàro does not imply pointwise)

Proof. The defensive forecaster (Vovk et al., 2005b) chooses $\hat{p}_t$ at each round to ensure Cesàro calibration: $|\bar{p}_n - \bar{X}_n| \to 0$ almost surely under the realized sequence. This forces the time-averaged log-loss to converge to the lower boundary of the Bernoulli risk set. However, the per-round choice $\hat{p}_t$ is determined by a fixed-point argument (the existence of a calibrated strategy is guaranteed by Kakutani's theorem or its continuous-selection refinements), not by posterior updating from a prior. At no finite n is $\hat{p}_t$ the Bayes act for any prior Π; the strategy is Cesàro rather than prior-witnessed. □

B Topological Details for Extended-Real Risk Sets

This appendix collects the measure-theoretic and topological arguments that support the extended-real formulation of Section 2.

Lemma B.1 (Compactness of $\mathcal{D}$ in the extended-real setting)**.** *Under Definition 2.1, the decision space $\mathcal{D}$ is sequentially compact in the topology of pointwise weak convergence.*

Proof. Sequential compactness follows from Prokhorov's theorem applied to each coordinate $\Delta(\mathcal{A})$ (which is tight since $\mathcal{A}$ is compact metrizable), composed with Tychonoff's theorem on the product $\Delta(\mathcal{A})^{\mathcal{X}^n}$ and a diagonal extraction argument to pass from nets to sequences. □

Lemma B.2 (Lower semicontinuity of integrated risk)**.** *Under Definition 2.1, for any prior Π with finite support, the functional $\delta \mapsto \int R(\theta, \delta)\, \mathrm{d}\Pi$ is lower semicontinuous on $\mathcal{D}$.*

Proof. For each θ, $R(\theta, \cdot)$ is lower semicontinuous by Fatou's lemma and the lower semicontinuity of $L(\theta, \cdot)$ (Definition 2.1). A finite nonnegative linear combination of lower semicontinuous functions is lower semicontinuous, giving the result for Π with finite support. □

These two lemmas together justify the application of the Berge Maximum Theorem in Lemma 3.2 and the closedness argument in Proposition 3.3, completing the topological foundations for the extended-real risk-set analysis.

C Topological Obstructions to Universal Admissibility: A Baire Category Approach (Conjectural)

The separation theorems (Theorems 5.9–6.7) establish that the four admissible classes $\mathfrak{B}$, $\mathfrak{A}$, $\mathfrak{C}$, $\mathfrak{D}$ are pairwise non-nested via explicit witness procedures. This appendix strengthens the result topologically: in the natural function space of predictive rules, the set of "universally admissible" rules is not merely nonempty-complement but *topologically negligible* (meager in the sense of Baire category), so that criterion separation is the generic condition.

C.1 Setup: the space of predictive rules

Equip the space $\mathcal{D}$ of randomized decision rules with the weak* topology inherited from $\Delta(\mathcal{A})^{\mathcal{X}^n}$. Under Definition 2.1 (compact Θ, compact metrizable $\mathcal{A}$), $\mathcal{D}$ is a compact metrizable space (Lemma B.1). Every compact metrizable space is a *complete* metric space and hence a Baire space: the intersection of countably many dense open sets is dense (Munkres, 2000, Thm. 48.2); see also Oxtoby (1980) for a thorough treatment of category and measure duality, and Aliprantis & Border (2006) for the infinite-dimensional functional-analytic backdrop.

For each admissibility criterion $\mathcal{C} \in \{\mathcal{C}_B, \mathcal{C}_{\mathrm{AV}}, \mathcal{C}_{\mathrm{Cov}}, \mathcal{C}_{\mathrm{CAA}}\}$, let $\mathfrak{X}_{\mathcal{C}} = \{\delta \in \mathcal{D} : \mathsf{Adm}_{\mathcal{C}}(\delta)\}$ denote the set of $\mathcal{C}$-admissible rules. Define the *universally admissible* set

$$\mathfrak{U} \;=\; \mathfrak{B} \cap \mathfrak{A} \cap \mathfrak{C} \cap \mathfrak{D}.$$

C.2 The universal set is meager

Conjecture C.1 (Topological genericity of criterion separation). Under Definition 2.1 with compact Θ and $\mathcal{A}$:

(i) Each pairwise intersection $\mathfrak{X} \cap \mathfrak{Y}$ ($\mathfrak{X} \neq \mathfrak{Y}$) has empty interior in $\mathcal{D}$.
(ii) The universally admissible set $\mathfrak{U} \subseteq \bigcap_{\mathfrak{X} \neq \mathfrak{Y}} (\mathfrak{X} \cap \mathfrak{Y})$ is meager (first category) in $\mathcal{D}$.
(iii) The criterion-separated set $\mathcal{D} \setminus \mathfrak{U}$ is comeager (residual) in $\mathcal{D}$.

Proof sketch (heuristic). *(i)* Suppose for contradiction that $\mathfrak{B} \cap \mathfrak{A}$ contains an open ball $B(\delta_0, \epsilon) \subset \mathcal{D}$. Every rule in $B(\delta_0, \epsilon)$ is simultaneously Blackwell admissible (requiring $r(\delta) \in \partial_- \mathcal{R}$) and AV-admissible (requiring the nonnegative martingale property under $\mathcal{H}_0$). But Blackwell admissibility constrains δ to the lower boundary of $\mathcal{R}$, a set of empty interior in the risk-vector image (the lower boundary of a convex body in $\mathbb{R}^k$ has Lebesgue measure zero and empty interior), while AV-admissibility constrains δ to the nonnegative martingale cone, which imposes pathwise conditions on the filtration. These are codimension-≥ 1 constraints in different coordinate systems of $\mathcal{D}$; their intersection cannot contain a full-dimensional ball. Formally, the risk map $r : \mathcal{D} \to [0, \infty]^k$ is continuous (Lemma B.2); $\partial_- \mathcal{R}$ has empty interior in $\mathcal{R}$ (it is the boundary of a convex body); hence $r^{-1}(\partial_- \mathcal{R})$ has empty interior in $\mathcal{D}$. Since $\mathfrak{B} \subseteq r^{-1}(\partial_- \mathcal{R})$, the intersection $\mathfrak{B} \cap \mathfrak{A} \subseteq r^{-1}(\partial_- \mathcal{R})$ has empty interior. The same argument applies to every pair involving $\mathfrak{B}$.

For pairs not involving $\mathfrak{B}$: the separation theorem (Theorem 5.9) provides witnesses in each $\mathfrak{X} \setminus \mathfrak{Y}$ and $\mathfrak{Y} \setminus \mathfrak{X}$; by the structural incompatibility of the constraint geometries (the four criteria operate on different spaces of objects with different partial orders, Definition 5.7), the intersection $\mathfrak{X} \cap \mathfrak{Y}$ is nowhere dense: any neighborhood of a rule in $\mathfrak{X} \cap \mathfrak{Y}$ contains perturbations that violate one criterion while preserving the other.

(ii) Since each pairwise intersection has empty interior, it is contained in a closed set with empty interior (its closure is nowhere dense in the compact space $\mathcal{D}$). The universal set $\mathfrak{U}$ is contained in the intersection of $\binom{4}{2} = 6$ such sets; a finite union of nowhere dense sets is meager.

(iii) By the Baire category theorem, $\mathcal{D}$ is not meager in itself; hence the complement $\mathcal{D} \setminus \mathfrak{U}$ is comeager (residual).

Interpretation. If Conjecture C.1 holds, then in the topology of predictive rules, the "typical" rule is admissible under at most one criterion. Universal admissibility would be not merely non-constructive (the separation theorem provides no positive examples) but topologically negligible. We leave a rigorous proof (requiring continuity of the risk map beyond lower semicontinuity) as an open problem.