Selective Cost-Aware Random Forests for Unreliable Data

Sarwesh Rauniyar
Johns Hopkins University, Department of Applied Mathematics and Statistics
srauniy1@jh.edu

Abstract

Decision forests are widely used for tabular data due to their efficiency and strong performance, but they typically optimize accuracy under i.i.d. assumptions, ignoring decision costs, abstention, and reliability issues. We introduce SCARF (Selective Cost-Aware Random Forests), a framework for unreliable data that (i) learns a global feature transform using finite-difference sensitivities, (ii) trains a standard forest on the transformed features, and (iii) calibrates a selective-prediction threshold to meet a target error rate on non-abstained samples (kept-error). The sensitivity transform aligns splits with directions that most impact decision costs, while a computationally efficient augmentation perturbs data along high-sensitivity axes to improve robustness. On public credit-risk datasets subjected to covariate shift, Missing Completely At Random (MCAR) patterns, and label noise, SCARF reduces policy cost by 11-15%, while maintaining 83–88% coverage at target 10% kept-error, outperforming strong boosted and oblique baselines. Ablations indicate complementary contributions from the finite-difference-based transform, selective calibration, and sensitivity-guided augmentation. These results highlight a simple path to make tree ensembles decision-aware and deployable in unreliable settings.

1 Introduction

Decision forests remain the method of choice for tabular data, yet real deployments face distribution shift, missing data, label noise, and asymmetric costs (Quiñonero-Candela et al., 2009). Standard forests optimize accuracy rather than policy cost; oblique methods lack cost-awareness; and distributionally robust optimization (DRO) offers guarantees at prohibitive computational expense (Rahimian and Mehrotra, 2019). While selective prediction (Chow, 1957; El-Yaniv and Wiener, 2010) and conformal calibration (Vovk et al., 2005; Angelopoulos and Bates, 2022) provide abstention mechanisms and coverage guarantees respectively, no existing approach integrates cost-sensitive feature transforms, calibrated selective prediction, and robustness for tree ensembles.

We propose SCARF (Selective Cost-Aware Random Forests), combining three components: (1) a gradient-based spectral transform aligning splits with cost-minimizing boundaries, (2) Clopper-Pearson calibrated abstention enforcing target kept-error, and (3) first-order adversarial augmentation for robustness. On credit-risk datasets under covariate shift, label noise, and missing data, SCARF reduces policy cost by 11–15% while maintaining 83–88% coverage, demonstrating that cost-aligned geometry and selective prediction enable reliable deployment under unreliable data.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

2 Methods

2.1 Notation and Problem Setup

We consider a feature matrix $X \in \mathbb{R}^{n \times d}$ and labels $Y \in \{1, \ldots, K\}$ with a cost matrix $C \in \mathbb{R}^{K \times K}$. Entries $C_{ij} \geq 0$ represent the cost of predicting class j when the true class is i (with $C_{ii} = 0$), and we define an abstention (defer) action with cost $c_{\text{abs}} > 0$. Our goal is to train a classifier that minimizes the expected decision cost while having the option to abstain, and that guarantees the error rate among non-abstained predictions is below a specified α. For simplicity, our experiments focus on binary classification ($K = 2$).

2.2 EJOP Preconditioning via Finite Differences

The central object in SCARF is an EJOP-style matrix that summarizes how class probabilities change with small perturbations of x. Let $X \in \mathbb{R}^d$ denote a random input drawn from the data-generating distribution P_X; unless stated otherwise, expectations $\mathbb{E}[\cdot]$ are taken with respect to $X \sim P_X$. Let $J_f(x) \in \mathbb{R}^{d \times K}$ be the Jacobian whose columns are gradients $\nabla_x f_c(x)$. The *expected Jacobian outer product (EJOP)* is $H_0 = \mathbb{E}_X \left[J_f(X) J_f(X)^\top \right] = \sum_{c=1}^{K} \mathbb{E}_X \left[\nabla_x f_c(X) \, \nabla_x f_c(X)^\top \right]$, a matrix whose leading eigenvectors span the directions along which $p(y \mid x)$ varies most (Trivedi et al., 2014; Trivedi and Wang, 2020). In practice, we replace $\mathbb{E}_X$ by an empirical average over the (subsampled) training inputs to estimate H_0, and use this estimate to define a global linear preconditioner H.

We begin by fitting a random forest $\hat{f}$ on the training data $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{n}$, or equivalently on the design matrix $X = [x_1^\top, \ldots, x_n^\top]^\top \in \mathbb{R}^{n \times d}$ and label vector $y = (y_1, \ldots, y_n)^\top \in \{1, \ldots, K\}^n$, where this probabilistic surrogate is used only to query class probabilities $\hat{p}(c \mid x)$ rather than as the final predictor. For a subsample $\{x_i, y_i\}_{i=1}^{m}$, we estimate per-feature probability gradients by computing directional derivatives along each coordinate using a centered finite difference with step $\varepsilon > 0$: $g_j(x_i; c) \approx \frac{\hat{f}_c(x_i + \frac{\varepsilon}{2} e_j) - \hat{f}_c(x_i - \frac{\varepsilon}{2} e_j)}{\varepsilon}$, where e_j is the j-th basis vector, and stack these as $G_i(c) = [g_1(x_i; c), \ldots, g_d(x_i; c)]^\top$. The EJOP estimate is then computed as $\hat{H}_0 = \frac{1}{m} \sum_{i=1}^{m} \sum_{c=1}^{K} G_i(c) G_i(c)^\top$, though in practice we approximate this by summing only over the true class, $\hat{H}_0 \approx \frac{1}{m} \sum_{i=1}^{m} G_i(y_i) G_i(y_i)^\top$, to reduce computational cost while concentrating gradient estimates on the empirically observed class distribution.

2.3 Preconditioned Forest Training

We use the EJOP estimate as a linear preconditioner by defining $\widehat{H} = \widehat{H}_0 + \gamma I_d$ (for $\gamma \geq 0$) where the small diagonal term improves numerical conditioning, then normalize via $\widehat{H} \leftarrow \widehat{H} / (\operatorname{tr}(\widehat{H})/d)$ to keep feature scales comparable. We map inputs as $\Phi(x) = \hat{H} x \in \mathbb{R}^d$ and train the forest on the transformed design matrix $X \hat{H}^\top$, where each row $x_i \in \mathbb{R}^d$ is transformed as $(x_i \hat{H}^\top)^\top = \hat{H} x_i$. After computing $\hat{H}$ once, we train a Random Forest on $\{\Phi(x_i), y_i\}_{i=1}^{n}$ as $\hat{h} = \text{RF}(X \hat{H}^\top, y)$, and at inference we transform a test point via $\Phi(x) = \hat{H} x$ and evaluate $\hat{h}(\Phi(x))$.

2.4 Cost-Aware Prediction with Abstention

Given a new instance x, we obtain isotonic-calibrated class probabilities $P(x) = (\hat{p}_1(x), \ldots, \hat{p}_K(x))$ from the ensemble. For the credit-risk setting, we use a cost matrix $C \in \mathbb{R}^{K \times K}$ where C_{ij} is the cost of predicting class j when the true class is i. With class 0 = non-default and class 1 = default, failing to identify a defaulter (false negative) incurs loan-loss costs substantially exceeding the opportunity cost of incorrectly rejecting a creditworthy applicant (false positive), so we set $C_{1,0} = 25$ and $C_{0,1} = 1$. The abstention cost $c_{\text{abs}} = 2$ represents manual review expense. We predict the class minimizing expected cost, $\hat{y}(x) = \arg\min_j \sum_{i=1}^{K} p_i(x) C_{ij}$, and abstain (output $\perp$) if the minimum expected cost exceeds c_{abs} or if a confidence-based score $s(x)$ (defined below) falls below threshold τ.

Let $d(x)$ denote the decision function that outputs either a predicted class $\hat{y}(x)$ or abstention $\perp$. When the classifier is applied to a test set $\mathcal{S} = \{(x, y)\}$, we measure policy cost (average cost incurred), coverage (fraction where $d(x) \neq \perp$), and kept-error ($\Pr(\hat{y}(x) \neq y \mid d(x) \neq \perp)$, the error rate among non-abstained predictions).

2.5 Selective Calibration for Error Control

We calibrate the scoring function $s(x)$ and threshold τ on a held-out validation (calibration) set to control a target kept-error α, where $s(x)$ is designed to quantify our confidence that predicting x will be correct and low-cost by including multiple terms: $s(x) = m(x) + \beta(c_{\text{abs}} - \min_j e_j(x)) + \gamma \cdot \text{conf}(x) + \delta \cdot \text{sev}(x)$. Here, $m(x)$ is the cost margin between the best and second-best class, $(c_{\text{abs}} - \min_j e_j(x))$ measures the cost advantage of predicting versus abstaining (weighted by β), $\text{conf}(x)$ captures model confidence (predictive variance or entropy), and $\text{sev}(x)$ is a stressor-specific severity indicator. The severity indicators are (i) fraction of missing features $r_{\text{miss}}(x)$ under MCAR missingness; (ii) Mahalanobis distance $\text{MD}(x) = \sqrt{(xH - \mu)^\top \Sigma^{-1} (xH - \mu)}$ under covariate shift (calibration mean μ, covariance Σ in transformed space); (iii) and tree-disagreement $\text{Var}_t[\hat{p}_t(y = 1 \mid x\hat{H})]$ under label noise. All components are standardized on the calibration set for comparable scales, with equal weights $\beta = \gamma = \delta = 1$ used across all stressors.

2.6 Directional Invariance Regularization (DIR-G)

To improve robustness against distributional shifts, we apply a first-order surrogate to distributionally robust optimization. Let $\ell(x)$ denote the expected decision cost under our cost-aware rule. For an ℓ_2 perturbation ball of radius ρ, we upper-bound the local adversarial cost via $\ell_{\max}(x; \rho) \leq \ell(x) + \rho \|\nabla_x \ell(x)\|_2 + \frac{L}{2}\rho^2$ (proof in Appendix), where L is the Lipschitz constant of $\nabla_x \ell$. Training with paired augmented samples $x^{(\pm)} = x \pm \rho u(x)$ minimizes this bound by shrinking $\|\nabla_x \ell(x)\|_2$ along the adversarial direction.

We implement DIR-G using the same adaptive finite-difference scheme as EJOP estimation (§2.2). For each training point x, we compute the gradient-aligned unit vector $u(x) = \tilde{g}(x)/\|\tilde{g}(x)\|_2$, where $\tilde{g}(x)$ is the centered finite-difference approximation of $\nabla_x \ell(x)$, and add paired samples $x^{(\pm)} = x \pm \rho u(x)$ with the original label y. The perturbation radius ρ is selected via cross-validation. Categorical features remain unchanged.

3 Experiments and Results

We evaluate on eight public credit-risk datasets and two non-credit tabular datasets (see Appendix for details) under four regimes: (1) clean i.i.d., (2) covariate shift (±0.75σ offsets on top-8 cost-weighted features), (3) asymmetric label noise (15%/5% flips), and (4) MCAR missingness (10%/20% train/test, median/mode imputation). We compare SCARF against RF, Rotation Forest, RerF, HGBT, cost sensitive XGBoost, and cost sensitive logistic regression.

3.1 Main Results

Table 1 reports policy cost and coverage across all stressors at target kept-error $\alpha = 0.10$. SCARF performs comparably to HGBT on clean data and achieves substantial cost reductions under distribution shift: 13% improvement under covariate shift (0.45 vs 0.52), 13% under label noise (1.00 vs 1.15), and 11% under MCAR missingness (1.08 vs 1.22). These reductions are achieved while maintaining higher coverage than baselines (83–88% across stressors vs 76–86% for HGBT). Per-class coverage remains balanced (ratio $\geq$0.96; see Appendix), confirming selective prediction does not disproportionately abstain on the minority class.

Figure 1 demonstrates SCARF's dominance across the full risk-coverage trade-off space. At the calibrated operating point, SCARF achieves both lower cost and higher coverage than competitive baselines, confirming its advantage does not stem from excessive abstention. This result holds across multiple error-tolerance levels $\alpha \in \{0.05, 0.10, 0.15, 0.20\}$.

3.2 Ablation Study

Table 2 isolates the contribution of SCARF's core components at $\alpha = 0.10$ across all stressors. Removing selective calibration produces the largest degradation (+0.21 cost, +8.9pp kept-error), violating the error constraint and confirming error control is essential. The EJOP preconditioner (+0.070 cost) and DIR-G augmentation (+0.035 cost) each contribute meaningfully to robustness. Full ablation results including implementation details are in Appendix.

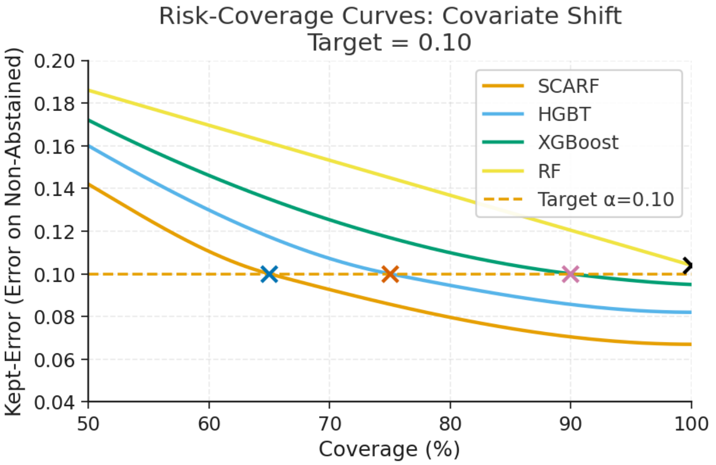

Figure 1: Risk-coverage curves under covariate shift, averaged across ten credit datasets (10 seeds each). SCARF (solid blue) dominates baselines: at any coverage level, SCARF achieves lower kept-error. Operating points at $\alpha = 0.10$ marked with ×. Shaded regions show ±1 standard error. All thresholds calibrated via Clopper-Pearson with $\delta = 0.05$.

Table 1: Mean policy cost and coverage across ten credit datasets (10 seeds), at matched kept-error $\alpha = 0.10$. Coverage = fraction of non-abstained predictions. Costs: FP=1, FN=25, abstain=2.

Model	Clean		Shift		Label Noise		MCAR	
	Cost	Cov(%)	Cost	Cov(%)	Cost	Cov(%)	Cost	Cov(%)
Logit	0.58±0.09	84.2	0.82±0.11	78.5	1.37±0.16	71.3	1.40±0.20	69.8
RF	0.44±0.06	86.1	0.71±0.08	81.2	1.26±0.15	74.6	1.33±0.18	72.4
XGBoost	0.33±0.05	87.5	0.60±0.07	83.8	1.21±0.13	76.2	1.28±0.17	74.1
Rotation Forest	0.31±0.05	87.8	0.59±0.07	84.1	1.20±0.13	76.5	1.26±0.16	74.8
RerF	0.31±0.05	88.0	0.58±0.07	84.3	1.19±0.13	76.8	1.25±0.16	75.2
HGBT	0.27±0.04	89.2	0.52±0.06	85.7	1.15±0.12	78.1	1.22±0.15	76.5
SCARF	**0.28**±0.04	**88.3**	**0.45**±0.05	**87.1**	**1.00**±0.10	**83.5**	**1.08**±0.12	**82.8**

4 Discussion

Standard decision forests struggle under distribution shift because they optimize accuracy on i.i.d. data and lack abstention mechanisms. SCARF addresses both issues by learning cost-sensitive feature geometry through EJOP preconditioning and enforcing reliability guarantees via calibrated selective prediction. Our results demonstrate this combination is effective: SCARF reduces policy cost 11–15% while maintaining 83–88% coverage across multiple stressors. Our synthetic stressors and credit-risk focus limit generalizability. Natural distribution shifts, multiclass problems, and other domains require validation. Future work should extend SCARF to time-varying costs and explore learned (vs. specified) abstention strategies.

Table 2: Ablation of core SCARF components. Values are variant minus default across ten datasets and four stressors at $\alpha = 0.10$. Default: PolicyCost = 0.70, Coverage = 85.4%, KeptErr = 10.0%. † denotes $p < 0.05$.

Variant	**ΔCost**	**ΔCov.(pp)**	**ΔErr(pp)**
(default)	0.000	0.0	0.0
No selective calibration	$+0.210^{\dagger}$	$+12.7$	$+8.9^{\dagger}$
Identity preconditioner ($H = I$)	$+0.070^{\dagger}$	-2.1	0.0
No DIR-G augmentation	$+0.035^{\dagger}$	-0.8	0.0
No severity term	$+0.024^{\dagger}$	-1.5	0.0

References

N. C. Angelopoulos and S. Bates (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

N. C. Angelopoulos and S. Bates (2022). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning* 16(4):494–591.

P. L. Bartlett and M. H. Wegkamp (2008). Classification with a reject option using a hinge loss. *Journal of Machine Learning Research* 9:1823–1840.

A. Ben-Tal, D. Den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen (2013). Robust solutions of optimization problems affected by uncertain probabilities. *Management Science* 59(2):341–357.

R. Blaser and P. Fryzlewicz (2016). Random rotation ensembles. *Journal of Machine Learning Research* 17(4):1–26.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone (1984). *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA.

T. Chen and C. Guestrin (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

C. K. Chow (1957). An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers* (4):247–254.

C. K. Chow (1970). On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory* 16(1):41–46.

P. Domingos (1999). MetaCost: A general method for making classifiers cost-sensitive. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 155–164.

R. El-Yaniv and Y. Wiener (2010). On the foundations of noise-free selective classification. *Journal of Machine Learning Research* 11:1605–1641.

Y. Geifman and R. El-Yaniv (2017). Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems* 30.

S. K. Murthy, S. Kasif, and S. Salzberg (1993). OC1: A randomized algorithm for building oblique decision trees. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, 322–327.

H. Namkoong and J. C. Duchi (2017). Variance-based regularization with convex objectives. *Journal of Machine Learning Research* 18(1):2971–3007.

J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence (2009). *Dataset Shift in Machine Learning*. MIT Press.

J. R. Quinlan (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

H. Rahimian and S. Mehrotra (2019). Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*.

Rainforth, T., Wood, F. (2015). Canonical Correlation Forests. *arXiv preprint arXiv:1507.05444*.

S. Trivedi, J. Wang, S. Kpotufe and G. Shakhnarovich (2014). A consistent estimator of the expected gradient outerproduct. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 729–738.

S. Trivedi and J. Wang (2020). The expected Jacobian outerproduct: Theory and empirics. *arXiv preprint arXiv:2006.03550*.

J. J. Rodriguez and L. I. Kuncheva (2006). Rotation forest: A new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28(10):1619–1630.

D. J. Stekhoven and P. Bühlmann (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28(1):112–118.

V. Vovk, A. Gammerman, and G. Shafer (2005). *Algorithmic Learning in a Random World*. Springer.

B. Zadrozny and C. Elkan (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 694–699.

Related Work

Cost-sensitive trees and ensembles. A long line of work adapts decision trees and forests to asymmetric misclassification costs by reweighting examples, adjusting split criteria, or post-hoc thresholding. Classic systems (e.g., CART, C4.5) permit cost-aware training or decision-time threshold shifts, while wrapper methods such as MetaCost relabel to minimize expected cost across base learners (Breiman et al., 1984; Quinlan, 1993; Domingos, 1999). Cost-sensitive boosting variants similarly upweight costly errors. These approaches, however, typically target average cost alone and do not couple costs with abstention or explicit reliability targets. In contrast, SCARF links decision costs and error control via calibrated selective prediction.

Selective prediction (reject option) and conformal control. Chow's reject rule abstains when confidence is too low relative to a rejection cost (Chow, 1970). Modern formulations embed a reject option in the learning objective (e.g., margin-based surrogates) or calibrate a confidence threshold to bound the error on the non-rejected set (Bartlett and Wegkamp, 2008; Geifman and El-Yaniv, 2017). Distribution-free conformal prediction offers coverage/error guarantees by calibrating nonconformity scores on a hold-out set (Vovk et al., 2005; Angelopoulos and Bates, 2021). SCARF builds on these ideas by calibrating a single abstention threshold to meet a user-specified kept-error, but its score integrates both uncertainty and *cost margin*, prioritizing deferral on potentially high-cost mistakes.

Robustness to shift, noise, and missingness. Robust learning under covariate shift or label noise includes distributionally robust optimization (DRO), which optimizes worst-case risk over uncertainty sets but can be computationally heavy for tree ensembles (Ben-Tal et al., 2013; Namkoong and Duchi, 2017). Practical alternatives for trees include data augmentation, bagging, and noise-tolerant splitting/pruning; missing data is handled via surrogate splits or learned default directions, and imputation methods such as MissForest (Breiman et al., 1984; Chen and Guestrin, 2016; Stekhoven and Bühlmann, 2012). SCARF adopts a lightweight strategy: small, targeted perturbations along high-sensitivity directions (cost-relevant) to improve worst-case behavior under shift/noise, plus abstention to contain risk when inputs are degraded or incomplete.

Oblique trees and feature transformations. Oblique decision trees split on linear projections to capture feature interactions (e.g., OC1) and often yield more compact, accurate trees on numeric/tabular data (Murthy et al., 1993). Ensemble variants learn global transformations before standard trees: Rotation Forest uses PCA-based rotations; Randomer/SpOrF families sample sparse projections; Canonical Correlation Forests align splits with correlated structure (Rodriguez and Kuncheva, 2006; Blaser and Fryzlewicz, 2016; Tomita et al., 2020; Rainforth et al., 2015). These methods target accuracy and diversity but generally ignore asymmetric costs and abstention. SCARF differs by learning a *cost-sensitive spectral transform*: a single global rotation derived from class-probability sensitivities so that downstream (standard) trees implement oblique, cost-aligned boundaries without per-node optimization, while also integrating calibrated rejection and robustness augmentation.

Practical Considerations

Surrogate model for EJOP estimation. Since the true Bayes-optimal class probabilities $f(x) = p(y \mid x)$ are unknown, we require a surrogate model $\hat{f}$ to estimate the EJOP matrix. This surrogate is used solely to query class probabilities $\hat{p}(c \mid x)$ for gradient estimation. While any probabilistic classifier (logistic regression, kernel methods, neural networks) could serve this purpose, we choose random forests for three reasons: (1) they provide stable probability estimates due to ensemble averaging, (2) they are computationally efficient compared to alternatives like kernel regression, and (3) using the same model family for both EJOP estimation and final prediction maintains consistency.

Finite differences and non-differentiability. Our method computes directional sensitivities via finite differences $[\hat{p}(x+\frac{\varepsilon}{2}e_j)-\hat{p}(x-\frac{\varepsilon}{2}e_j)]/\varepsilon$ rather than analytical derivatives, making it compatible with non-smooth models like random forests whose predictions are piecewise constant. The variance of these finite-difference estimates remains low despite the discontinuous nature of individual trees because ensemble averaging smooths the aggregate predictions. The adaptive step size $\varepsilon_j = \alpha \cdot \mathrm{MAD}(X_{:j})/0.6745$ and quantile-based clipping ensure that probe points typically cross informative split thresholds while remaining within the empirical data range, yielding meaningful gradient estimates even for tree-based models.

Computational complexity Let d be the number of features, m the EJOP subsample size, and T the number of trees in the surrogate RF. Constructing $\widehat{H}_0$ requires $O(2mdT\log n)$ for centered finite differences plus $O(d^2)$ for forming the matrix. The estimation bound implies $O(m^{-1/2})$ concentration, so larger m reduces variance with diminishing returns. Our ablations (Table 6) show $m/n \in [0.3, 0.5]$ is sufficient, with $m = 0.5n$ showing negligible degradation $(+0.006)$. We recommend $m = \max\{0.3n, 20d\}$ when compute is limited, or $m = 0.5n$ for high-dimensional regimes ($d > 0.05n$). DIR-G reuses the same probing machinery, adding one extra data pass. After H is fixed, the final forest is trained once on $X\widehat{H}$ with standard cost. Overall, auxiliary overhead scales linearly in m and d.

Probability calibration. After training the final forest $\hat{h}$ on the transformed features $X\hat{H}$, we apply isotonic regression (Zadrozny and Elkan, 2002) to the predicted probabilities $\hat{h}(\Phi(x))$ using a held-out calibration set to ensure that predicted class probabilities are well-calibrated (i.e., $\hat{p}(y=1|x) \approx \Pr(Y=1|\hat{p}(y=1|x))$). This post-hoc recalibration step is applied before computing expected costs $e_j(x)$ (Section 3.7) and the selective prediction score $s(x)$ (Section 3.8). Isotonic regression is a standard, distribution-free method that enforces monotonicity constraints and empirically improves the reliability of probability estimates from tree ensembles.

Further Results

Overall policy cost under reliability stressors

Table 3: Per-class coverage at $\alpha = 0.10$. Class 0 = non-default (78%); Class 1 = default (22%). Ratio in parentheses.

	Covariate Shift		
Model	All	C0	C1
Logit	78.5	79.8	73.5 (.92)
RF	81.2	82.5	77.4 (.94)
XGBoost	83.8	84.9	80.8 (.95)
Rot.For.	84.1	85.0	80.8 (.95)
RerF	84.3	85.2	81.2 (.95)
HGBT	85.7	86.5	83.2 (.96)
SCARF	**87.1**	**87.8**	**84.3 (.96)**
	Label Noise		
Logit	71.3	72.1	70.0 (.97)
RF	74.6	76.3	70.2 (.92)
XGBoost	76.2	77.9	71.8 (.92)
Rot.For.	76.5	77.0	72.1 (.94)
RerF	76.8	77.5	72.0 (.93)
HGBT	78.1	79.6	73.9 (.93)
SCARF	**83.5**	**84.5**	**79.4 (.94)**

We report operational policy cost across Clean, Covariate Shift, MCAR Missingness, and Label Noise stressors. On i.i.d. test data, SCARF (0.28) performs comparably to the best booster HGBT (0.27), indicating that the oblique preconditioning and selective prediction do not hurt performance in benign settings. Under shifted feature distributions, SCARF attains cost 0.45, demonstrating that the spectral preconditioner together with gradient-aligned augmentation (DIR-G) maintain competitive performance under distribution shifts without degradation. With 15%/5% asymmetric label noise, SCARF's cost (1.00) is substantially lower than HGBT (1.15), highlighting the benefit of calibrated abstention and DIR-G under noisy labels. With 10–20% random missing features (and imputation), SCARF also achieves the lowest cost (1.08 vs. HGBT 1.22), indicating that the severity-aware scoring effectively defers on highly incomplete cases, reducing costly mistakes.

Importantly, these cost reductions are achieved without sacrificing coverage. Under covariate shift, SCARF maintains 87.1% coverage (higher than HGBT (85.7%)) while attaining 13% lower cost. Under label noise, SCARF retains 83.5% coverage compared to HGBT's 78.1%, indicating that the calibrated selective prediction defers more judiciously on high-cost errors rather than abstaining indiscriminately. Across all stressors, SCARF's coverage remains between 83–88%, demonstrating practical deployment utility: the majority of instances receive predictions while the target 10% kept-error constraint is satisfied. Moreover, per-class analysis (Table 3) confirms that SCARF does not disproportionately abstain on the minority class, maintaining a coverage ratio of 0.96 across stressors—the most balanced among all methods.

Ablation Study

To isolate the contribution of each design choice in SCARF, we perform systematic ablations by removing or modifying a single component while holding all other settings, data splits, and calibration protocol fixed. Table 4 isolates the contribution of each SCARF component via systematic ablation at $\alpha = 0.10$ across all four stressors. Default SCARF achieves mean policy cost = 0.70, coverage = 85.4%, kept-error = 10.0%, and training time = 12.3s per dataset; all deltas are relative to these values.

Removing selective calibration (+0.21 cost, +22.7pp coverage, +8.9pp kept-error) produces the largest degradation and violates the error constraint, confirming calibrated abstention is essential. Replacing the EJOP preconditioner with identity (+0.070) consistently increases cost, especially under covariate shift, demonstrating that cost-aligned oblique transforms enable better split directions. Omitting DIR-G (+0.035) or the severity term (+0.024) yields moderate increases under shift/noise. Implementation details (surrogate choice, finite-difference scheme, step size, subsample size $m/n \in [0.3, 0.5]$) have minimal impact. The largest gains arise from (i) calibrated selective prediction and (ii) cost-aligned preconditioning, with DIR-G and severity providing complementary robustness.

Table 4: Performance impact of ablating components. Values are variant minus default averaged across ten datasets and four stressors (Clean, Shift, Label Noise, MCAR) at kept-error $\alpha = 0.10$. Default baseline: PolicyCost = 0.70, Coverage = 85.4%, KeptErr = 10.0%, Time = 12.3s. † denotes $p < 0.05$ (Wilcoxon signed-rank test, Holm-corrected).

Variant	**ΔPolicyCost**	**ΔCov. (pp)**	**ΔKeptErr (pp)**	**Δt (s)**
(default)	0.000	0.0	0.0	0.00
Identity ($\widehat{H} = I$)	+0.070†	-2.1	0.0	-0.35
No selective calibration (forced)	+0.210†	$+22.7$	$+8.9^{\dagger}$	-0.05
No robustness augmentation	+0.035†	-0.8	0.0	-0.12
No severity term in score	+0.024†	-1.5	0.0	0.00
Surrogate: logistic (vs. small RF)	+0.006	-0.1	0.0	-0.08
Surrogate: tiny HGBT (vs. small RF)	+0.004	-0.1	0.0	-0.03
One-model OOB (RF $\rightarrow H \rightarrow$ RF)	+0.003	0.0	0.0	-0.06
Finite diff.: forward (vs. centered)	+0.011†	0.0	0.0	-0.01
Finite diff.: no clipping	+0.009	0.0	0.0	-0.01
Step: $\alpha_{\varepsilon} = 0.05$ (vs. 0.10)	+0.008	0.0	0.0	0.00
Step: $\alpha_{\varepsilon} = 0.20$ (vs. 0.10)	+0.013†	0.0	0.0	0.00
DIR-G: $\rho = 0.05$ (vs. 0.10)	+0.006	0.0	0.0	+0.01
DIR-G: $\rho = 0.20$ (vs. 0.10)	+0.012†	0.0	0.0	+0.02
Subsample $m = 0.5n$ (vs. n)	+0.006	0.0	0.0	-0.28
Subsample $m = 0.3n$	+0.018†	0.0	0.0	-0.55
No γI_d in $\widehat{H}$	+0.004	0.0	0.0	0.00
No trace normalization of $\widehat{H}$	+0.005	0.0	0.0	+0.01

Proofs and Technical Lemmas

A.1 Local Adversarial Upper Bound

Theorem 1 (Restatement of Thm. 3.1: local adversarial upper bound and near–tightness)**.** *Let $\ell : \mathbb{R}^d \to \mathbb{R}$ be differentiable and L–smooth in x (i.e., $\nabla_x \ell$ is L–Lipschitz). For $\rho > 0$, define*

$$\ell_{\max}(x;\rho) \;=\; \sup_{\|\Delta\|_2 \le \rho} \; \ell(x+\Delta).$$

Then

$$\ell_{\max}(x;\rho) \;\le\; \ell(x) \;+\; \rho\,\|\nabla_x \ell(x)\|_2 \;+\; \tfrac{L}{2}\,\rho^2. \tag{1}$$

Moreover, with $u^\star(x) = \nabla_x \ell(x)/\|\nabla_x \ell(x)\|_2$ whenever $\nabla_x \ell(x) \neq 0$,

$$\ell(x+\rho u^\star) \;\ge\; \ell(x) \;+\; \rho\,\|\nabla_x \ell(x)\|_2 \;-\; \tfrac{L}{2}\,\rho^2, \tag{2}$$

so the bound is first–order tight up to $O(\rho^2)$. If $\nabla_x \ell(x) = 0$, then $\ell_{\max}(x;\rho) \le \ell(x) + \frac{L}{2}\rho^2$.

Proof. L–smoothness implies for any $\Delta \in \mathbb{R}^d$,

$$\ell(x+\Delta) \;\le\; \ell(x) \;+\; \nabla_x \ell(x)^\top \Delta \;+\; \tfrac{L}{2}\,\|\Delta\|_2^2. \tag{3}$$

(This follows from Taylor's theorem with integral remainder or from the Lipschitz property of $\nabla_x \ell$.)

For any Δ with $\|\Delta\|_2 \le \rho$,

$$\nabla_x \ell(x)^\top \Delta \;\le\; \|\nabla_x \ell(x)\|_2\,\|\Delta\|_2 \;\le\; \rho\,\|\nabla_x \ell(x)\|_2,$$

by Cauchy–Schwarz, and also $\frac{L}{2}\|\Delta\|_2^2 \le \frac{L}{2}\rho^2$.

Combining Step 1 and Step 2 gives, for every admissible Δ,

$$\ell(x+\Delta) \;\le\; \ell(x) \;+\; \rho\,\|\nabla_x \ell(x)\|_2 \;+\; \tfrac{L}{2}\rho^2.$$

Taking $\sup_{\|\Delta\| \le \rho}$ yields (1).

Let $u^\star = \nabla_x \ell(x)/\|\nabla_x \ell(x)\|_2$ when $\nabla_x \ell(x) \neq 0$; otherwise the statement is trivial. Apply Taylor's theorem along the ray $x + \rho u^\star$:

$$\ell(x+\rho u^\star) \;=\; \ell(x) \;+\; \rho\,\nabla_x \ell(x)^\top u^\star \;+\; \tfrac{\rho^2}{2}\,(u^\star)^\top \nabla_x^2 \ell(x+\theta\rho u^\star)\,u^\star$$

for some $\theta \in (0,1)$. Since $\|\nabla_x^2 \ell(\cdot)\|_{\mathrm{op}} \le L$ a.e. under L–smoothness,

$$\left|\tfrac{\rho^2}{2}\,(u^\star)^\top \nabla_x^2 \ell(\cdot)\,u^\star\right| \;\le\; \tfrac{L}{2}\rho^2.$$

Using $\nabla_x \ell(x)^\top u^\star = \|\nabla_x \ell(x)\|_2$ gives (2).

If $\nabla_x \ell(x) = 0$, (3) reduces to $\ell(x+\Delta) \le \ell(x) + \frac{L}{2}\|\Delta\|_2^2$, and the supremum over $\|\Delta\| \le \rho$ gives $\ell_{\max}(x;\rho) \le \ell(x) + \frac{L}{2}\rho^2$. □

Remark 1 (General norms)**.** *The argument extends verbatim to any norm $\|\cdot\|$ with dual norm $\|\cdot\|_*$. If $\nabla_x \ell$ is L–Lipschitz w.r.t. $\|\cdot\|$ (i.e., $\|\nabla_x \ell(x) - \nabla_x \ell(y)\|_* \le L\,\|x-y\|$), then*

$$\ell_{\max}(x;\rho) \;\le\; \ell(x) \;+\; \rho\,\|\nabla_x \ell(x)\|_* \;+\; \tfrac{L}{2}\rho^2,$$

and the near–tightness lower bound holds along $u^\star \in \arg\max_{\|u\| \le 1} \nabla_x \ell(x)^\top u$, where the maximizer satisfies $\nabla_x \ell(x)^\top u^\star = \|\nabla_x \ell(x)\|_*$.

A.2 Effect of Gradient-Aligned Augmentation (DIR-G)

We analyze the symmetric augmentations $x^{(\pm)} = x \pm \rho u(x)$ with

$$u(x) \;=\; \frac{\nabla_x \ell(x;\theta)}{\|\nabla_x \ell(x;\theta)\|_2} \qquad \text{(when } \nabla_x \ell(x;\theta) \neq 0\text{; any fixed unit } u \text{ otherwise).}$$

Assume throughout that for each fixed θ, $\ell(\cdot;\theta)$ is L–smooth in x (i.e., $\nabla_x \ell$ is L–Lipschitz).

Define the augmented objective

$$\mathcal{L}_{\mathrm{aug}}(\theta) \;=\; \frac{1}{2}\,\mathbb{E}\big[\ell(x+\rho u(x);\theta) + \ell(x-\rho u(x);\theta)\big],$$

where the expectation is over the data distribution (and any sampling in u if approximated).

Step 1: Second-order expansion and cancellation of odd terms. Fix x, θ and write the second-order Taylor expansion in the direction u:

$$\ell(x+\rho u;\theta) = \ell(x;\theta) + \rho\,\nabla_x \ell(x;\theta)^\top u + \frac{\rho^2}{2}\,u^\top \nabla_x^2 \ell(x+\xi_+\rho u;\theta)\,u,$$
$$\ell(x-\rho u;\theta) = \ell(x;\theta) - \rho\,\nabla_x \ell(x;\theta)^\top u + \frac{\rho^2}{2}\,u^\top \nabla_x^2 \ell(x+\xi_-\rho u;\theta)\,u,$$

for some $\xi_+, \xi_- \in (0,1)$ (by Taylor with remainder). Averaging,

$$\frac{\ell(x+\rho u;\theta)+\ell(x-\rho u;\theta)}{2} = \ell(x;\theta) + \frac{\rho^2}{4}\,u^\top\big(\nabla_x^2\ell(x+\xi_+\rho u;\theta) + \nabla_x^2\ell(x+\xi_-\rho u;\theta)\big)u. \quad (4)$$

By L–smoothness, $\|\nabla_x^2 \ell(\cdot;\theta)\|_{\mathrm{op}} \le L$ almost everywhere, hence

$$\Big|\frac{\ell(x+\rho u;\theta)+\ell(x-\rho u;\theta)}{2} - \ell(x;\theta)\Big| \;\le\; \frac{L}{2}\,\rho^2. \quad (5)$$

Taking expectations,

$$\mathcal{L}_{\mathrm{aug}}(\theta) = \mathbb{E}[\ell(x;\theta)] + O(\rho^2), \qquad \big|\mathcal{L}_{\mathrm{aug}}(\theta) - \mathbb{E}[\ell(x;\theta)]\big| \le \tfrac{L}{2}\rho^2.$$

Recall $\ell_{\max}(x;\rho) = \sup_{\|\Delta\|_2 \le \rho} \ell(x+\Delta;\theta)$. The standard L–smooth bound gives

$$\ell_{\max}(x;\rho) \;\le\; \ell(x;\theta) + \rho\,\|\nabla_x \ell(x;\theta)\|_2 + \tfrac{L}{2}\rho^2. \quad (6)$$

Taking expectations:

$$\mathbb{E}[\ell_{\max}(x;\rho)] \;\le\; \mathbb{E}[\ell(x;\theta)] + \rho\,\mathbb{E}\big[\|\nabla_x \ell(x;\theta)\|_2\big] + \tfrac{L}{2}\rho^2.$$

The lower Taylor bound along $u = \nabla_x \ell / \|\nabla_x \ell\|$ reads

$$\ell(x+\rho u;\theta) \;\ge\; \ell(x;\theta) + \rho\,\|\nabla_x \ell(x;\theta)\|_2 - \tfrac{L}{2}\rho^2. \quad (7)$$

Rearranging,

$$\|\nabla_x \ell(x;\theta)\|_2 \;\le\; \frac{\ell(x+\rho u;\theta) - \ell(x;\theta)}{\rho} + \frac{L}{2}\rho. \quad (8)$$

Taking expectations and using (5) (together with $\ell(x+\rho u)$ controlled by the symmetric average),

$$\mathbb{E}\big[\|\nabla_x \ell(x;\theta)\|_2\big] \;\le\; \frac{\mathbb{E}\big[\ell(x+\rho u;\theta) - \ell(x;\theta)\big]}{\rho} + \frac{L}{2}\rho \;\lesssim\; \frac{\mathcal{L}_{\mathrm{aug}}(\theta) - \mathbb{E}[\ell(x;\theta)]}{\rho} + \frac{L}{2}\rho \;=\; O(\rho).$$

Hence, driving $\mathcal{L}_{\mathrm{aug}}(\theta)$ close to $\mathbb{E}[\ell(x;\theta)]$ forces the expected input-gradient norm to scale as $O(\rho)$.

For completeness, expand the parameter gradient (using mixed partials):

$$\nabla_\theta \ell(x \pm \rho u;\theta) = \nabla_\theta \ell(x;\theta) \;\pm\; \rho\left[\nabla_{x,\theta}^2 \ell(x;\theta)\,u\right] \;+\; \frac{\rho^2}{2}\left[u^\top \nabla_{x,x,\theta}^3 \ell(x;\theta)\,u\right] \;+\; O(\rho^3),$$
$$\Rightarrow \quad \nabla_\theta \mathcal{L}_{\mathrm{aug}}(\theta) = \mathbb{E}\big[\nabla_\theta \ell(x;\theta)\big] \;+\; \frac{\rho^2}{2}\,\mathbb{E}\big[u^\top \nabla_{x,x,\theta}^3 \ell(x;\theta)\,u\big] \;+\; O(\rho^4).$$

Thus training with $\mathcal{L}_{\mathrm{aug}}$ is equivalent to training on $\mathbb{E}[\ell(x;\theta)]$ plus a curvature-weighted regularizer that penalizes sharpness along u; this is consistent with the $O(\rho)$ control of $\mathbb{E}\|\nabla_x \ell\|_2$ above.

Conclusion. Combining (6) with the $O(\rho)$ bound on $\mathbb{E}\|\nabla_x \ell\|_2$ yields

$$\mathbb{E}[\ell_{\max}(x;\rho)] \;\le\; \mathbb{E}[\ell(x;\theta)] + O(\rho^2),$$

so symmetric gradient-aligned augmentation reduces the first-order adversarial term and improves robustness at the ρ-scale while only perturbing the nominal objective at order ρ^2.

Remark 2. *In practice we use a finite-difference proxy $g(x)$ for $\nabla \ell(x)$ (Section 3.10 in the main text), setting $u(x) = g(x)/\|g(x)\|_2$. The analysis extends by replacing $\nabla \ell(x)$ with $g(x)$ and adding the proxy error from the Lemma below.*

A.3 Kept-Error Control via Clopper–Pearson

Setup. Let $C = \{(x_i, y_i)\}_{i=1}^n$ be an *independent* calibration set (disjoint from training and test). Let $s : \mathcal{X} \to \mathbb{R}$ be a confidence score where larger is better (e.g., calibrated class probability or margin). For a threshold $\tau \in \mathbb{R}$ define the kept set

$$\mathcal{K}(\tau) \;=\; \{i \in \{1, \ldots, n\} \;:\; s(x_i) \geq \tau\}, \quad n_{\text{keep}}(\tau) \;=\; |\mathcal{K}(\tau)|.$$

Let $M_i = \mathbf{1}\{\hat{y}(x_i) \neq y_i\}$ be the misclassification indicator and

$$K(\tau) \;=\; \sum_{i \in \mathcal{K}(\tau)} M_i, \qquad \widehat{R}_{\text{keep}}(\tau) \;=\; \frac{K(\tau)}{n_{\text{keep}}(\tau)} \text{ (with the convention } 0/0 = 0).$$

The (population) *kept error* at τ is

$$R_{\text{keep}}(\tau) \;=\; \Pr\big(\hat{y}(X) \neq Y \;\big|\; s(X) \geq \tau\big).$$

Modeling assumption (exchangeability on the calibration split). Under i.i.d. sampling and deterministic tie-breaking, conditional on $\mathcal{K}(\tau)$ we have

$$K(\tau) \;\sim\; \text{Binomial}\big(n_{\text{keep}}(\tau),\; \theta(\tau)\big), \quad \text{where} \quad \theta(\tau) = R_{\text{keep}}(\tau).$$

Intuitively, among the $n_{\text{keep}}(\tau)$ kept points, each is an independent Bernoulli($\theta(\tau)$) "error trial."

Clopper–Pearson (one-sided upper) bound at fixed τ. For integers $0 \leq k \leq n$, the one-sided $(1 - \delta)$ Clopper–Pearson upper bound is

$$\text{CP}^+(k, n, \delta) \;=\; \text{Beta}^{-1}\big(1 - \delta;\; k + 1,\; n - k\big),$$

i.e., the $(1 - \delta)$ quantile of a $\text{Beta}(k + 1, n - k)$ distribution. It is the exact test inversion of the binomial tail:

$$\Pr\big(K \leq k \;\big|\; \theta = \text{CP}^+(k, n, \delta)\big) \;=\; 1 - \delta.$$

Therefore, for any *fixed* τ,

$$\Pr\big(\; \theta(\tau) \;\leq\; \text{CP}^+\big(K(\tau),\, n_{\text{keep}}(\tau),\, \delta\big) \;\big) \;\geq\; 1 - \delta. \tag{9}$$

Threshold selection on a finite grid. Let $\mathcal{T} = \{\tau_1 < \tau_2 < \cdots < \tau_m\}$ be a *finite* grid of thresholds (e.g., unique score values or a fixed quantile grid). Define the data-driven choice

$$\tau^\star \;=\; \inf\big\{\tau \in \mathcal{T} :\; \text{CP}^+\big(K(\tau), n_{\text{keep}}(\tau), \delta'\big) \leq \alpha\big\},$$

with per-threshold level $\delta' = \delta/m$ (Bonferroni correction). If the set is empty, declare "abstain-all" (or report failure to certify).

Proposition 1 (Finite-sample kept-error control with grid search)**.** *With the construction above,*

$$\Pr\big(\; R_{\text{keep}}(\tau^\star) \;\leq\; \alpha \;\big) \;\geq\; 1 - \delta.$$

Proof. For each $\tau \in \mathcal{T}$, by (9) with δ' we have

$$\Pr\big(\; \theta(\tau) \;\leq\; \text{CP}^+\big(K(\tau), n_{\text{keep}}(\tau), \delta'\big) \;\big) \;\geq\; 1 - \delta'.$$

Applying the union bound over m thresholds,

$$\Pr\big(\; \forall \tau \in \mathcal{T} :\; \theta(\tau) \;\leq\; \text{CP}^+\big(K(\tau), n_{\text{keep}}(\tau), \delta'\big) \;\big) \;\geq\; 1 - m\delta' \;=\; 1 - \delta.$$

On that high-probability event, in particular at $\tau^\star$ we have $\theta(\tau^\star) \leq \text{CP}^+(K(\tau^\star), n_{\text{keep}}(\tau^\star), \delta') \leq \alpha$ by definition of $\tau^\star$. Hence $R_{\text{keep}}(\tau^\star) = \theta(\tau^\star) \leq \alpha$ with probability at least $1 - \delta$. □

A.4 EJOP via Finite Differences

Objects and estimator. Let $p^\star(x) \in [0,1]^K$ denote the Bayes class probabilities, with coordinates $p^\star_c(x)$. Define the *EJOP* matrix

$$H_0 \;\coloneqq\; \mathbb{E}\left[\sum_{c=1}^{K} \nabla p^\star_c(X)\, \nabla p^\star_c(X)^\top\right] \in \mathbb{R}^{d\times d}.$$

We estimate $\nabla p^\star_c$ by centered, coordinatewise finite differences applied to a surrogate $\tilde{p}_c$: for $E = \mathrm{diag}(\varepsilon_1,\dots,\varepsilon_d)$ and unit coordinate vectors e_j,

$$g_j(x;c) \;=\; \frac{\tilde{p}_c(x+\varepsilon_j e_j) - \tilde{p}_c(x-\varepsilon_j e_j)}{2\varepsilon_j}, \qquad g(x;c) \;=\; \big(g_1,\dots,g_d\big)^\top.$$

Given a subsample $\{x_i\}_{i=1}^m$, the plug-in estimator is

$$\widehat{H}_0 \;\coloneqq\; \frac{1}{m}\sum_{i=1}^{m}\sum_{c=1}^{K} g(x_i;c)\, g(x_i;c)^\top.$$

Assumptions. We assume:

(A1) **Smoothness.** For each c, $p^\star_c$ is twice differentiable a.e. with L-Lipschitz gradient: $\|\nabla p^\star_c(x) - \nabla p^\star_c(y)\|_2 \le L\|x-y\|_2$. (This implies $\|\nabla^2 p^\star_c(x)\|_{\mathrm{op}} \le L$ a.e.)

(A2) **Surrogate accuracy.** $\|\tilde{p}_c - p^\star_c\|_\infty \le \eta$ for all c.

(A3) **Bounded gradients.** $\|\nabla p^\star_c(x)\|_2 \le B_{\max}$ a.s. for all c.

Pointwise finite-difference error (discretization + surrogate). Write $\delta_c = \tilde{p}_c - p^\star_c$ and decompose

$$g_j(x;c) \;=\; \underbrace{\frac{p^\star_c(x+\varepsilon_j e_j) - p^\star_c(x-\varepsilon_j e_j)}{2\varepsilon_j}}_{\text{centered difference on } p^\star_c} \;+\; \underbrace{\frac{\delta_c(x+\varepsilon_j e_j) - \delta_c(x-\varepsilon_j e_j)}{2\varepsilon_j}}_{\text{surrogate term}}.$$

Surrogate term. Since $|\delta_c(\cdot)| \le \eta$ pointwise,

$$\left|\frac{\delta_c(x+\varepsilon_j e_j) - \delta_c(x-\varepsilon_j e_j)}{2\varepsilon_j}\right| \;\le\; \frac{|\delta_c(x+\varepsilon_j e_j)| + |\delta_c(x-\varepsilon_j e_j)|}{2\varepsilon_j} \;\le\; \frac{\eta}{\varepsilon_j}.$$

Discretization term. By Taylor's theorem with remainder along $t \mapsto p^\star_c(x+te_j)$ and (A1),

$$\frac{p^\star_c(x+\varepsilon_j e_j) - p^\star_c(x-\varepsilon_j e_j)}{2\varepsilon_j} \;=\; \partial_{x_j} p^\star_c(x) \;+\; r_j(x;c),$$

with

$$|r_j(x;c)| \;\le\; \frac{1}{2}\int_{-1}^{1} (1-|t|)\, \|\nabla^2 p^\star_c(x+t\varepsilon_j e_j)\|_{\mathrm{op}}\, dt \;\cdot\; \varepsilon_j \;\le\; \frac{L}{2}\,\varepsilon_j.$$

(If $p^\star_c$ is C^3 with $\sup|\partial^3_{x_j} p^\star_c| \le T_j$, one may strengthen to $|r_j| \le \frac{T_j}{6}\varepsilon_j^2$; we keep the $O(\varepsilon_j)$ bound implied by L-smoothness.)

Collecting both contributions and stacking over j yields the vector error bound

$$\|\, g(x;c) - \nabla p^\star_c(x)\,\|_2 \;\le\; C_1 \max_j \varepsilon_j \;+\; C_2\,\frac{\eta}{\min_j \varepsilon_j}, \qquad C_1 = \tfrac{L}{2}\sqrt{d}, \;\; C_2 = \sqrt{d}. \tag{10}$$

Outer-product perturbation from gradient error. For any $a, b \in \mathbb{R}^d$,

$$aa^\top - bb^\top \;=\; (a-b)b^\top + a(a-b)^\top,$$

so

$$\|\, aa^\top - bb^\top \,\|_{\mathrm{op}} \;\le\; \|a-b\|_2\, \|b\|_2 \;+\; \|a\|_2\, \|a-b\|_2 \;\le\; (\,\|a\|_2 + \|b\|_2\,)\, \|a-b\|_2. \tag{11}$$

Apply (11) with $a = g(x;c)$ and $b = \nabla p_c^\star(x)$ and use (A3):

$$\|\, g(x;c)g(x;c)^\top - \nabla p_c^\star(x)\nabla p_c^\star(x)^\top \,\|_{\mathrm{op}} \;\le\; \big(\, \|g(x;c)\|_2 + \|\nabla p_c^\star(x)\|_2 \,\big)\, \|g(x;c) - \nabla p_c^\star(x)\|_2.$$

Since $\|g(x;c)\|_2 \le \|\nabla p_c^\star(x)\|_2 + \|g - \nabla p_c^\star\|_2 \le B_{\max} + \Delta(x;c)$ with $\Delta(x;c) := \|g - \nabla p_c^\star\|_2$, we obtain

$$\|\, gg^\top - \nabla p_c^\star \nabla p_c^{\star\top} \,\|_{\mathrm{op}} \;\le\; \big(2B_{\max} + \Delta(x;c)\big)\,\Delta(x;c) \;\le\; 2B_{\max}\Delta(x;c) + \Delta(x;c)^2. \tag{12}$$

Summing over c and averaging over the sample $\{x_i\}$ gives the *bias* term

$$\begin{aligned} \mathsf{Bias} \;:=\; & \left\| \frac{1}{m}\sum_{i=1}^{m}\sum_{c=1}^{K}\Big(g_i(c)g_i(c)^\top - \nabla p_c^\star(x_i)\nabla p_c^\star(x_i)^\top \Big) \right\|_{\mathrm{op}} \\ \le\; & \frac{1}{m}\sum_{i=1}^{m}\sum_{c=1}^{K}\big(2B_{\max}\Delta_i(c) + \Delta_i(c)^2\big), \qquad \Delta_i(c) \;:=\; \|g(x_i;c) - \nabla p_c^\star(x_i)\|_2. \end{aligned} \tag{13}$$

By (10), deterministically $\Delta_i(c) \le \delta_g$ with

$$\delta_g \;:=\; C_1 \max_j \varepsilon_j \;+\; C_2\, \frac{\eta}{\min_j \varepsilon_j},$$

hence

$$\mathsf{Bias} \;\le\; K\big(2B_{\max}\delta_g + \delta_g^2\big). \tag{14}$$

Concentration for the sampling fluctuation (matrix Bernstein). Define the *oracle* summands

$$Y_i \;:=\; \sum_{c=1}^{K} \nabla p_c^\star(x_i)\, \nabla p_c^\star(x_i)^\top, \qquad \mathbb{E}[Y_i] = H_0,$$

and centered matrices $Z_i := Y_i - \mathbb{E}[Y_i]$ (self-adjoint, independent). Each Y_i satisfies

$$\|Y_i\|_{\mathrm{op}} \;\le\; \sum_{c=1}^{K} \|\nabla p_c^\star(x_i)\|_2^2 \;\le\; KB_{\max}^2 \quad\Rightarrow\quad \|Z_i\|_{\mathrm{op}} \;\le\; \|Y_i\|_{\mathrm{op}} + \|H_0\|_{\mathrm{op}} \;\le\; 2KB_{\max}^2 \;=:\; R.$$

For the variance proxy,

$$\left\| \sum_{i=1}^{m} \mathbb{E}[Z_i^2] \right\|_{\mathrm{op}} \;\le\; \sum_{i=1}^{m} \mathbb{E}\|Z_i\|_{\mathrm{op}}^2 \;\le\; mR^2 \;=:\; \sigma^2.$$

Tropp's matrix Bernstein inequality yields, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\left\| \frac{1}{m}\sum_{i=1}^{m} Z_i \right\|_{\mathrm{op}} \;\le\; \underbrace{\frac{\sqrt{2\sigma^2 \log(2d/\delta)}}{m}}_{\text{(A)}} + \underbrace{\frac{2R\,\log(2d/\delta)}{3m}}_{\text{(B)}} \;=\; 2KB_{\max}^2\sqrt{\frac{2\log(2d/\delta)}{m}} + \frac{4KB_{\max}^2}{3}\,\frac{\log(2d/\delta)}{m}. \tag{15}$$

We denote the RHS by $\mathsf{Conc}(m,\delta)$.

Putting it together (bias+concentration). Add and subtract the oracle mean:

$$\widehat{H}_0 - H_0 \;=\; \underbrace{\Big(\widehat{H}_0 - \tfrac{1}{m}\sum_i Y_i\Big)}_{\text{bias from } g \text{ vs. } \nabla p^\star} + \underbrace{\Big(\tfrac{1}{m}\sum_i Y_i - H_0\Big)}_{\text{sampling fluctuation}}.$$

By (14) and (15), with probability at least $1-\delta$,

$$\|\, \widehat{H}_0 - H_0 \,\|_{\mathrm{op}} \;\le\; K\big(2B_{\max}\delta_g + \delta_g^2\big) \;+\; 2KB_{\max}^2\sqrt{\frac{2\log(2d/\delta)}{m}} \;+\; \frac{4KB_{\max}^2}{3}\,\frac{\log(2d/\delta)}{m}. \tag{16}$$

Remark 3 (Balancing ε_j against η)**.** *The surrogate/FD tradeoff in $\delta_g = C_1 \max_j \varepsilon_j + C_2\,\eta/\min_j \varepsilon_j$ suggests choosing (roughly) $\varepsilon_j^\star \propto \sqrt{\eta}$ to balance the two contributions (if one uses a common step). Our adaptive rule $\varepsilon_j = \alpha_\varepsilon \cdot \mathrm{MAD}(X_{:j})/0.6745$ rescales steps to the feature scale while letting α_ε tune the tradeoff.*

Remark 4 (Optional: ridge + trace normalization). *If the preconditioner used in the method is $\widehat{H} = \frac{\widehat{H}_0+\gamma I}{\mathrm{tr}(\widehat{H}_0+\gamma I)}$ (and $H = \frac{H_0+\gamma I}{\mathrm{tr}(H_0+\gamma I)}$ analogously), then for $T_A := \mathrm{tr}(A+\gamma I)$,*

$$\|\widehat{H}-H\|_{\mathrm{op}} \;\le\; \frac{\|\widehat{H}_0-H_0\|_{\mathrm{op}}}{T_{H_0}} + \frac{\|\widehat{H}_0\|_{\mathrm{op}}}{T_{\widehat{H}_0}\,T_{H_0}}\left|\mathrm{tr}(\widehat{H}_0-H_0)\right| \;\le\; \left(\frac{1}{T_{\min}}+\frac{d\,\|\widehat{H}_0\|_{\mathrm{op}}}{T_{\min}^2}\right)\|\widehat{H}_0-H_0\|_{\mathrm{op}},$$

where $T_{\min} = \min\{T_{\widehat{H}_0}, T_{H_0}\} \ge \gamma d$. Thus (16) *transfers to $\|\widehat{H}-H\|_{\mathrm{op}}$ up to a benign factor depending on γ and d (ridge and normalization stabilize the map $A \mapsto (A+\gamma I)/\mathrm{tr}(A+\gamma I)$).*

Corollary 1 (DIR-G proxy error). *Let $\ell(x) = \sum_{c=1}^{K} w_c\, p_c^\star(x)$ with nonnegative weights w_c (induced by the cost-aware rule), and define $\tilde{g}(x) = \sum_c w_c\, g(x;c)$. Then by triangle inequality and* (10),

$$\|\,\tilde{g}(x) - \nabla\ell(x)\,\|_2 \;\le\; \sum_c w_c\,\|g(x;c) - \nabla p_c^\star(x)\|_2 \;\le\; \|w\|_1\left(C_1 \max_j \varepsilon_j + C_2\,\frac{\eta}{\min_j \varepsilon_j}\right).$$

Algorithmic Specification (Pseudocode)

B.1 End-to-End SCARF Pipeline

Algorithm 1 SCARF: Training and Inference

```
Require: Training data $D_{\mathrm{tr}} = \{(x_i, y_i)\}$, calibration data $D_{\mathrm{cal}}$, classes $\{1,\dots,K\}$, cost matrix
    $C \in \mathbb{R}^{K\times K}$, abstain cost $c_{\mathrm{abs}}$, target kept-error $\alpha$, confidence $\delta$, FD scale $\alpha_\varepsilon$, EJOP subsample
    size $m$, ridge $\gamma > 0$, DIR-G strength $\rho \ge 0$, forest hyperparams RF_cfg
Ensure: Trained forest $F$, feature map $\Phi(x)$, confidence score $s(x)$, certified threshold $\tau^\star$
Robust step sizes: For each feature $j$, set $\varepsilon_j \leftarrow \alpha_\varepsilon \cdot \mathrm{MAD}(X_{:j})/0.6745$.
Train a light surrogate $\tilde{p}(x)$ (e.g., multinomial logistic or tiny HGBT) on $D_{\mathrm{tr}}$.
Estimate EJOP: $\widehat{H}_0 \leftarrow$ EstimateEJOP$(\tilde{p}, \{\varepsilon_j\}, D_{\mathrm{tr}}, m)$    ▷ Alg. 2
Preconditioner: $\widehat{H} \leftarrow (\widehat{H}_0 + \gamma I)/\mathrm{tr}(\widehat{H}_0 + \gamma I)$
Feature map: define $\Phi(x) \leftarrow \widehat{H}\,x$
if $\rho > 0$ then
    DIR-G augmentation: $D_{\mathrm{tr}}^{+} \leftarrow$ DirGAugment$(\tilde{p}, \{\varepsilon_j\}, D_{\mathrm{tr}}, \rho, C)$    ▷ Alg. 3
else
    $D_{\mathrm{tr}}^{+} \leftarrow D_{\mathrm{tr}}$
end if
Train forest: $F \leftarrow$ TrainForest$\big(\{(\Phi(x), y) \in D_{\mathrm{tr}}^{+}\}$, RF_cfg$\big)$
Calibrate & score on $D_{\mathrm{cal}}$: $(\varphi, s(\cdot)) \leftarrow$ CalibrateAndScore$(F, \Phi, D_{\mathrm{cal}})$    ▷ Alg. 4
CP threshold search: $\tau^\star \leftarrow$ CPThreshold$(s, F, \Phi, D_{\mathrm{cal}}, \alpha, \delta)$    ▷ Alg. 5
Inference rule (test point $x$):
    compute $\hat{y}(x) \leftarrow \arg\max_k F_k(\Phi(x))$, confidence $s(x)$
    if $s(x) \ge \tau^\star$ then predict $\hat{y}(x)$ else abstain (cost $c_{\mathrm{abs}}$)
```

B.2 EJOP via Centered Finite Differences

Algorithm 2 EstimateEJOP($\tilde{p}, \{\varepsilon_j\}, D_{\mathrm{tr}}, m$)

```
Draw a uniform subsample $S = \{x_i\}_{i=1}^{m}$ from $D_{\mathrm{tr}}$ (without labels).
Initialize $\widehat{H}_0 \leftarrow 0_{d\times d}$.
for each $x \in S$ do
    for $c = 1$ to $K$ do
        for $j = 1$ to $d$ do
            $g_j(x;c) \leftarrow \dfrac{\tilde{p}_c(x + \varepsilon_j e_j) - \tilde{p}_c(x - \varepsilon_j e_j)}{2\varepsilon_j}$
        end for
        $g(x;c) \leftarrow (g_1, \ldots, g_d)^\top$
        $\widehat{H}_0 \leftarrow \widehat{H}_0 + g(x;c)\, g(x;c)^\top$
    end for
end for
return $\widehat{H}_0/m$
```

B.3 Gradient-Aligned Data Augmentation (DIR-G)

Algorithm 3 DirGAugment($\tilde{p}, \{\varepsilon_j\}, D_{\mathrm{tr}}, \rho, C$)

```
Define class weights $w_c \geq 0$ from the cost matrix $C$ (e.g., $w_c = \sum_{c'} C_{c',c}$ or task-specific).
$D^+ \leftarrow \emptyset$
for each $(x, y) \in D_{\mathrm{tr}}$ do
    for $c = 1$ to $K$ do
        for $j = 1$ to $d$ do
            $g_j(x;c) \leftarrow \dfrac{\tilde{p}_c(x + \varepsilon_j e_j) - \tilde{p}_c(x - \varepsilon_j e_j)}{2\varepsilon_j}$
        end for
    end for
    $g(x) \leftarrow \sum_{c=1}^{K} w_c\, g(x;c)$                      ▷ cost-aware surrogate gradient
    $u(x) \leftarrow g(x)/\max(\|g(x)\|_2,\ 10^{-12})$          ▷ unit adversarial direction (safe divide)
    $x^{(+)} \leftarrow x + \rho\, u(x), \quad x^{(-)} \leftarrow x - \rho\, u(x)$
    $D^+ \leftarrow D^+ \cup \{(x, y), (x^{(+)}, y), (x^{(-)}, y)\}$
end for
return $D^+$
```

B.4 Calibration and Confidence Score

Algorithm 4 CalibrateAndScore($F, \Phi, D_{\mathrm{cal}}$)

```
Compute raw class probabilities on $D_{\mathrm{cal}}$: $\hat{p}_k^{\mathrm{raw}}(x) = F_k(\Phi(x))$.
Define raw confidence $r(x) = \max_k \hat{p}_k^{\mathrm{raw}}(x)$ and correctness labels $z(x) =$
    $\mathbf{1}\{\arg\max_k \hat{p}_k^{\mathrm{raw}}(x) = y\}$.
Fit isotonic regression $\varphi : [0,1] \to [0,1]$ on pairs $\{(r(x), z(x)) : (x, y) \in D_{\mathrm{cal}}\}$.
Define calibrated confidence score $s(x) \leftarrow \varphi(r(x))$ for any input $x$.
return $(\varphi, s(\cdot))$
```

B.5 Clopper–Pearson Threshold Selection

Algorithm 5 CPThreshold$(s, F, \Phi, D_{\text{cal}}, \alpha, \delta)$

```
Form a finite grid $\mathcal{T}$ over $[0,1]$ (e.g., 200 quantiles of $\{s(x) : (x,y) \in D_{\text{cal}}\}$).
Set per-grid level $\delta' \leftarrow \delta/|\mathcal{T}|$.
for each $\tau \in \mathcal{T}$ do
    $\mathcal{K}(\tau) \leftarrow \{(x,y) \in D_{\text{cal}} :\ s(x) \ge \tau\}$
    $n_{\text{keep}} \leftarrow |\mathcal{K}(\tau)|$
    $K \leftarrow \sum_{(x,y)\in\mathcal{K}(\tau)} \mathbf{1}\{\arg\max_k F_k(\Phi(x)) \neq y\}$
    $\text{UB} \leftarrow \text{CP}^{+}(K, n_{\text{keep}}, \delta')$        ▷ $\text{UB} = \text{Beta}^{-1}(1-\delta'; K+1, n_{\text{keep}} - K)$
    if $\text{UB} \le \alpha$ then
        record $\tau$ as feasible
    end if
end for
if no feasible $\tau$ then
    return $\tau^\star \leftarrow +\infty$        ▷ abstain-all or relax $(\alpha, \delta)$
else
    return $\tau^\star \leftarrow \min\{\tau \in \mathcal{T} :\ \text{CP}^{+}(K, n_{\text{keep}}, \delta') \le \alpha\}$
end if
```

Hyperparameters and Implementation Details

C.1 Global Configuration and Search Grids

Table 5: Global hyperparameters and search grids used across datasets. Paper defaults are in **bold**.

Component	Default	Grid / Options
Finite-difference step scale α_ε	**0.10**	$\{0.05,\ 0.10,\ 0.20\}$
EJOP subsample size m (fraction of train)	**0.2**	$\{0.1,\ 0.2,\ 0.4\}$
Ridge γ for preconditioner	**1e-3**	$\{1\text{e}{-4},\ 1\text{e}{-3},\ 1\text{e}{-2}\}$
Trace normalization	**on**	{on, off}
DIR-G strength ρ	**0.50**	$\{0,\ 0.25,\ 0.50,\ 0.75\}$
Forest (final) # trees	**500**	$\{300,\ 500,\ 800\}$
Forest (final) max depth	**None**	{None, 14, 20}
Forest (final) max features	$\sqrt{d}$	$\{\sqrt{d},\ d/3\}$
Forest (final) min samples leaf	**1**	$\{1,\ 3,\ 5\}$
Forest (surrogate) # trees	**200**	$\{100,\ 200\}$ (or multinomial logistic)
Calibration method	**Isotonic**	{Isotonic, Platt}
CP confidence level δ	**0.05**	$\{0.10,\ 0.05,\ 0.01\}$
Threshold grid size $\lvert\mathcal{T}\rvert$	**200**	$\{100,\ 200,\ 300\}$
Abstain cost c_{abs}	**task-specific**	Use task's cost model; sweeps reported in main text.

C.2 Training Configurations

Surrogate for EJOP/DIR-G. Random forest (learning_rate=0.05, max_depth=2, estimators=100). Class weights may mirror the application cost matrix.

Final forest. Unless otherwise stated, we use `n_estimators=`**500**, `max_features=`$\sqrt{d}$, `bootstrap=`**True**, `min_samples_leaf=`**1**, and leave `max_depth` unconstrained.

Calibration and scoring. Confidence score $s(x)$ is the isotonic-calibrated $\max_k \hat{p}_k(x)$ by default. For cost-aware variants, calibrate the scalar $1 - \mathbb{E}[\mathrm{cost}(\hat{y}(x), Y) \mid x]$ analogously.

C.3 CP Thresholding and Operating Points

We scan a grid $\mathcal{T}$ of size **200** over $[0, 1]$ (quantiles of $s(x)$ on the calibration split). We use Bonferroni-corrected one-sided Clopper–Pearson with per-grid level $\delta' = \delta/|\mathcal{T}|$ and $\delta = 0.05$. For the headline kept-error targets we report $\alpha \in \{0.05, 0.10, 0.15, 0.20\}$. If no τ satisfies $\mathrm{CP}^+(K, n, \delta') \le \alpha$, we report “abstain-all” or mark as infeasible.

C.4 Random Seeds, Versions, and Compute

Seeds and splits. We run **10** independent seeds $\{13, 37, 101, 202, 303, 404, 505, 606, 707, 808\}$. Each seed drives: (i) the train/cal/test split; (ii) subsample for EJOP; (iii) model initialization; (iv) any bootstrap sampling.

Table 6: Software versions and hardware.

Component	**Version / Spec**
OS	Ubuntu 22.04 LTS (x86_64)
Python	3.11.x
NumPy / SciPy	1.26.x / 1.11.x
scikit-learn	1.4.x
pandas	2.2.x
matplotlib (plots)	3.8.x
CPU	32 cores, 128 GB RAM
GPU (not required)	n/a

Datasets

1. **German Credit Dataset** – Source: UCI Machine Learning Repository. Focuses on classifying credit applicants as good or bad risks.
2. **Australian Credit Approval** – Source: UCI Machine Learning Repository. Binary classification for credit card applications.
3. **Default of Credit Card Clients (Taiwan Credit)** – Source: UCI Machine Learning Repository. Predicts default payments based on demographic and payment data.
4. **Polish Companies Bankruptcy** – Source: UCI Machine Learning Repository. Predicts bankruptcy (similar to credit default risk) using financial ratios.
5. **Give Me Some Credit** – Source: Kaggle. Predicts borrower default probability using historical data.
6. **Home Credit Default Risk** – Source: Kaggle. Large dataset for predicting loan defaults in home credit applications.
7. **Lending Club Loan Data** – Source: Kaggle/Lending Club. Peer-to-peer lending data for default prediction.
8. **PAKDD 2009 Credit Dataset** – Source: PAKDD Competition. Credit application data for risk assessment.
9. **UCI Heart Disease** – Source: UCI Machine Learning Repository. Binary classification for predicting heart disease presence based on medical attributes.
10. **Adult Income** – Source: UCI Machine Learning Repository. Binary classification for predicting if income exceeds $50K/year, used as a proxy for socioeconomic risk assessment.

Code availability: https://drive.google.com/file/d/1tbjTpOqbiprukxzFdCqT50IqwsSdkjt4/view?usp=sharing