THE PRICE OF PREDICTIVE MULTIPLICITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Classifiers with the same population performance may assign correct predictions to different people. We consider the problem of model selection in a scenario where individuals privately value correctness. Using Gibbs selection of a fixed library, we show that payments for truthful critical values can scale as $\min\{V\sqrt{n}, nV^2/T\}$ when every classifier is population optimal (n is even, V is an upper bound for values, T is temperature). Each payment depends not only on the report induced change in the probability that a model is chosen, but also on disagreement among almost optimal classifiers. Moreover, if all optimal classifiers agree on the correctness of individuals with positive values, aggregate payments vanish asymptotically along each predetermined schedule $T_n = o(n)$. We further show that each person's law of joint correctness and payment is characterized by normalized universal truthfulness, while the dependence across people is available for design. Based on these observations, we propose Blockwise Accuracy Auctions, which leave each individual's law invariant, while reducing the quantile of aggregate payment from linear to $\log n/\log\log n$ with confidence $1 - n^{-c}$ on a balanced instance for fixed $c > 1$. On Adult and Bank Marketing, independent deployment reduces 99% expected shortfall from 10.23 to 1.10 and from 6.62 to 1.02 without changing expected welfare or distributions of individual payments.

1 INTRODUCTION

Classifiers with the same population accuracy can make errors on different people. Studies of predictive multiplicity and Rashomon sets highlight the disagreement between well-performing models (Breiman, 2001; Fisher et al., 2019; Marx et al., 2020). Model selection therefore decides who receives a correct prediction (Black et al., 2022). If people value correctness differently, these values matter to the model selection. Even if population welfare is equal, it is not clear who will benefit.

A platform that selects classifiers based on reported correctness values must take into account strategic reports; a participant could have a higher chance of a correct prediction by reporting an inflated correctness value. We study welfare-based Gibbs selection from a fixed library (Huang & Kannan, 2012). Critical payments render truthful reporting optimal for all realizations of the platform's randomness. We study the burden of these payments: how it depends on the library and the data at hand, and how it is distributed among participants.

We show that payments can increase even when all classifiers are population optimal. Consider the two constant binary classifiers with fair labels and equal correctness values. These classifiers are both population optimal but correct for disjoint sets of participants. For even sample sizes, we give matching upper and lower bounds on the expected total payment at all temperatures. The rate of expected total payment is $\Theta(\sqrt{n})$ at a fixed temperature. More generally, any prescribed Gibbs temperature schedule that is welfare consistent across distributions incurs diverging expected total payment at even sample sizes. The population excess welfare is zero for every sample, including those that favor one classifier over the other.

We explain the payment burden by analyzing how one person's report changes the distribution over classifiers. Expected payment is a temperature-scaled KL divergence between the model distributions with and without that report. It also admits an integral representation in terms of correctness disagreement along the reporting path, which we use to bound payments through disagreements between near-optimal classifiers. If all population-optimal classifiers agree on the correctness of

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

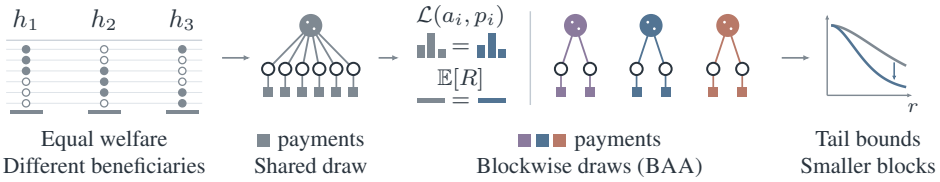


Figure 1: Shared and blockwise deployment in BAA. Filled matrix entries denote correct predictions. Each block uses all reports. Independent block draws preserve individual joint correctness–payment laws and expected total payment, while smaller blocks tighten Chernoff tail bounds.

positive-value individuals, the expected total payment instead tends to zero for every predetermined schedule $T_n = o(n)$. This condition allows multiple distinct but imperfect population-optimal classifiers. Therefore, populations can have opposite payment limits, even though every population optimum within each library achieves the same population welfare.

The platform also has another choice: which participants should share a random model draw (Figure 1). Once the selection rule is specified over all report profiles, normalized universal truthfulness fixes each person's joint law of correctness and payment. Dependence across people remains available for design, and shared draws can make large groups pay together. To address this, we propose Blockwise Accuracy Auctions (BAA), which use all reports in every block and change only the sharing of model draws. They preserve individual laws and incentives, and use block size to control the aggregate payment concentration. On a balanced profile, singleton blocks reduce polynomial-confidence payment quantiles from linear to $\log n / \log \log n$. Experiments with independently fitted logistic regression and decision tree libraries show that singleton deployment reduces 99% expected shortfall from 10.23 to 1.10 on Adult and from 6.62 to 1.02 on Bank Marketing, with expected welfare unchanged.

Our contributions are fourfold. First, we characterize individual allocation and payment laws, and bound the payments via information change and near-optimal correctness disagreement (Section 3). Second, we prove vanishing payments under beneficiary consensus, and matching all-temperature rates for conflicting beneficiaries, establishing a consistency-payment tradeoff within the Gibbs family (Section 4). Third, we characterize concentrated payments under shared decisions and equal billing, and construct a blockwise mechanism that preserves individual laws while guaranteeing aggregate tail bounds (Section 5). Fourth, we empirically validate the effectiveness of the proposed mechanism (Section 6).

2 Problem Formulation

Given a finite library $H = \{h_1, \ldots, h_K\}$ of deterministic classifiers $h : \mathcal{X} \to \mathcal{Y}$, with possibly multiclass label space, participant i has correctness allocation and utility

$$a_i(h) = \mathbf{1}\{h(x_i) = y_i\}, \qquad u_i = v_i a_i - p_i, \tag{1}$$

where $v_i \in [0, V]$ is private, $V > 0$ is public, and p_i is the payment. Each participant reports $b_i \in [0, V]$. Features and labels are fixed inputs. Utilities concern only the participant's own correctness and payment. We normalize $p_i(0, b_{-i}; \omega) = 0$ for every fixed random seed ω. A mechanism is *universally truthful* if truthful reporting maximizes (1) for every ω and every profile of other reports (Nisan & Ronen, 2001).

We condition on the library and any data used for its construction. No report will modify the set of its members. We consider statistical results for $S = \{(X_i, Y_i, v_i)\}_{i=1}^n$ iid and independent of library construction. We define

$$W(h) = \mathbb{E}[v\mathbf{1}\{h(X) = Y\}], \qquad W(q) = \mathbb{E}_{h\sim q} W(h), \qquad W^* = \max_{h \in H} W(h). \tag{2}$$

All population guarantees are relative to H. We are interested in total payment $R = \sum_i p_i$ and payer count $N = \sum_i \mathbf{1}\{p_i > 0\}$. Conditional expectations over mechanism randomness are denoted $\bar{p}_i$, $\rho = \mathbb{E}_\omega R$, and $\mu = \mathbb{E}_\omega N$. Statistical expectations are also averaged over S.

Model selection. Define the *Gibbs distribution* for total reported welfare $s_b(h) = \sum_i b_i a_i(h)$, $T > 0$ as

$$q_b(h) = \frac{e^{s_b(h)/T}}{\sum_{g \in H} e^{s_b(g)/T}}. \tag{3}$$

It is the unique distribution that maximizes the entropy-regularized welfare objective $\mathbb{E}_q s_b(h) - T\,\mathrm{KL}(q \| \mathrm{Unif}(H))$ (McSherry & Talwar, 2007; Huang & Kannan, 2012). Temperature applies to *total* welfare, so the temperature coefficient for the average welfare objective is T/n. We will use a *predetermined schedule* T_n, meaning that the temperature is fixed under any counterfactual report in the auction, including a zero report.

Blockwise Accuracy Auctions. Let $I_1, \ldots, I_G$ be a partition that does not depend on the reports. We draw independent standard Gumbels $G_{g,h}$ for each block g and set

$$h_g = \arg\max_{h \in H} \{s_b(h) + T G_{g,h}\}. \tag{4}$$

Then, each block uses all the reports and has marginal $h_g \sim q_b$, by the Gumbel-max trick (Maddison et al., 2014; Huijben et al., 2023). For $i \in I_g$, remove the participant's score contribution and take the maxima over correct and incorrect classifiers,

$$r_{i,h} = s_b(h) - b_i a_i(h) + T G_{g,h}, \quad M_i^+ = \max_{a_i(h)=1} r_{i,h}, \quad M_i^- = \max_{a_i(h)=0} r_{i,h}. \tag{5}$$

The payment is $p_i = a_i(h_g)[M_i^- - M_i^+]_+$. If either correctness group is empty, $p_i = 0$. In case of model ties, a fixed ordering is applied. We call this mechanism *Blockwise Accuracy Auctions* (BAA). One block corresponds to shared deployment; singleton blocks correspond to independent deployment.

Proposition 1 (Truthful implementation)**.** BAA *is normalized and universally truthful. At truthful reports,* $0 \le p_i \le v_i a_i$ *for every realization. Each agent receives a model with marginal distribution* q_b.

Conditioning on the random draws of the mechanism, with both correctness groups nonempty, a participant receives a correct prediction if $b_i + M_i^+ > M_i^-$. By reversing the inequality, we obtain an incorrect prediction; the fixed model ordering is used to break ties at equality. Critical-value pricing is truthful and individually rational, as shown in Appendix A. Using $C_{ik} = a_i(h_k)$, Algorithm 1 in Appendix A.2 implements the mechanism in $O(nK)$ time. Blocks can use different members of the library.

3 Payments and beneficiary disagreement

3.1 Individual payment distributions

Fix the other reports and write $\alpha(t) = \Pr(a = 1 \mid b_i = t)$. For a monotone allocation response, expected payments are determined by normalized single-parameter truthfulness (Myerson, 1981; Archer & Tardos, 2001). Universal truthfulness also determines their distribution.

Theorem 1 (Distributional payment equivalence)**.** *For any continuous, nondecreasing map* $\alpha : [0, V] \to [0, 1]$*, any normalized, universally truthful implementation of the response curve satisfies for all reports* b*,*

$$\Pr(p > x) = \alpha(b) - \alpha(x) \quad (0 \le x < b), \qquad \Pr(a = 1, p = 0) = \alpha(0). \tag{6}$$

Together with $\Pr(a = 0, p = 0) = 1 - \alpha(b)$*, these uniquely characterize the joint law of* (a, p).

Fix the random seed. Correctness then follows a threshold rule, with its critical payment. Aside from endpoint ties of probability zero, $p > x$ exactly when the threshold is in $(x, b]$. Averaging over seeds gives the theorem. This law determines every moment of each individual's payment; the implementation can still choose the dependence among participants.

For Gibbs selection, let $q_{i,t} = q_{(t,b_{-i})}$ and $\alpha_i(t) = \mathbb{E}_{q_{i,t}} a_i$. If both correctness groups are nonempty, then $\alpha_i(t) = \sigma(\gamma_i + t/T)$ for some report-independent offset γ_i where σ is logistic. Constant correctness rows have constant responses and zero payment. In both cases,

$$\alpha_i'(t) = \frac{\alpha_i(t)(1 - \alpha_i(t))}{T}, \qquad \bar{p}_i = \int_0^{b_i} \frac{t}{T} \mathrm{Var}_{q_{i,t}}(a_i)\, dt = T\,\mathrm{KL}(q_b \| q_{i,0}). \tag{7}$$

Differentiating the log partition function and integrating the threshold payment give this identity. Appendix B proves

$$\bar{p}_i + T\,\mathrm{KL}(q_{i,0}\|q_b) = b_i[\alpha_i(b_i) - \alpha_i(0)]. \tag{8}$$

Equation (7) thus expresses expected payment in terms of the change in model-selection probabilities induced by one report.

The current correctness probability also gives an explicit payment formula. With $a = \alpha_i(b_i)$ and $\beta = b_i/T$, we obtain

$$\alpha_i(0) = \frac{ae^{-\beta}}{1-a+ae^{-\beta}}, \qquad \bar{p}_i = b_i a + T\log(1-a+ae^{-\beta}) \tag{9}$$

Changing b_i multiplies every correct classifier's Gibbs weight by the same factor and preserves relative weights within each correctness group. The model-level KL in (7) therefore reduces to a Bernoulli KL. Given cached C, computing $s = C^\top b$ and $\alpha = Cq_b$ yields all expected payments in $O(nK)$ time.

Let $d_i(q_b) = \Pr_{h,h' \overset{\text{iid}}{\sim} q_b}(a_i(h) \neq a_i(h'))$. We have

$$e^{-b_i/T}\frac{b_i^2 d_i(q_b)}{4T} \le \bar{p}_i \le e^{b_i/T}\frac{b_i^2 d_i(q_b)}{4T}. \tag{10}$$

When $b_i/T \ll 1$, the payment is approximately $b_i^2 d_i(q_b)/(4T)$. Unanimous correctness and unanimous error both yield zero payment, since a report can only change the relative mass of the two correctness groups. Different incorrect labels do not contribute to d_i in multiclass problems.

3.2 Payments from near-optimal disagreement

For a fixed report profile and $\varepsilon \ge 0$, define

$$H_\varepsilon = \{h : s_b(h) \ge \max_g s_b(g) - \varepsilon\}, \qquad I_i^\varepsilon = \mathbf{1}\{\exists h, g \in H_\varepsilon \text{ with } a_i(h) \neq a_i(g)\}. \tag{11}$$

Let $M = K - |H_\varepsilon|$.

Theorem 2 (Payment localization)**.** *The conditional total payment satisfies*

$$\rho \le \frac{1}{8T}\sum_i b_i^2 I_i^\varepsilon + Me^{-(\varepsilon - V)/T}\sum_{i:I_i^\varepsilon=0}\min\left(b_i, \frac{b_i^2}{2T}\right). \tag{12}$$

Removing one report changes a pairwise score difference by at most V. The Gibbs mass outside H_ε is at most $Me^{-(\varepsilon-V)/T}$ along the reporting path. When near-optimal models agree on a participant's correctness, that mass bounds correctness variance, and hence payment. Each remaining participant contributes at most $b_i^2/(8T)$, and so (12) holds. The mechanism still samples from the full, report-independent library H.

4 Statistical payment bounds

Payment identities relate elicitation costs to generalization. For a library independent of the sample, let $e_m = V\sqrt{\log(2K)/(2m)}$. Finite-class concentration yields

$$\mathbb{E}[W^* - W(q_S)] \le 2e_n + \frac{T\log K}{n}, \tag{13}$$

$$\mathbb{E}R \le \min\left\{nV, \frac{nV^2}{8T}, ne_n + 2ne_{n-1} + \frac{nT\log K}{n-1}\right\}. \tag{14}$$

Here q_S is the Gibbs distribution at truthful values, $\widehat{W}_S(q) = n^{-1}\sum_i v_i \mathbb{E}_{h\sim q}a_i(h)$ is empirical welfare, and $n \ge 2$. The distribution q_{-i} removes only participant i's weight while leaving the library and temperature fixed. The distribution $q_{S_{n-1};T}$ uses an independent sample of size $n-1$. Summing (8) yields

$$\frac{1}{n}\mathbb{E}\left[R + T\sum_i \mathrm{KL}(q_{-i}\|q_S)\right] = \mathbb{E}\widehat{W}_S(q_S) - \mathbb{E}W(q_{S_{n-1};T}). \tag{15}$$

The removed observation is independent of q_{-i}, so its expected weighted correctness is population welfare. Finite-class concentration bounds the in-sample term. The identity in (15) relates welfare at different training sizes and includes both payment and a reverse-KL term. Appendix C proves these bounds, which imply $O(\sqrt{n})$ total payment and $O(n^{-1/2})$ population welfare gap for fixed T.

4.1 Consensus among optimal classifiers

Let $H^* = \arg\max_h W(h)$, $r = |H^*|$, and $M = K - r$. We say that H^* has *beneficiary consensus* if

$$\Pr\left(v > 0,\ a(h; X, Y) \neq a(g; X, Y)\right) = 0 \quad \text{for all } h, g \in H^*. \tag{16}$$

This allows different and imperfect classifiers to predict different incorrect labels or to be in disagreement only for participants with zero value.

Theorem 3 (Vanishing payments under consensus)**.** *Assume the consensus condition* (16) *and evaluate the mechanism at truthful reports. For $M > 0$, define $\Delta = \min_{h \notin H^*}[W^* - W(h)] > 0$. If $n \geq 4V/\Delta$, let*

$$\eta_n = \frac{M}{r} e^{-n\Delta/(4T)}, \qquad \delta_n = M e^{-n\Delta^2/(8V^2)}.$$

Then

$$\mathbb{E}R \leq n \min\left(V, \frac{V^2}{2T}\right) \eta_n + n \min\left(V, \frac{V^2}{8T}\right) \delta_n, \tag{17}$$

$$\mathbb{E}N \leq n \min\left(1, \frac{V}{T}\right) \eta_n + n \min\left(1, \frac{V}{4T}\right) \delta_n. \tag{18}$$

For any predetermined schedule $T_n/n \to 0$ the expectations converge to zero, as does $\Pr(N_n > 0)$. If $M = 0$ then the payment is zero almost surely.

On an event of probability at least $1 - \delta_n$, all nonoptimal classifiers have sample score at least $n\Delta/2$ lower than the optimal score. Under beneficiary consensus, all r optimal classifiers have the same value-weighted sample score. If any one report is reduced by at most V, the resulting advantage is at least $n\Delta/4$. Hence the total Gibbs probability of nonoptimal classifiers remains at most η_n along the reporting path, leading to (17)–(18). If $T_n = o(n)$, the factor $(n/T_n)e^{-cn/T_n}$ vanishes. For bounded or vanishing temperature, individual rationality leads to the exponentially small bound without the factor $1/T_n$.

Population-optimal classifiers need not be unique or error-free. Agreement on correctness for positive-value individuals is sufficient.

4.2 Payments under population ties

Theorem 4 (A matching all-temperature rate)**.** *Let $H = \{h_+, h_-\}$ be the two constant binary classifiers, $v_i = V$, and Y_i be iid fair signs. For every even $n \geq 2$ and every $T > 0$,*

$$\mathbb{E}R \asymp \min\left(V\sqrt{n}, \frac{nV^2}{T}\right), \qquad \mathbb{E}N \asymp \min\left(\sqrt{n}, \frac{nV}{T}\right), \tag{19}$$

with absolute constants. However, $W^* - W(q_S) = 0$ for every sample and temperature.

The classifiers have the same population welfare, but the benefits are distributed differently. Hence, reports can change the probability of a correct prediction, even though either selection is population optimal. Balanced samples explain the payment burden: for fixed $V, T > 0$, they occur with probability $\Theta(n^{-1/2})$ for even n and have conditional expected total payment $\Theta(n)$ by (23). Their contribution to $\mathbb{E}R$ alone is therefore $\Theta(\sqrt{n})$, before adding unbalanced samples.

The proof involves the representation

$$\mathbb{E}R = n \int_0^V \frac{t}{T} \mathbb{E}\left[\sigma'\left(\frac{t + VS_{n-1}}{T}\right)\right] dt, \tag{20}$$

where S_{n-1} is a sum of independent fair signs. Its largest atom is $O(n^{-1/2})$. For even n, the intervals $[Vs, Vs + V]$ for admissible s are disjoint; integrating the logistic density over them proves the

upper bound $V\sqrt{n}$. The upper bound $nV^2/(8T)$ is proved using $\sigma' \le 1/4$. The lower bounds from the central binomial mass match these on either side of $T \asymp V\sqrt{n}$. Appendix D proves the bounds uniformly in T and collects the lattice parity condition.

Corollary 1 (Consistency and total payment)**.** *Welfare consistency for every iid population holds within two-classifier Gibbs families with prescribed temperature schedules if and only if $T_n/n \to 0$. Every such schedule satisfies $\mathbb{E}R_n \to \infty$ on the fair-label population for even sample sizes.*

The condition is necessary because, on an all-positive-label population, the incorrect classifier is selected with probability $\sigma(-nV/T_n)$. On the tied population, Theorem 4 shows diverging payments. The divergence can be arbitrarily slow: take $T_n = n/g(n)$ with slowly diverging $g(n)$. The conclusion holds for Gibbs selection with a predetermined temperature schedule. Under the same consistency condition, Theorem 3 shows vanishing payments when the optimal classifiers agree on correctness, whereas the tied population in Theorem 4 has the opposite payment limit.

5 Deployment and aggregate payment risk

The previous statistical bounds average over data and mechanism randomness. For fixed data and reports, the aggregate payment risk depends on how model draws are shared among participants. For a given continuous allocation response to all counterfactual reports, normalized universal truthfulness fixes each participant's joint law of correctness and payment (Theorem 1). However, implementations may vary in their dependence among participants. Blockwise implementation changes this dependence while preserving the expected payment burden (Section 4).

5.1 Shared decisions and equal billing

We consider $n = 2m$ agents with identical features: m agents have positive labels, m agents have negative labels, and all agents bid V. The shared model selection rule takes the form

$$\Pr(h_+ \mid b) = F\left(\sum_{Y_i=+1} b_i - \sum_{Y_i=-1} b_i\right), \tag{21}$$

where $F : \mathbb{R} \to [0,1]$ is continuous and nondecreasing, and $F(-z) = 1 - F(z)$. In each group of agents with a given label, all agents are charged the same bill for each fixed random realization. The shared Gumbel mechanism satisfies this property.

Theorem 5 (Shared equal billing)**.** *Every normalized, universally truthful implementation of* (21) *based on one shared classifier and realization-wise equal billing obeys*

$$\Pr(R > mx) = 1 - 2F(x - V), \qquad \Pr(N = m) = 1 - 2F(-V), \tag{22}$$

for $0 \le x < V$. For Gibbs selection ($F(z) = \sigma(z/T)$) we have

$$\mathbb{E}_\omega R = nT \log\cosh\frac{V}{2T}, \qquad \Pr(R > nx/2) = \tanh\frac{V-x}{2T}. \tag{23}$$

At the balanced profile, the response of each participant is $\alpha_i(t) = F(t - V)$ and Theorem 1 gives the payment tail. The shared classifier is correct for exactly one label group, and equal billing gives its members a common payment. Adding the probabilities of the two group events yields (22).

For $T = n\tau$ with fixed $\tau > 0$, by (23) we have

$$\mathbb{E}_\omega R \to \frac{V^2}{8\tau}, \qquad \mathbb{E}_\omega N \to \frac{V}{4\tau}, \qquad \Pr(R > nV/4) = \Theta(1/n). \tag{24}$$

An aggregate bill of linear size occurs with probability $\Theta(1/n)$, even though the expected total payment is bounded. Individual rationality limits the payment to each participant's value, but the shared classifier and equal billing mean that the winning group pays together.

5.2 Concentration under blockwise deployment

Theorem 6 (Preservation and payment tails)**.** *For any report-independent partition, algorithm* BAA *preserves each participant's full joint law* (a_i, p_i) *from shared deployment. Let* s *be the largest block size and* $\psi(z) = z \log z - z + 1$.

$$\Pr(N \geq x) \leq e^{-(\mu/s)\psi(x/\mu)}, \qquad \Pr(R \geq y) \leq e^{-(\rho/(sV))\psi(y/\rho)} \tag{25}$$

for $x \geq \mu$ *and* $y \geq \rho$*. A zero mean implies an almost surely zero variable.*

Each block uses the marginal perturbation law of the shared mechanism, and the draws are independent across blocks. Hence, conditional on the data and reports, the block payment totals are independent, and each is bounded by sV. By convexity, their moment generating functions are bounded by those of scaled Poisson variables, which gives (25). Proposition 1 shows universal truthfulness, and the conditional means are unchanged.

Corollary 2 (Sharp quantile comparison)**.** *On the balanced profile with* $T = n\tau$*, singleton* BAA *gives for every* $c > 0$

$$Q_{1-n^{-c}}(N) \sim \frac{c \log n}{\log \log n}, \qquad Q_{1-n^{-c}}(R) \sim \frac{cV \log n}{\log \log n}, \tag{26}$$

while for shared implementations of the Gibbs response covered by Theorem 5 and $c > 1$ *the quantiles are* $n/2$ *and asymptotically* $nV/2$*.*

Under singleton deployment, N is binomial with mean converging to $V/(4\tau)$. Its logarithmic tail is $-k \log k + O(k)$, which implies (26). The same argument applied to the number of payments above any fixed $r < V$ gives the payment quantile. The bounds are conditioned on the balanced profile, whereas the statistical lower bound in Theorem 4 is an average over samples.

Test welfare under blockwise deployment. Let m be the size of an independent test batch with block sizes $m_1, \ldots, m_G$ independent of test observations. Let $U(h, Z) = v\mathbf{1}\{h(X) = Y\}$ and $\sigma_h^2 = \operatorname{Var}_Z U(h, Z)$. Given the learned model distribution q,

$$\operatorname{Var}(\widehat{W}_{\mathrm{dep}} \mid q) = \frac{\mathbb{E}_q \sigma_h^2}{m} + \frac{\sum_g m_g^2}{m^2} \operatorname{Var}_q W(h). \tag{27}$$

The first term is sampling variation conditional on the classifier. The second is variation in population welfare across model draws; it vanishes if all classifiers have equal population welfare, even if their individual predictions differ. For the two constant classifiers evaluated on independent fair test labels, the correctness count is $\mathrm{Bin}(m, 1/2)$ for every block size. Hence all deployments in Corollary 2 have the same test-accuracy distribution. Appendix E.7 derives (27) by conditioning on the model drawn for each block.

5.3 Alternative payment rules

The conditional mean fee can be used in place of the realized critical price if truthfulness in expectation is enough. Keep the Gibbs allocation but charge the participant with a correct prediction

$$\widetilde{p}_i = \frac{\bar{p}_i}{\alpha_i(b_i)} \leq \min\left(b_i, \frac{b_i^2}{2T}\right), \tag{28}$$

and zero fee otherwise. The expected payment at each report is $\bar{p}_i$, so the mechanism is truthful in expectation for risk-neutral participants. If $T = n\tau$, every aggregate bill is at most $V^2/(2\tau)$.

For some seeds, the classifier remains correct if a participant reduces a positive report to zero. This deviation does not change the correctness, but removes the conditional fee, so the mechanism is not universally truthful. Blockwise implementation enforces the critical price for all seeds and can assign different realized classifiers to different participant groups. Appendix F.1 gives the payment bound and deviation argument. The comparison is based on the incentive guarantees of the implementations.

6 Experiments

We compare the extent of payment concentration on independently fitted classifier libraries on two datasets. We fix the models, reporting profiles, and temperature across deployments.

Table 1: Prediction quality and payments on Adult with label-associated values. Table entries represent means and standard deviations over five splits.

Method	Accuracy (%)	Test welfare	Residual benefit	ρ	μ
Accuracy-only	84.9±0.2	.3142±.0025	.3174±.0050	0	0
Welfare ERM	80.4±1.4	.3433±.0032	.3458±.0038	1.36±1.71	4.8±6.7
Free Gibbs	83.0±0.2	.3226±.0025	.3241±.0044	0	0
BAA	81.9±0.1	.3289±.0024	.3307±.0042	.102±.005	.270±.010

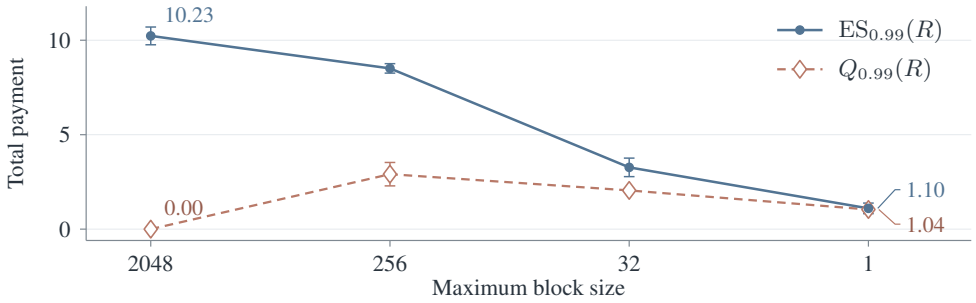


Figure 2: Payment risk across block sizes on Adult (mean ± standard deviation over five splits).

Data and models. We use Adult and Bank Marketing data from the UCI repository (Becker & Kohavi, 1996; Moro et al., 2014). Each of the five random splits uses 40% of the data for construction, 20% for an independent reference set, 20% for the auction pool, and 20% for testing. The default auction size is $n = 2048$. On the construction set, we fit four logistic regressions with inverse regularization parameters $\{10^{-2}, 10^{-1}, 1, 10\}$ and four trees with depths $\{2, 3, 4, 6\}$. Thresholding each estimator at 0.25 and 0.5 yields $K = 16$ classifiers. Appendix G describes the preprocessing and other details.

Values and comparison methods. The baseline values are $v = 0.25 + 0.75\mathbf{1}\{Y = 1\}$, $V = 1$, $T = 0.1n$ (uniform and age-based assignments are also considered). All use utility weights on real features and labels. The same fitted library is used for comparisons. Accuracy-only selection ignores reports and charges nothing. Welfare ERM maximizes reported welfare and charges deterministic critical payments. Free Gibbs charges nothing and capped reports are weakly dominant for positive-value participants. Shared, blockwise, and singleton BAA have the same model marginals and individual payment laws. Smoothed bills have the same expected payments but only guarantee truthfulness in expectation.

Evaluation. Test accuracy, welfare, expected payment ρ, and expected payer count μ are computed using cached predictions. The residual cohort benefit is $n^{-1}\sum_i(v_i\mathbb{E}_\omega a_i - \bar{p}_i)$. For each split and condition, 4096 deployment draws are used to estimate the 0.99 payment quantile and expected shortfall

$$\mathrm{ES}_{0.99}(R) = \inf_t\{t + 100\mathbb{E}[(R - t)_+]\}. \tag{29}$$

The mean and standard deviation are taken across five splits. The expected shortfall is estimated using the identity $\mathbb{E}[(R - t)_+] = \rho - \mathbb{E}[\min(R, t)]$, combining the analytic mean total payment with an estimate of a bounded, truncated term. Only the latter term is estimated from deployment draws.

Welfare and mean payment. BAA increases test welfare by 0.0147 and decreases accuracy by 3.0 percentage points compared to accuracy-only selection (Table 1). BAA decreases welfare by 0.0144, increases accuracy by 1.5 percentage points, and decreases mean total payment from 1.36 to 0.102 compared to welfare ERM. Bank Marketing has welfare differences of $+0.0080$ and -0.0085 in the same order. The tradeoffs depend on the value-weighted objective and temperature. Block size does not change welfare or mean payment.

Aggregate payment risk. On Adult, the quantile at the 0.99 level is zero under shared deployment, but expected shortfall is 10.23 (Figure 2). Under singleton deployment, the quantile is 1.04 and expected shortfall is 1.10; expected shortfall is 8.51 and 3.27 for block sizes 256 and 32. The reduction in expected shortfall from shared to singleton deployment is 9.13 on Adult and 5.60 on Bank Marketing. The shared-minus-singleton quantile differences are -1.04 and -0.94. Thus, at the 0.99 level, the quantile ranks shared deployment more favorably despite its higher expected shortfall. Corollary 2 concerns the balanced profile at confidence levels $1 - n^{-c}$ with $c > 1$. It does not order every fixed quantile on arbitrary data.

On Adult, independent deployment reduces total payment variance from 5.960 to 0.064. The second moments of individual payments are unchanged, so the 5.896 reduction is exactly the covariance contribution under shared deployment, $2\sum_{i<j}\mathrm{Cov}(p_i, p_j)$. On Bank Marketing, the reduction is 2.063. Independence eliminates cross-participant covariances, but preserves each participant's payment law. Payment and payer count statistics are in Appendix G.9.

Sensitivity analyses. Increasing T from 1 to $0.1n$ decreases Adult's mean payment by 4.56, increases the empirical test-welfare gap by 0.0142, and increases effective library size by 12.8. Uniform, label-associated, and age-associated values have mean payments 0.242, 0.102, and 0.164. Their near-optimal disagreement fractions are 8.1%, 16.9%, and 11.3%. Disagreement frequency alone does not suffice to order the mean payments, because (7) also depends on the values and Gibbs mass. Compared to the full library, the logistic-only and tree-only libraries change test welfare by $+0.0019$ and -0.0020, and mean payment by -0.017 and $+0.011$. These comparisons use the fitted classifiers.

7 Related work

Rashomon set analyses identify alternatives among well-performing models (Fisher et al., 2019; Semenova et al., 2022) and predictive multiplicity measures the disagreement between model predictions (Marx et al., 2020; Hsu & Calmon, 2022; Watson-Daniels et al., 2023). Algorithmic monoculture refers to the dependence introduced by shared models and components (Kleinberg & Raghavan, 2021; Bommasani et al., 2022). We relate correctness disagreement to the payments induced by reports and investigate the impact of sharing model draws on the aggregate payment risk. Information-theoretic generalization bounds rely on mutual information and KL divergence (Xu & Raginsky, 2017; Steinke & Zakynthinou, 2020; Aminian et al., 2021) and our payment identity provides a monetary interpretation of the KL divergence caused by removing a report.

Incentive and allocation mechanisms are based on critical pricing and exponential mechanism welfare selection (Myerson, 1981; McSherry & Talwar, 2007; Huang & Kannan, 2012). Private correctness values also arise in accuracy auctions, where learning stability bounds payments (Sadi et al., 2026). Our finite library results are related to optimal model disagreement and deployment dependence. Manipulation is studied in strategic classification (Hardt et al., 2016; Ghalme et al., 2021), distribution shift due to deployment in performative prediction (Perdomo et al., 2020), and learning provider incentives in delegated classification (Saig et al., 2023). Residual surplus mechanisms optimize benefits minus payments over general outcome spaces (Hartline & Roughgarden, 2008; Fotakis et al., 2016; Ezra, 2026). Our sample size results apply to the predetermined temperature Gibbs family.

8 Discussion and Conclusion

Elicitation costs can arise even when all classifiers in a fixed library are population optimal. Under the beneficiary consensus, the expected total payment vanishes at every predetermined Gibbs temperature schedule $T_n = o(n)$. We prove matching payment bounds at every temperature for the tied two-classifier population. This illustrates the source of elicitation costs in the way near-optimal classifiers distribute correctness among participants.

Blockwise Accuracy Auctions preserve the joint law of correctness and payment of every individual, but allow dependence between individuals to be changed. Smaller blocks lead to concentration bounds on aggregate payments. With learned models, they reduce expected shortfall at the same expected welfare.

REFERENCES

Carlo Acerbi and Dirk Tasche. On the coherence of expected shortfall. *Journal of Banking & Finance*, 26(7):1487–1503, 2002.

Gholamali Aminian, Yuheng Bu, Laura Toni, Miguel Rodrigues, and Gregory Wornell. An exact characterization of the generalization error for the Gibbs algorithm. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8106–8118, 2021.

Aaron Archer and Éva Tardos. Truthful mechanisms for one-parameter agents. In *Proceedings of the IEEE Symposium on Foundations of Computer Science*, pp. 482–491, 2001.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

Moshe Babaioff, Robert D. Kleinberg, and Aleksandrs Slivkins. Truthful mechanisms with implicit payment computation. *Journal of the ACM*, 62(2):10:1–10:37, 2015.

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6(58):1705–1749, 2005.

Barry Becker and Ronny Kohavi. *Adult*. UCI Machine Learning Repository, 1996.

George Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

Emily Black, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pp. 850–863, 2022.

Rishi Bommasani, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky, and Percy Liang. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, second edition, 2006.

André Elisseeff, Theodoros Evgeniou, and Massimiliano Pontil. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(3):55–79, 2005.

Tomer Ezra. Optimal prior-free mechanisms for consumer surplus. *arXiv preprint arXiv:2608.01693*, 2026.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

Dimitris Fotakis, Dimitris Tsipras, Christos Tzamos, and Emmanouil Zampetakis. Efficient money burning in general domains. *Theory of Computing Systems*, 59(4):619–640, 2016.

Pascal Germain, Alexandre Lacasse, François Laviolette, Mario Marchand, and Jean-Francis Roy. Risk bounds for the majority vote: From a PAC-Bayesian analysis to a learning algorithm. *Journal of Machine Learning Research*, 16(26):787–860, 2015.

Ganesh Ghalme, Vineet Nair, Itay Eilat, Inbal Talgam-Cohen, and Nir Rosenfeld. Strategic classification in the dark. In *Proceedings of the International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3672–3681, 2021.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the Conference on Innovations in Theoretical Computer Science*, pp. 111–122, 2016.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, et al. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020.

Jason D. Hartline and Tim Roughgarden. Optimal mechanism design and money burning. In *Proceedings of the ACM Symposium on Theory of Computing*, pp. 75–84, 2008.

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

Hsiang Hsu and Flavio Calmon. Rashomon capacity: A metric for predictive multiplicity in classification. In *Advances in Neural Information Processing Systems*, volume 35, pp. 28988–29000, 2022.

Zhiyi Huang and Sampath Kannan. The exponential mechanism for social welfare: Private, truthful, and nearly optimal. In *Proceedings of the IEEE Symposium on Foundations of Computer Science*, pp. 140–149, 2012.

Iris A. M. Huijben, Wouter Kool, Max B. Paulus, and Ruud J. G. van Sloun. A review of the Gumbel-max trick and its extensions for discrete stochasticity in machine learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1353–1371, 2023.

Jon Kleinberg and Manish Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, 2021.

Lucien Le Cam. An approximation theorem for the Poisson binomial distribution. *Pacific Journal of Mathematics*, 10(4):1181–1197, 1960.

Chris J. Maddison, Daniel Tarlow, and Tom Minka. A* sampling. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *Proceedings of the International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6765–6774, 2020.

David A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the Conference on Computational Learning Theory*, pp. 164–170, 1999.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *Proceedings of the IEEE Symposium on Foundations of Computer Science*, pp. 94–103, 2007.

Sérgio Moro, Paulo Rita, and Paulo Cortez. *Bank Marketing*. UCI Machine Learning Repository, 2014.

Roger B. Myerson. Optimal auction design. *Mathematics of Operations Research*, 6(1):58–73, 1981.

Noam Nisan and Amir Ronen. Algorithmic mechanism design. *Games and Economic Behavior*, 35(1–2):166–196, 2001.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7599–7609, 2020.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41, 2000.

Bana Sadi, Eden Saig, and Nir Rosenfeld. Welfare-optimal classification with accuracy auctions. In *Proceedings of the International Conference on Machine Learning*, 2026.

Eden Saig, Inbal Talgam-Cohen, and Nir Rosenfeld. Delegated classification. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Lesia Semenova, Cynthia Rudin, and Ronald Parr. On the existence of simpler machine learning models. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pp. 1827–1858, 2022.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(90):2635–2670, 2010.

Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. In *Proceedings of the Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3437–3452, 2020.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

Jamelle Watson-Daniels, David C. Parkes, and Berk Ustun. Predictive multiplicity in probabilistic classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10306–10314, 2023.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

A TRUTHFUL IMPLEMENTATION AND INDIVIDUAL PAYMENT LAWS

A.1 NOTATION AND CONDITIONING

Table 2 summarizes the notation used in the proofs. In the comparisons of reports, we fix four objects: the classifier library, the temperature, the prior, and the distribution of random seeds. Setting b_i to zero removes only that person's score contribution; the temperature remains the same as the one assigned to the original cohort size.

Table 2: Notation for the allocation rule, individual payments, and population guarantees. Conditional moments keep the observations and reports fixed. Population expectations furthermore average over an iid sample independent of the classifier library.

Symbol	Definition
H, K	Fixed classifier library and its cardinality
$a_i(h)$	Indicator that classifier h is correct for participant i
v_i, b_i, V	Private value, report, and public upper bound
T, T_n	Total-score temperature and its predetermined sample-size schedule
$s_b(h)$	Reported welfare $\sum_i b_i a_i(h)$
q_b, $q_{i,t}$	Gibbs distributions at reports b and (t, b_{-i})
$\alpha_i(t)$	Probability of correctness under $q_{i,t}$
p_i, $\bar{p}_i$	Realized and conditional expected payment
R, N	Total payment and number of positive payments
ρ, μ	Conditional means $\mathbb{E}_\omega R$ and $\mathbb{E}_\omega N$
H_ε	Classifiers within ε of the largest reported score
H^*, r	Population-optimal classifiers and their number
Δ	Smallest population welfare gap outside H^*
s	Maximum number of participants sharing one model draw
$Q_u(R)$	Left quantile $\inf\{x : \Pr(R \le x) \ge u\}$

For the correctness matrix for mechanism statements, we may take any fixed binary matrix. There are no distributional assumptions on the rows. For statistical statements, we assume triples (X_i, Y_i, v_i) are iid conditional on an independently constructed library. We show conditional payment concentration after observing this sample, using the same symbols ρ and μ throughout the dependence proofs.

Universal truthfulness requires truthful reporting to maximize a participant's utility for every fixed seed and every profile of other reports. Truthfulness in expectation requires utility maximization after averaging over the random seed. Normalization states that a participant's report of zero leads to a zero payment in every realization. Normalized universal truthfulness for binary allocations results in the critical price representation.

A.2 PSEUDOCODE

The algorithm for matrix form construction is presented in Algorithm 1. Here, the score vector includes all reports, and the partition indicates which participants share a Gumbel vector. The partition, temperature, Gumbel draws and model ordering are held fixed for counterfactual comparisons (only the score contributions differ). Below, we use the normalized critical-payment characterization for this implementation.

A.3 CRITICAL PAYMENTS UNDER NORMALIZATION

Fix the other reports and a realization of mechanism randomness. The normalized single-parameter characterization (Myerson, 1981) applies to the allocation $a(t) \in \{0, 1\}$ and payment $p(t)$ at report t. For types $t > u$, the two incentive constraints yield

$$ta(t) - p(t) \ge ta(u) - p(u), \qquad ua(u) - p(u) \ge ua(t) - p(t).$$

Adding the two inequalities gives $(t - u)[a(t) - a(u)] \ge 0$, so a is nondecreasing. If two reports have the same allocation, then the same two incentive constraints imply equal payments. There is thus one payment on the losing interval and one on the winning interval.

Algorithm 1 Blockwise Accuracy Auction

Require: Correctness matrix $C \in \{0,1\}^{n\times K}$, bids b, temperature T, partition $\{I_g\}$

$s \leftarrow C^\top b$
for each block I_g **do**
  Draw independent standard Gumbels $G_{g,1}, \dots, G_{g,K}$
  $z_k \leftarrow s_k + TG_{g,k}$ for $k = 1, \dots, K$, and $k_g \leftarrow \arg\max_k z_k$
  for each $i \in I_g$ **do**
    $a_i \leftarrow C_{i,k_g}, \quad p_i \leftarrow 0$
    if row C_i contains both 0 and 1 **then**
      $r_{i,k} \leftarrow z_k - b_i C_{i,k}$ for $k = 1, \dots, K$
      $p_i \leftarrow a_i \big[\max_{k:C_{i,k}=0} r_{i,k} - \max_{k:C_{i,k}=1} r_{i,k}\big]_+$
    end if
  end for
end for
return Allocations a, payments p, and block classifiers $\{h_{k_g}\}$

If the losing interval is non-empty, then normalization requires that the payment for the losing interval be zero. Let c denote the success threshold. Incentive compatibility requires that the payment for success be no more than t for every successful type t, and no less than u for every unsuccessful type u. Taking the limits as $t \downarrow c$ and $u \uparrow c$ gives the critical payment c. An allocation that is constant over the reporting domain has payment zero everywhere. For the continuous-response curves we consider, the choice of threshold endpoints is immaterial.

A.4 Proof of Proposition 1

For a fixed Gumbel vector in participant i's block, the maximum score of a correct classifier is $b_i + M_i^+$, and the maximum score of an incorrect classifier is M_i^-. Thus correctness is defined by the inequality $b_i + M_i^+ > M_i^-$ and failure by $b_i + M_i^+ < M_i^-$. In the case of equality, the specified model ordering determines the allocation, with critical payment unchanged. The above construction of critical payment implies universal truthfulness. Truthful participants pay at most their truthful value, and unsuccessful participants pay zero.

For the model marginal, transform each standard Gumbel G_h to $E_h = e^{-G_h}$, where $E_h \sim \mathrm{Exp}(1)$. Maximizing $s_b(h)/T + G_h$ is the same as minimizing $e^{-s_b(h)/T}E_h$. The random variables are independent exponentials with rates $e^{s_b(h)/T}$. The minimum is at h with probability $q_b(h)$. This establishes the marginal in each block. The utility of each player depends only on the outcome of their own block. Conditioning on the random draws in all blocks preserves the individual threshold argument. □

A.5 Proof of Theorem 1

For each fixed realization, the mechanism is a normalized threshold rule, with constant allocations having zero payment. Let c denote the threshold for a nonconstant realization. Continuity of the response implies that an active threshold tie at any fixed report is a null event. Outside of these ties, the event $p > x$ for $0 \le x < b$ is precisely $\{x < c \le b\}$. Averaging over realizations yields

$$\Pr(p > x) = \alpha(b) - \alpha(x).$$

The remaining probability masses are $\Pr(a = 0, p = 0) = 1 - \alpha(b)$ and $\Pr(a = 1, p = 0) = \alpha(0)$. The response curve determines the entire joint distribution. In particular, for any $r > 0$,

$$\mathbb{E}p^r = r \int_0^b x^{r-1}[\alpha(b) - \alpha(x)]\, dx. \tag{A.1}$$

The moment formula follows by integrating the tail of p^r. Continuity of α rules out positive threshold atoms in the interior of the reporting interval. Normalization is crucial: adding a report-independent transfer would leave incentives unchanged while modifying the payment law. □

A.6 Expected utility and deviations

Let the true value of a participant be v, and let the function $\alpha(t)$ be the response given a report, with the other reports fixed. Then the distributional payment equivalence is

$$\bar{p}(b) = b\alpha(b) - \int_0^b \alpha(t)\, dt. \tag{A.2}$$

Then the expected utility at truthful reporting is

$$U(v; v) = v\alpha(v) - \bar{p}(v) = \int_0^v \alpha(t)\, dt. \tag{A.3}$$

We have included the utility obtained at zero report. A participant with a constant response function $\alpha(t) = 1$ obtains utility v and pays nothing.

The expression in (A.2) on substitution of the alternative report $b \leq v$ implies

$$U(v; v) - U(v; b) = \int_b^v [\alpha(t) - \alpha(b)]\, dt \geq 0. \tag{A.4}$$

The corresponding expression when $b \geq v$ is

$$U(v; v) - U(v; b) = \int_v^b [\alpha(b) - \alpha(t)]\, dt \geq 0. \tag{A.5}$$

Monotonicity establishes the above inequalities. These expressions specify expected gains and losses under the same allocation response. The seedwise threshold argument proves the stronger universal guarantee for BAA.

Summing (A.3) at truthful reports gives the following cohort-level identity,

$$\mathbb{E}_\omega \sum_i (v_i a_i - p_i) = \sum_i \int_0^{v_i} \alpha_i(t)\, dt. \tag{A.6}$$

Each term is nonnegative. This quantity refers to the set of participants who make the reports. Test welfare, calculated from new data under the learned model distribution, does not have to equal the cohort quantity.

B Payment identities and disagreement bounds

B.1 Partition-function calculation

Fix participant i. Group the partition function by correctness,

$$Z_t = Ae^{t/T} + B, \quad A = \sum_{h:a_i(h)=1} e^{\sum_{j\neq i} b_j a_j(h)/T}, \quad B = \sum_{h:a_i(h)=0} e^{\sum_{j\neq i} b_j a_j(h)/T}. \tag{B.1}$$

When both groups exist, $A, B > 0$ and

$$\alpha_i(t) = \frac{Ae^{t/T}}{Ae^{t/T} + B}, \qquad \alpha_i'(t) = \frac{\alpha_i(t)[1 - \alpha_i(t)]}{T}.$$

The payment law in Theorem 1 yields

$$\begin{aligned} \bar{p}_i &= b_i\alpha_i(b_i) - \int_0^{b_i} \alpha_i(t)\, dt = \int_0^{b_i} t\, d\alpha_i(t) \\ &= b_i\alpha_i(b_i) - T \log \frac{Z_{b_i}}{Z_0}, \end{aligned} \tag{B.2}$$

as $d(T \log Z_t)/dt = \alpha_i(t)$. Moreover,

$$\log \frac{q_b(h)}{q_{i,0}(h)} = \frac{b_i a_i(h)}{T} - \log \frac{Z_{b_i}}{Z_0}.$$

The expectation under q_b gives $\bar{p}_i = T\,\mathrm{KL}(q_b \| q_{i,0})$. Applying the reverse expectation and adding gives

$$\bar{p}_i + T\,\mathrm{KL}(q_{i,0} \| q_b) = b_i[\alpha_i(b_i) - \alpha_i(0)].$$

In both groups, the conditional model distribution is not dependent on b_i. Using the chain rule for the KL divergence (Cover & Thomas, 2006), we obtain

$$\mathrm{KL}(q_b \| q_{i,0}) = \mathrm{KL}\left(\mathrm{Ber}(\alpha_i(b_i)) \| \mathrm{Ber}(\alpha_i(0))\right). \tag{B.3}$$

The odds ratio between report b_i and report zero increases by the factor $e^{b_i/T}$. Using $a = \alpha_i(b_i)$ and $\beta = b_i/T$, we have

$$\alpha_i(0) = \frac{ae^{-\beta}}{1 - a + ae^{-\beta}}, \qquad \bar{p}_i = b_i a + T\log(1 - a + ae^{-\beta}).$$

These expressions can be extended continuously to the constant correctness rows, in which the payment is zero.

B.2 Payment density and moments

For a Gibbs response with both correctness groups nonempty, (6) implies the density of positive payments,

$$f_i(x) = \frac{1}{T}\alpha_i(x)(1 - \alpha_i(x)), \qquad 0 < x < b_i. \tag{B.4}$$

Its integral is $\alpha_i(b_i) - \alpha_i(0)$. The remaining mass is at zero,

$$\Pr(p_i = 0) = 1 - \alpha_i(b_i) + \alpha_i(0).$$

Among zero payments, mass $\alpha_i(0)$ corresponds to correct predictions and mass $1 - \alpha_i(b_i)$ to incorrect predictions. For every $k > 0$,

$$\begin{aligned}\mathbb{E}_\omega p_i^k &= \int_0^{b_i} \frac{x^k}{T}\alpha_i(x)(1 - \alpha_i(x))\,dx \\ &= k\int_0^{b_i} x^{k-1}[\alpha_i(b_i) - \alpha_i(x)]\,dx.\end{aligned} \tag{B.5}$$

The second equality can be proved by integration by parts. In particular,

$$\mathrm{Var}_\omega(p_i) \le \mathbb{E}_\omega p_i^2 \le b_i\bar{p}_i, \qquad \mathbb{E}_\omega p_i^2 \le \frac{b_i^3}{12T}. \tag{B.6}$$

The first estimate follows from $0 \le p_i \le b_i$. The second follows from (B.4) and $\alpha_i(1 - \alpha_i) \le 1/4$. Both estimates hold for every implementation from Theorem 1.

B.3 Curvature of the welfare objective

The log-partition derivative identities (Wainwright & Jordan, 2008) for exponential families imply that $\Phi(b) = T\log\sum_h \exp(s_b(h)/T)$ satisfies

$$\frac{\partial\Phi}{\partial b_i} = \alpha_i(b_i), \qquad \frac{\partial^2\Phi}{\partial b_i \partial b_j} = \frac{1}{T}\,\mathrm{Cov}_{h\sim q_b}(a_i(h), a_j(h)). \tag{B.7}$$

For any $z \in \mathbb{R}^n$,

$$z^\top \nabla^2\Phi(b) z = \frac{1}{T}\,\mathrm{Var}_{q_b}\left(\sum_i z_i a_i(h)\right) \ge 0.$$

The diagonal entries measure sensitivity to an individual's own report. An off-diagonal entry can be positive or negative. Its sign indicates whether correctness for one participant is positively or negatively correlated with correctness for another, given random selection of classifier. This covariance is over candidate classifiers under q_b. Aggregate payment covariances also depend on the mechanism's coupling of the model draws of different participants.

Noting that $b^{(-i)} = (0, b_{-i})$, the Bregman divergence induced by Φ (Banerjee et al., 2005):

$$D_\Phi(x, y) = \Phi(x) - \Phi(y) - \langle \nabla\Phi(y), x - y\rangle,$$

implies that (A.2) becomes

$$\bar{p}_i = D_\Phi(b^{(-i)}, b). \tag{B.8}$$

The argument order is consistent with the forward model-distribution KL identity (7). Equation (B.8) is derived by integrating the first derivative of Φ along the one-coordinate reporting path.

B.4 Sharp individual bounds

The elementary bound $\alpha(1-\alpha) \le 1/4$, together with individual rationality, implies

$$\bar{p}_i \le \min\left(b_i, \frac{b_i^2}{8T}\right), \qquad \Pr(p_i > 0) \le \min\left(1, \frac{b_i}{4T}\right). \tag{B.9}$$

It is also possible to determine the maximum over correctness-group offsets. For $\beta = b_i/T > 0$ define $x_\beta = \beta/(1 - e^{-\beta})$. Then

$$\Pr(p_i > 0) \le \tanh(\beta/4), \qquad \bar{p}_i \le T[x_\beta - 1 - \log x_\beta]. \tag{B.10}$$

For the probability bound, maximize $\sigma(\gamma+\beta) - \sigma(\gamma)$ over γ. The maximum is attained at $\gamma = -\beta/2$. For payment, set $k = 1 - e^{-\beta}$. Dividing (9) by T yields the strictly concave function

$$f(a) = \beta a + \log(1 - ka), \qquad a \in [0, 1].$$

The maximizer $a^* = 1/k - 1/\beta$ lies in $(0, 1)$. Substitution yields (B.10). Both bounds are sharp over logistic offsets $\gamma \in \mathbb{R}$, but with different offsets that maximize them. Attainment in a given classifier library is contingent upon what group scores are offered.

B.5 Bounds from current disagreement

Correctness disagreement is related to voter-disagreement quantities as studied by Germain et al. (2015) in the context of majority vote analysis. The random output is one classifier drawn from the set and the integrals below are used to quantify the price of changing its correctness probability.

Let $g(t) = \alpha_i(t)[1 - \alpha_i(t)]$. For nonconstant correctness rows, we have

$$\left|\frac{d}{dt}\log g(t)\right| = \frac{|1 - 2\alpha_i(t)|}{T} \le \frac{1}{T}.$$

Therefore for $0 \le t \le b_i$,

$$e^{-b_i/T} g(b_i) \le g(t) \le e^{b_i/T} g(b_i).$$

Integration of $tg(t)/T$ over the reporting interval, together with the fact that $d_i(q_b) = 2g(b_i)$, yields (10). For constant correctness rows, both sides of (10) are zero.

B.6 Additive bounds for small reports

There is an additive version of the multiplicative estimate (10). Let $g_i(t) = \alpha_i(t)(1 - \alpha_i(t))$. Then

$$g_i'(t) = \frac{1}{T} g_i(t)(1 - 2\alpha_i(t)), \qquad |g_i'(t)| \le \frac{1}{4T}.$$

We get

$$\left|\bar{p}_i - \frac{b_i^2}{2T} g_i(b_i)\right| \le \int_0^{b_i} \frac{t}{T}\frac{b_i - t}{4T}\, dt = \frac{b_i^3}{24T^2}, \tag{B.11}$$

$$\left|\Pr(p_i > 0) - \frac{b_i}{T} g_i(b_i)\right| \le \int_0^{b_i} \frac{b_i - t}{4T^2}\, dt = \frac{b_i^2}{8T^2}. \tag{B.12}$$

Now, $d_i(q_b) = 2g_i(b_i)$, so the leading terms are $b_i^2 d_i(q_b)/(4T)$ for payment and $b_i d_i(q_b)/(2T)$ for the payer probability. These bounds apply to fixed reports and require no distributional assumption on the observations. The results bound the approximation error by an absolute amount, even if the disagreement is small.

B.7 Proof of Theorem 2

Suppose the current score maximizer is h_0. Then, for all $h \notin H_\varepsilon$, we have $s_b(h_0) - s_b(h) > \varepsilon$. We decrease b_i to some $t \in [0, b_i]$. Then the change in the pairwise gap is no more than $b_i - t \le V$. The partition function includes a term for h_0. Hence, for all t along the path, we have

$$q_{i,t}(H \setminus H_\varepsilon) \le M e^{-(\varepsilon - V)/T} =: \eta. \tag{B.13}$$

The inequality does not assume that h_0 is the score maximizer for all t.

If $I_i^\varepsilon = 0$, the models in H_ε have the same correctness $c_i \in \{0, 1\}$. If $c_i = 1$, then $1 - \alpha_i(t) \le \eta$. If $c_i = 0$, then $\alpha_i(t) \le \eta$. In both cases we have

$$\mathrm{Var}_{q_{i,t}}(a_i) \le \eta, \qquad \bar{p}_i \le \frac{b_i^2}{2T}\eta, \qquad \Pr(p_i > 0) \le \frac{b_i}{T}\eta.$$

The endpoints also satisfy $\alpha_i(b_i) - \alpha_i(0) \le \eta$. Since $p_i \le b_i \mathbf{1}\{p_i > 0\}$, this implies $\bar{p}_i \le b_i \eta$. For $I_i^\varepsilon = 1$ we can use (B.9). Summing these bounds proves Theorem 2. The same argument with (B.10) gives

$$\mu \le \sum_{i: I_i^\varepsilon = 1} \tanh \frac{b_i}{4T} + M e^{-(\varepsilon - V)/T} \sum_{i: I_i^\varepsilon = 0} \min\left(1, \frac{b_i}{T}\right). \tag{B.14}$$

□

C Statistical payment bounds

C.1 Report removal and finite-class concentration

Leave-one-observation-out comparisons are a common method used in the study of stability and learnability (Bousquet & Elisseeff, 2002; Elisseeff et al., 2005; Shalev-Shwartz et al., 2010). Report removal still uses the same library and temperature and results in the identity used for payment.

Summing (8) at truthful reports yields

$$\mathbb{E}_\omega R + T \sum_i \mathrm{KL}(q_{-i} \| q_S) = \sum_i v_i \left[\mathbb{E}_{q_S} a_i - \mathbb{E}_{q_{-i}} a_i\right], \tag{C.1}$$

where q_{-i} only omits the ith weight and uses the same temperature T. The library is independent of the auction sample, so that q_{-i} is independent of $Z_i = (X_i, Y_i, v_i)$. Exchangeability yields

$$\frac{1}{n}\mathbb{E}\left[R + T \sum_i \mathrm{KL}(q_{-i} \| q_S)\right] = \mathbb{E}\widehat{W}_S(q_S) - \mathbb{E}W(q_{S_{n-1};T}). \tag{C.2}$$

The second term uses the original temperature, but the two sample sizes differ.

For completeness, let $D_m = \sup_{h \in H} |\widehat{W}_m(h) - W(h)|$. An application of Hoeffding's exponential-moment inequality and a union bound over both signs gives

$$\mathbb{E}e^{\lambda D_m} \le 2K \exp\left(\frac{\lambda^2 V^2}{8m}\right).$$

Applying Jensen's inequality and minimizing over $\lambda > 0$ gives

$$\mathbb{E}D_m \le V\sqrt{\frac{\log(2K)}{2m}} = e_m. \tag{C.3}$$

The Gibbs variational identity is

$$T \log\left(\frac{1}{K}\sum_h e^{s_S(h)/T}\right) = \max_q \left\{\mathbb{E}_q s_S(h) - T\,\mathrm{KL}(q \| \mathrm{Unif}(H))\right\}.$$

Comparing with the point mass at a population-optimal classifier h^* gives

$$\widehat{W}_S(q_S) \ge \widehat{W}_S(h^*) - \frac{T \log K}{n}.$$

Two applications of (C.3) establish (13).

For payments, we drop the nonnegative reverse KL term in (C.2). Then

$$\mathbb{E}\widehat{W}_S(q_S) \le W^* + e_n, \qquad \mathbb{E}W(q_{S_{n-1};T}) \ge W^* - 2e_{n-1} - \frac{T\log K}{n-1}.$$

These estimates together imply the third bound in (14). The other two bounds follow from (B.9). □

C.2 High-probability welfare and growing libraries

Conditioning on a library that is independent of the sample used for the auction, we use Hoeffding's inequality and union bound to obtain

$$\Pr\left(D_n > V\sqrt{\frac{\log(2K/\delta)}{2n}}\right) \le \delta, \qquad D_n = \sup_{h\in H}|\widehat{W}_S(h) - W(h)|. \tag{C.4}$$

The Gibbs variational comparison is valid on every sample, so, together with (C.4), we have that, with probability at least $1-\delta$,

$$W^* - W(q_S) \le 2V\sqrt{\frac{\log(2K/\delta)}{2n}} + \frac{T\log K}{n}. \tag{C.5}$$

The probability distribution is over the training sample, and $W(q_S)$ already has the expectation taken over the selected classifier.

The same proof holds for any fixed sequence of independently constructed libraries H_n. Let $K_n = |H_n|$. Then it suffices that

$$\log K_n = o(n), \qquad T_n \log K_n = o(n)$$

for convergence in expected welfare to the optimum in H_n. Equation (14) also shows that the expected payment per participant vanishes under these two conditions. To see this, divide the statistical term by n to obtain $e_n + 2e_{n-1} + T_n \log K_n/(n-1)$, with the same K_n in both deviation terms. The leave-one-report-out comparison involves H_n and T_n on both sides. A population comparison to a larger classifier class also adds the approximation error $\sup_h W(h) - \max_{g\in H_n} W(g)$.

C.3 Proof of Theorem 3

Fix $h_0 \in H^*$. By beneficiary consensus, all $h \in H^*$ have the same value-weighted score almost surely. This equality also holds even after any one truthful report is decreased. Zero-value participants have zero score and pay zero.

For each nonoptimal classifier h, define $X_i(h) = v_i[a_i(h_0) - a_i(h)]$. Its range is $[-V, V]$ and its expectation is at least Δ. Hoeffding's inequality (Hoeffding, 1963) implies that

$$\Pr\left(\sum_i X_i(h) < \frac{n\Delta}{2}\right) \le e^{-n\Delta^2/(8V^2)}.$$

Taking a union bound over the M nonoptimal classifiers yields an event $\mathcal{E}$ of probability at least $1-\delta_n$ on which all gaps are at least $n\Delta/2$. If one report is removed, the gap decreases by at most V. Since $n \ge 4V/\Delta$, every counterfactual gap remains at least $n\Delta/4$. There are r equally scored optimal classifiers in the partition function, so

$$q_{i,t}(H \setminus H^*) \le \frac{M}{r}e^{-n\Delta/(4T)} = \eta_n.$$

The argument of Appendix B now applies with the optimal set replacing H_ε. On $\mathcal{E}$, we have

$$\bar{p}_i \le \min\left(V, \frac{V^2}{2T}\right)\eta_n, \qquad \Pr_\omega(p_i > 0) \le \min\left(1, \frac{V}{T}\right)\eta_n.$$

On $\mathcal{E}^c$, (B.9) gives

$$\bar{p}_i \le \min\left(V, \frac{V^2}{8T}\right), \qquad \Pr_\omega(p_i > 0) \le \min\left(1, \frac{V}{4T}\right).$$

Averaging gives (17) and (18). At the current reports, there is no report-removal penalty, so the same proof gives the population welfare bound

$$\mathbb{E}[W^* - W(q_S)] \le V \min\left\{1, \frac{M}{r} e^{-n\Delta/(2T)} + \delta_n\right\}. \tag{C.6}$$

If $M = 0$, then every positive-value participant has the same correctness outcome under every library member, whether correct or incorrect. The payment and population excess welfare are both zero.

Now let $T_n/n \to 0$. The bounds on the bad-event terms are $nV\delta_n$ and $n\delta_n$, both of which are exponentially small. When $T_n < V$, the bound on the good-event payment term is

$$nV(M/r)e^{-n\Delta/(4V)}.$$

When $T_n \ge V$, choose $x_n = n/T_n \to \infty$. The resulting bound is

$$\frac{V^2 M}{2r} x_n e^{-\Delta x_n/4} \longrightarrow 0.$$

The proof for the payer-count is the same. To finish, apply $\Pr(N_n > 0) \le \mathbb{E}N_n$. □

With fixed temperature the payment probabilities are summable in n. Therefore, for any joint construction with these sample-size marginal probabilities, the first Borel-Cantelli lemma implies that the positive payment events happen at only finitely many sample sizes almost surely. Thus the summability conclusion relies on fixed temperature. The expectation limits hold for every $T_n = o(n)$.

D Matching bounds for conflicting beneficiaries

D.1 Integral representation

Fix a participant, and orient labels so that their true label is positive. Let S_{n-1} be the sum of the other $n-1$ relative labels. These are independent fair signs. The correctness response derivative is

$$\alpha_i'(t) = \frac{1}{T}\sigma'\left(\frac{t + VS_{n-1}}{T}\right).$$

Averaging the payment integral gives (20). Replacing t/T by $1/T$ gives the corresponding expression for $\mathbb{E}N$.

D.2 Proof of Theorem 4

Upper bounds. When n is even, the S_{n-1} distribution has odd integer values. The largest atom size is

$$p_{\max} = 2^{-n}\binom{n}{n/2} \le \frac{1}{\sqrt{n+1}}.$$

The intervals $[Vs, Vs + V]$ for odd integers s are disjoint. Because $t \le V$,

$$\frac{\mathbb{E}R}{n} \le V p_{\max} \sum_{s \text{ odd}} \int_0^V \frac{1}{T}\sigma'\left(\frac{t + Vs}{T}\right) dt \le V p_{\max}.$$

Integrating the density over disjoint intervals, the total mass cannot exceed one, so $\mathbb{E}R \le V\sqrt{n}$. Omitting the t term, we have $\mathbb{E}N \le \sqrt{n}$. Temperature bounds come from $\sigma' \le 1/4$.

Low temperature. If $0 < T \le V$, keep $S_{n-1} = -1$ and integrate over $t \in [V - T/2, V]$. There $t \ge V/2$ and $\sigma'((t-V)/T) \ge \sigma'(1/2)$. Using the central binomial inequality $\Pr(S_{n-1} = -1) \ge 1/\sqrt{2n}$, we get

$$\mathbb{E}R \ge \frac{\sigma'(1/2)}{4\sqrt{2}} V\sqrt{n}. \tag{D.1}$$

Intermediate temperature. Since $V \le T \le V\sqrt{n}$ we may set $L = T/V$ and $m = n - 1$. By Chebyshev's inequality, $\Pr(|S_m| \le 2\sqrt{m}) \ge 3/4$. By symmetry and unimodality, central lattice points have at least the average mass of a larger central interval. The interval $|s| \le L$ contains at least $L/2$ admissible points. The interval $|s| \le 2\sqrt{m}$ contains at most $4\sqrt{n}$. Therefore

$$\Pr(|S_m| \le L) \ge \frac{3L}{32\sqrt{n}}.$$

On this event, and for $0 \le t \le V$, we have $|(t + VS_m)/T| \le 2$. Substituting into (20), we obtain

$$\mathbb{E}R \ge \frac{3\sigma'(2)}{64} V\sqrt{n}. \tag{D.2}$$

High temperature. For $T \ge V\sqrt{n}$, with probability at least $3/4$, the event $|S_m| \le 2\sqrt{m}$ still holds. The logistic argument has magnitude at most 3, so

$$\mathbb{E}R \ge \frac{3\sigma'(3)}{8} \frac{nV^2}{T}. \tag{D.3}$$

The minimum of the positive constants in (D.1)–(D.3) finishes the proof of the uniform lower bound

$$c_0 \min\left(V\sqrt{n}, \frac{nV^2}{T}\right) \le \mathbb{E}R \le \min\left(V\sqrt{n}, \frac{nV^2}{8T}\right).$$

This is sufficient to establish the payer-count lower bound, as $R \le VN$. Both constant classifiers have population welfare $V/2$, so every mixture classifier has zero excess welfare. □

The even-sample assumption determines the position of the lattice point at which the success threshold occurs at small T. The theorem does not introduce a continuous normal approximation in place of this discrete calculation.

D.3 Proof of Corollary 1

If $T_n/n \to 0$, then equation (13) gives consistency when K is fixed and finite. Conversely, let the two-constant classifier family with $v = V$ and $Y = +1$ almost surely be given. Then, we make the wrong choice with probability $\sigma(-nV/T_n)$. If $T_n/n \ge \epsilon > 0$ on some subsequence, then the excess welfare on this subsequence is at least $V\sigma(-V/\epsilon) > 0$. Thus, to give consistency for every distribution of i.i.d. samples, we must have $n/T_n \to \infty$.

In the fair-label population, both $V\sqrt{n}$ and nV^2/T_n diverge. Theorem 4 covers the even sample sizes. Using $T_n = n/g(n)$ with $g(n) \to \infty$ sufficiently slowly, the expected payment is of order $g(n)$. There is no uniform polynomial lower rate over all consistent schedules. □

D.4 Nonunique optimal models with consensus

Consider a three-class example with two different optimal models with the same correctness allocations. Let $Y = 0$, $v = V$, and $\Pr(X = 0) = 3/4$. Let the two models predict 0 when $X = 0$ and predict different incorrect labels, 1 and 2, when $X = 1$, respectively. Let another model always predict 1. The first two models are different, though both imperfect, and their correctness allocations agree everywhere. Both have welfare $3V/4$ while the third has welfare 0. Theorem 3 holds for $r = 2$, $M = 1$, and $\Delta = 3V/4$.

For this example and Theorem 4, there are multiple optimal classifiers. Their payment limits differ since one set of optimal classifiers agrees on correctness and the other does not.

D.5 Finite-sample formulas for biased labels

For r the number of positive labels in a sample of size n, with all values equal to V, the two constant classifiers have scores rV and $(n - r)V$. Define

$$a_r = \sigma\left(\frac{(2r - n)V}{T}\right), \qquad F_{V,T}(a) = Va + T\log(1 - a + ae^{-V/T}).$$

The conditional expectations of payment and number of payers are

$$\rho_r = rF_{V,T}(a_r) + (n-r)F_{V,T}(1-a_r), \tag{D.4}$$
$$\mu_r = rG_{V,T}(a_r) + (n-r)G_{V,T}(1-a_r), \tag{D.5}$$

where

$$G_{V,T}(a) = a - \frac{ae^{-V/T}}{1-a+ae^{-V/T}}.$$

With positive-label probability p, the unconditional expectations are obtained from

$$\mathbb{E}R = \sum_{r=0}^{n} \binom{n}{r} p^r (1-p)^{n-r} \rho_r, \qquad \mathbb{E}N = \sum_{r=0}^{n} \binom{n}{r} p^r (1-p)^{n-r} \mu_r. \tag{D.6}$$

The population welfare given r is $V[(1-p) + (2p-1)a_r]$. If $p \geq 1/2$ then its gap to the optimal constant classifier is $V(2p-1)(1-a_r)$. If $p \leq 1/2$ the gap is $V(1-2p)a_r$. For $p = 1/2$ both expressions are zero for every r.

D.6 Parity at vanishing temperature

The even-sample assumption in Theorem 4 relates to the discrete location of critical reports. The low-temperature limits make this dependence explicit. Given n, let $T \downarrow 0$ under the fair-label model.

If n is even, put $c_n = 2^{-n}\binom{n}{n/2}$. Then

$$\lim_{T\downarrow 0} \mathbb{E}R = \frac{nV}{2} c_n, \qquad \lim_{T\downarrow 0} \mathbb{E}N = \frac{n}{2} c_n. \tag{D.7}$$

To see this, write the integral representation (20). Since S_{n-1} is odd, only $S_{n-1} = -1$ puts a transition at the endpoint $t = V$. Its probability is c_n. For this value,

$$\int_0^V \frac{t}{T} \sigma'\left(\frac{t-V}{T}\right) dt = \int_{-V/T}^{0} (V + Tu)\sigma'(u)\, du \longrightarrow \frac{V}{2}.$$

The analogous integral for payer probability tends to $1/2$. Every other admissible value of S_{n-1} stays a positive distance from a transition, giving a vanishing contribution. Stirling's formula gives the subsequent sample-size limits $\mathbb{E}R \sim V\sqrt{n/(2\pi)}$ and $\mathbb{E}N \sim \sqrt{n/(2\pi)}$ in this order of limits.

Let $c_{n-1} = 2^{-(n-1)}\binom{n-1}{(n-1)/2}$ if n is odd. Then

$$\lim_{T\downarrow 0} \frac{\mathbb{E}R}{nT} = c_{n-1} \log 2, \qquad \lim_{T\downarrow 0} \mathbb{E}N = \frac{n}{2} c_{n-1}. \tag{D.8}$$

S_{n-1} is even, and the transition at $S_{n-1} = 0$ is at $t = 0$, with the payment contribution

$$T \int_0^{V/T} u\sigma'(u)\, du.$$

By integration by parts, $\int_0^\infty u\sigma'(u)\, du = \log 2$. The payer contribution is $\sigma(V/T) - 1/2 \to 1/2$. The other terms decay exponentially in $1/T$. The odd samples can have positive payer probability as the positive bills themselves go to zero.

E Dependence and aggregate payment tails

E.1 Proof of Theorem 5

At the balanced profile, set all other bids to V. A positive-label participant has response $\alpha_i(t) = F(t-V)$. The response of a negative-label participant is the same because $F(-z) = 1 - F(z)$. Theorem 1 implies

$$\Pr(p_i > x) = \frac{1}{2} - F(x-V), \qquad 0 \leq x < V.$$

A shared classifier makes one label group correct. The losing group pays zero. Equal billing assigns the same payment to each of the m winners. $R > mx$ is the disjoint union of the events that a representative of the positive group or a representative of the negative group pays more than x. The sum of their probabilities is $1 - 2F(x - V)$, which is (22).

For $F(z) = \sigma(z/T)$, the tail is $\tanh((V - x)/(2T))$. Integrating yields

$$\mathbb{E}_\omega R = m \int_0^V \tanh \frac{V - x}{2T}\, dx = nT \log\cosh \frac{V}{2T}.$$

Furthermore, $N = mB$, with $B \sim \mathrm{Ber}(\tanh(V/(2T)))$. For $T = n\tau$, one has $\tanh z \sim z$ and $\log\cosh z \sim z^2/2$, leading to (24). In particular,

$$\Pr(R > nV/4) = \tanh \frac{V}{4n\tau}, \qquad \mathrm{Var}_\omega(N) = \frac{n^2}{4} p_n(1 - p_n) = \Theta(n),$$

where $p_n = \tanh(V/(2n\tau))$. □

E.2 Proof of Theorem 6

Tail bounds can be deduced from the bounded-variable exponential-moment argument that is used for Bennett's inequality (Bennett, 1962).

Each participant's block-specific Gumbel vector has the same distribution as the shared mechanism's vector. The correctness and critical payment are both functions of this vector at a fixed report profile. Replacing the shared draw with the block's draw thus does not change their joint distribution for that participant. The threshold proof still applies to each participant for any fixed set of block vectors. This proves universal truthfulness.

Let X_g be the number of payers in block g. These random variables are conditionally independent given the sample and partition and take values in $[0, s]$. For $\lambda \geq 0$ by convexity

$$e^{\lambda X_g} \leq 1 + \frac{X_g}{s}(e^{\lambda s} - 1).$$

Taking expectations and multiplying over blocks yields

$$\mathbb{E}_\omega e^{\lambda N} \leq \exp\left[\frac{\mu}{s}(e^{\lambda s} - 1)\right].$$

Markov's inequality with $\lambda = s^{-1}\log(x/\mu)$ implies the first bound in (25). The second follows by replacing X_g with the sum of payments in block g, which lies in $[0, sV]$.

A useful inverted form is implied by Bernstein's inequality (Boucheron et al., 2013). The block variances satisfy

$$\sum_g \mathrm{Var}(X_g) \leq s\mu, \qquad \sum_g \mathrm{Var}(R_g) \leq sV\rho.$$

For $u = \log(1/\delta)$, each of the inequalities

$$N \leq \mu + \sqrt{2s\mu u} + \frac{2su}{3}, \qquad R \leq \rho + \sqrt{2sV\rho u} + \frac{2sVu}{3} \tag{E.1}$$

holds with probability at least $1 - \delta$. A simultaneous statement can be made by applying the union bound with $\delta/2$ for each event. If the partition is itself random and independent of the reports, the conditional means remain μ, ρ, so averaging preserves these bounds. □

E.3 Proof of Corollary 2

For blocks of size one on the balanced profile,

$$N \sim \mathrm{Bin}(n, p_n), \qquad p_n = \frac{1}{2}\tanh \frac{V}{2n\tau}, \qquad np_n \longrightarrow \frac{V}{4\tau} > 0.$$

For $k = O(\log n/\log\log n)$, expansion of the binomial mass and Stirling's formula give

$$\log \Pr(N = k) = -k \log k + O(k), \tag{E.2}$$

with errors controlled by $k^2/n \to 0$ and np_n converging to a positive constant. Fix $\epsilon \in (0, c)$. For

$$k_+ = \left\lceil (c+\epsilon)\frac{\log n}{\log\log n} \right\rceil, \qquad k_- = \left\lfloor (c-\epsilon)\frac{\log n}{\log\log n} \right\rfloor,$$

Chernoff's bound gives $\Pr(N \geq k_+) \leq n^{-(c+\epsilon)+o(1)} < n^{-c}$. Retaining the single mass at k_- in (E.2) gives $\Pr(N \geq k_-) > n^{-c}$. Letting $\epsilon \downarrow 0$ proves the count quantile in (26).

For any fixed $r \in (0, V)$, let $M_r = \sum_i \mathbf{1}\{p_i > r\}$. Given singleton deployment this is binomial with

$$n \Pr(p_i > r) = \frac{n}{2} \tanh \frac{V-r}{2n\tau} \longrightarrow \frac{V-r}{4\tau} > 0.$$

Hence the same estimate applies for quantiles of M_r. From $rM_r \leq R \leq VN$, the constant for R lies between cr and cV. Passing $r \uparrow V$ proves the payment quantile.

For shared equal billing, we have $\Pr(N = n/2) = \Theta(1/n)$; since $c > 1$, for n sufficiently large, this is greater than n^{-c}, thus $Q_{1-n^{-c}}(N) = n/2$. We can then solve the tail for payments as

$$Q_{1-n^{-c}}(R) = \frac{n}{2}\left[V - 2n\tau \operatorname{artanh}(n^{-c})\right] \sim \frac{nV}{2}.$$

□

E.4 A COMPOUND POISSON LIMIT FOR INDEPENDENT BILLS

Independent positive-payment indicators satisfy the rare event assumptions underlying the Poisson approximation (Le Cam, 1960) and the payment amounts define the jump distribution below.

Consider balanced profiles with $T = n\tau$, where V and τ are fixed. Then for the singleton BAA, all payment variables are independent, with a positive-payment density

$$f_n(x) = \frac{1}{n\tau}\sigma'\left(\frac{x-V}{n\tau}\right), \qquad 0 < x < V. \tag{E.3}$$

We have $nf_n(x) \to 1/(4\tau)$ uniformly over the interval. This means that the number of positive bills tends in law to a Poisson variable with mean $\lambda = V/(4\tau)$, and a positive bill converges in law to $\mathrm{Unif}(0, V)$.

The joint limit follows immediately from the transform. For $0 \leq z \leq 1$ and $\theta \geq 0$,

$$\begin{aligned}\mathbb{E}_\omega[z^N e^{-\theta R}] &= \left[1 + \int_0^V (ze^{-\theta x} - 1) f_n(x)\, dx\right]^n \\ &\longrightarrow \exp\left[\frac{1}{4\tau}\int_0^V (ze^{-\theta x} - 1)\, dx\right].\end{aligned} \tag{E.4}$$

Thus

$$(N, R) \implies \left(M, \sum_{j=1}^{M} U_j\right), \quad M \sim \mathrm{Poisson}(\lambda), \quad U_j \overset{\text{iid}}{\sim} \mathrm{Unif}(0, V), \tag{E.5}$$

where M and the U_j are independent, and where we interpret the empty sum as zero.

The same finite-sample integrals show the limiting moments

$$\begin{aligned}\mathbb{E}_\omega R &\longrightarrow \frac{V^2}{8\tau}, & \mathrm{Var}_\omega(R) &\longrightarrow \frac{V^3}{12\tau}, \\ \mathrm{Var}_\omega(N) &\longrightarrow \frac{V}{4\tau}, & \mathrm{Cov}_\omega(N, R) &\longrightarrow \frac{V^2}{8\tau}.\end{aligned} \tag{E.6}$$

For instance, $n\mathbb{E}p_i^2 \to V^3/(12\tau)$ by integrating $x^2 n f_n(x)$, and $n(\mathbb{E}p_i)^2 \to 0$. These are statements about a fixed balanced sample and independent deployment. They concern the full limiting distribution of the bill used in the quantile calculation of Corollary 2.

E.5 Second moments under a shared model

The same balanced profile leads to different aggregate second moments, under shared equal billing. We write $m = n/2$ and $\pi_n = \tanh(V/(2n\tau))$. By Theorem 5, $N = mB_n$ with $B_n \sim \mathrm{Ber}(\pi_n)$ so

$$\mathrm{Var}_\omega(N) = \frac{n^2}{4}\pi_n(1-\pi_n) \sim \frac{nV}{8\tau}. \tag{E.7}$$

Integration of the tail in (23) gives

$$\mathbb{E}_\omega R^2 = 2m^2 \int_0^V x \tanh\left(\frac{V-x}{2n\tau}\right) dx \sim \frac{nV^3}{24\tau}. \tag{E.8}$$

The expectation converges to $V^2/(8\tau)$ and its square is bounded. Thus $\mathrm{Var}_\omega(R) \sim nV^3/(24\tau)$.

The individual laws are identical in the shared and the singleton cases. Summing the individual variances and expanding the aggregate square gives

$$\mathrm{Var}_\omega(R_{\mathrm{shared}}) - \mathrm{Var}_\omega(R_{\mathrm{singleton}}) = 2\sum_{i<j} \mathrm{Cov}_{\mathrm{shared}}(p_i, p_j). \tag{E.9}$$

The order-n variance term in (E.8) is generated by this sum. The individual covariances can have either sign for arbitrary libraries. The conditional concentration bounds in Theorem 6 do not rely on a common sign being assigned.

E.6 Unequal blocks and weighted aggregate payments

The block-size bound may also be stated in terms of each block's own range. Let $d_i \geq 0$ be report-independent coefficients, and let

$$R_d = \sum_i d_i p_i, \quad Y_g = \sum_{i \in I_g} d_i p_i, \quad B_g = \sum_{i \in I_g} d_i b_i, \quad r_g = \mathbb{E}_\omega Y_g.$$

Then $0 \leq Y_g \leq B_g$, and terms with $B_g = 0$ vanish. The conditional independence of the block randomizations and the convex secant bound imply, for $\lambda \geq 0$,

$$\log \mathbb{E}_\omega e^{\lambda R_d} \leq \sum_{g: B_g > 0} \frac{r_g}{B_g}(e^{\lambda B_g} - 1), \tag{E.10}$$

and optimizing Markov's inequality yields

$$\Pr(R_d \geq x \mid S, b) \leq \inf_{\lambda \geq 0} \exp\left[-\lambda x + \sum_{g: B_g > 0} \frac{r_g}{B_g}(e^{\lambda B_g} - 1)\right]. \tag{E.11}$$

The associated variance bound is $\mathrm{Var}_\omega(R_d) \leq \sum_g B_g r_g$. Setting every $d_i = 1$ and replacing each B_g by sV yields the main-text bound. Group indicators $d_i \in \{0, 1\}$ yield bounds on the aggregate payments of a given subpopulation. The coefficients enter the reported risk statistic but not the participants' utilities or the allocation rule.

E.7 Deployment variance and test accuracy

Let m iid observations be split independently into blocks with sizes $m_1, \ldots, m_G$. Each block independently draws a classifier from the learned distribution q. Put

$$U(h, Z) = v\mathbf{1}\{h(X) = Y\}, \qquad \sigma_h^2 = \mathrm{Var}_Z\, U(h, Z).$$

Conditional on the block model h_g, its benefit sum has mean $m_g W(h_g)$ and variance $m_g \sigma_{h_g}^2$. Applying the law of total variance, the variance conditional on q is

$$\mathrm{Var}(\text{block } g \mid q) = m_g \mathbb{E}_q \sigma_h^2 + m_g^2\, \mathrm{Var}_q\, W(h).$$

Summing independent blocks and dividing by m^2 we obtain

$$\mathrm{Var}(\widehat{W}_{\mathrm{dep}} \mid q) = \frac{\mathbb{E}_q \sigma_h^2}{m} + \frac{\sum_g m_g^2}{m^2}\, \mathrm{Var}_q\, W(h). \tag{E.12}$$

With maximum block size s, we can use Hoeffding's inequality and $\sum_g m_g^2 \le sm$ to show that

$$\Pr\left(|\widehat{W}_{\mathrm{dep}} - W(q)| \ge \eta \mid q\right) \le 2\exp\left(-\frac{2m\eta^2}{sV^2}\right).$$

In the case of two constant classifiers on independent fair test labels, condition on the entire vector of test predictions. Each correctness indicator is an independent $\mathrm{Ber}(1/2)$ variable, and this conditional law is independent of the prediction vector. Integration over any deployment coupling yields that the count has law $\mathrm{Bin}(m, 1/2)$.

F Additional mechanism properties

F.1 Smoothed bills

The guarantee for truthfulness in expectation works on expected utilities instead of individual outcomes, as used in randomized payment computation (Babaioff et al., 2015). Keep the same Gibbs allocation, and charge $\widetilde{p}_i = \bar{p}_i/\alpha_i(b_i)$ when correct and zero otherwise. The fee is zero when $\alpha_i(b_i) = 0$. The expected payment at any report is still $\bar{p}_i$, and so the expected utility is identical to that of the truthful threshold mechanism, establishing truthfulness in expectation for risk-neutral players.

By the payment integral, $\bar{p}_i \le b_i\alpha_i(b_i)$. By monotonicity,

$$\bar{p}_i \le \int_0^{b_i} \frac{t}{T}\alpha_i(t)\,dt \le \alpha_i(b_i)\frac{b_i^2}{2T}.$$

The fees charged for correct predictions are at most $\min(b_i, b_i^2/(2T))$. For $T = n\tau$, the realized total is at most $V^2/(2\tau)$, and truthful players are individually rational.

Universal truthfulness does not necessarily hold. Assume that both correctness groups exist and that $b_i > 0$. Then $\bar{p}_i > 0$. There is a set of random seeds with positive probability that already lead to a correct prediction at report zero. For these seeds, reducing the report to zero leaves the prediction correct and removes the positive smoothed fee. The fixed seed incentive constraint is violated.

F.2 Population welfare and participant surplus

With a fixed library and any fixed $T_n = o(n)$, the statistical term in (14) satisfies

$$\frac{\mathbb{E}R_n}{n} \le e_n + 2e_{n-1} + \frac{T_n \log K}{n-1} \longrightarrow 0.$$

Average charges can converge to zero even when total receipts diverge, as in the tied population of Theorem 4. Beneficiary consensus implies the stronger convergence $\mathbb{E}R_n \to 0$ because a population welfare gap creates sample-score differences of order n. Changing a report can affect such a difference by at most V.

Utility model values are expressed in a common unit, which measures the benefits from correct predictions. The outside option is zero, and (A.6) shows that the expected surplus of reporting participants is nonnegative. A free prediction service or a binding budget imposes an additional participation constraint. Payments retained by the platform are transfers and reported separately from the population prediction welfare and the participants' residual benefits. In blockwise deployment, the utility of an individual depends only on the model and the payment assigned to that individual. Preferences over the predictions of other participants are not captured by the utility model.

F.3 Fixed priors and duplicate classifiers

For a report-independent positive prior π, let $q_b(h) \propto \pi(h)e^{s_b(h)/T}$, and add $T\log\pi(h)$ to the perturbed score in Algorithm 1. The threshold argument and individual KL identities remain the same. The uniform prior used in our main results assigns mass $1/K$ to each member of the library.

The prior masses of those classifiers with the same correctness vector on the auction sample can be summed for computing the payment. The conditional model identities of these classifiers can be kept for evaluation on the test set. The response rule is changed when we assign a new uniform prior after aggregating the classifiers.

F.4 Prior-weighted variational and payment formulas

Consider a fixed prior $\pi(h) > 0$ and let

$$Z_\pi(b) = \sum_h \pi(h) e^{s_b(h)/T}, \qquad q_b^\pi(h) = \frac{\pi(h) e^{s_b(h)/T}}{Z_\pi(b)}.$$

Given a fixed reference prior as used in PAC-Bayesian model averaging (McAllester, 1999), the Gibbs variational identity (Wainwright & Jordan, 2008) is given by

$$T \log Z_\pi(b) = \max_q \{\mathbb{E}_q s_b(h) - T \,\mathrm{KL}(q\|\pi)\}. \tag{F.1}$$

This maximum is attained at q_b^π. This can be shown by writing

$$T \,\mathrm{KL}(q\|q_b^\pi) = T \,\mathrm{KL}(q\|\pi) - \mathbb{E}_q s_b(h) + T \log Z_\pi(b)$$

and using that the KL is always non-negative. The score perturbation $s_b(h) + T \log \pi(h) + TG_h$ realizes this distribution.

The prior weights remain the same on a reporting individual's path. Grouping correct and incorrect classifiers again gives a logistic response of slope $1/T$. Equations (7)–(9) and Theorem 1 hold with q^π instead of q. With population-optimal h^*, (F.1) compared to the point mass at h^* yields

$$\mathbb{E}[W^* - W(q_S^\pi)] \le 2e_n + \frac{T \log(1/\pi(h^*))}{n}. \tag{F.2}$$

For several population-optimal classifiers, the one with the largest prior probability yields the most restrictive of these point-mass bounds.

Classifiers with the same correctness vector on the auction sample receive the same multiplicative likelihood factor at any report profile. Such classifiers can be represented by a class whose prior mass is the sum of the prior masses of the original classifiers. The common likelihood factor and normalization give the combined posterior mass. Conditional on choosing this class, choose an original classifier in proportion to its prior mass. This recovers the original model marginal, including the predictions on test inputs. Aggregation affects only the computation.

F.5 Information disclosed by model draws

Order two report profiles so that $b_i - b_i' = \delta \in [0, V]$, with all other reports equal. Increasing the bid from b_i' to b_i changes every model score by an amount in $[0, \delta]$, so the change in the log partition function is in $[0, \delta/T]$. For each model,

$$\left|\log \frac{q_b(h)}{q_{b'}(h)}\right| \le \frac{V}{T}.$$

For G independent draws, likelihood ratios multiply, so their logarithms add and the absolute log-likelihood-ratio bound is GV/T. This is the model-identity likelihood-ratio calculation used by the exponential mechanism (McSherry & Talwar, 2007), here under bid adjacency. The bound applies to the release of model identities. Payment records require a separate likelihood analysis of the transaction mechanism.

G Experimental details and additional results

G.1 Data preparation and independent splits

The learned model experiment is Adult and the replication is Bank Marketing. We use the Adult income label and the term-deposit subscription label from `bank-full.csv` (Becker & Kohavi, 1996; Moro et al., 2014). We combine the two source partitions from Adult, remove trailing punctuation from the income labels, and remove duplicates from both labeled datasets. We do not use Adult's `fnlwgt` and Bank's call duration variable. The latter is only available after the call whose outcome is predicted.

Missing categorical values are represented as a separate category. Numeric columns are median-imputed and standardized. Categorical columns are one-hot encoded, and unknown categories are ignored at application time. The preprocessing is fitted only to the construction partition.

Random seeds $0, 1, 2, 3, 4$ are used for five random splits. 40% of the data are used for construction, 20% for reference, 20% for the auction pool, and 20% for the final test. A seeded permutation of the auction pool is used as the default pool of $n = 2048$ participants. Sample size comparisons use the prefixes of the same permutation with sample sizes 512 and 1024. The model library and test partition are unchanged in sample size comparisons. The reference partition is used for the median age of the covariate associated value assignment. Auction bids and test labels are not used to set model library parameters or temperature.

G.2 Classifier library

We construct a uniformly weighted library of $K = 16$ classifiers from eight fitted estimators. Four are L_2-regularized logistic regressions with inverse regularization $C \in \{10^{-2}, 10^{-1}, 1, 10\}$. The other four are CART probability estimators with maximum depths $\{2, 3, 4, 6\}$. Logistic regression uses an intercept, no class weighting, L-BFGS, tolerance 10^{-8}, and at most 3000 iterations. CART uses Gini impurity and minimum leaf size 50. Each estimator yields two deterministic classifiers by applying positive-class thresholds 0.25 and 0.5. The logistic-only and tree-only ablations use eight classifiers each, without fitting any new estimator.

The full evaluation contains 80 estimator fits on two data sets with five splits each. Cached predictions can be used for all mechanism conditions. We use Python 3.11, NumPy 1.26.4 (Harris et al., 2020), SciPy 1.13.1 (Virtanen et al., 2020), and scikit-learn 1.5.2 (Pedregosa et al., 2011). The seeds for splitting, fitting, partitioning, and mechanism sampling are generated independently based on the split seed.

G.3 Experimental configuration

Table 3 shows the fixed choices used in the learned-model evaluation. Temperature and deployment configurations use the same cached correctness matrices. The independent reference partition fixes the age threshold before values are set for auction and test observations.

Table 3: Configuration of the learned-model experiments. The settings for Adult and Bank Marketing are the same. Comparisons between model families keep the estimators set in advance for the entire library. We reuse the fitted estimators when changing temperature, values, or block size.

Component	Setting
Data splits	Five splits with seeds $0, 1, 2, 3, 4$
Partition proportions	40% construction, 20% reference, 20% auction, 20% test
Default auction size	$n = 2048$ with nested sizes 512 and 1024
Logistic regressions	$C \in \{0.01, 0.1, 1, 10\}$, L-BFGS, L_2 regularization
Optimization	Tolerance 10^{-8} and at most 3000 iterations
Decision trees	Depths $\{2, 3, 4, 6\}$, Gini impurity, minimum leaf size 50
Classification thresholds	0.25 and 0.5 for each of the eight estimators
Library prior	Uniform over the resulting 16 classifiers
Temperature comparison	$T \in \{1, \sqrt{n}, 0.1n\}$
Deployment comparison	Maximum block sizes n, 256, 32, and 1
Deployment repetitions	4096 draws per split and condition
Risk summaries	Left 0.99 quantile and 99% expected shortfall

All eight fitted estimators will use the construction-set preprocessing. Model families comparisons will also use this transformation. Neither fitting nor threshold selection will rely on current value reports. The two thresholds for probability classification are part of the definition of each classifier, thus all the auction results will be determined by selecting one of the deterministic classifiers.

G.4 VALUES AND BASELINES

Values assume $V = 1$ in all three cases:

$$\begin{aligned} v_{\text{uniform}} &= 1, \\ v_{\text{label}} &= 0.25 + 0.75\mathbf{1}\{Y = 1\}, \\ v_{\text{age}} &= 0.25 + 0.75\mathbf{1}\{\text{age} \geq m_{\text{ref}}\}, \end{aligned} \tag{G.1}$$

where m_{ref} is the median age in the reference set. The same function is used in the auction and test sets. These assignments set utility weights in the model. The near-optimal disagreement fraction in the value comparison is $n^{-1}\sum_i I_i^{\varepsilon}$, where $\varepsilon = 0.01n$ and H_ε is evaluated at truthful reports.

The comparison rules are described in Table 4. All use the same library and fixed model tie ordering. Accuracy-only selection ignores bids. In free Gibbs, every positive value participant weakly prefers to report the cap because correctness is non-decreasing in the report and there is no fee. The capped report profile is the same across all value assignments. Welfare ERM uses the deterministic version of Algorithm 1, with the perturbations removed. The smoothed-bill construction is described in Appendix F.

Table 4: Comparison rules on the fixed classifier library.

Method	Allocation and payment	Incentive guarantee
Accuracy-only	Maximum unweighted accuracy, zero payment	Report-independent
Welfare ERM	Maximum reported welfare, critical payment	Dominant-strategy truthful
Free Gibbs	Gibbs at capped reports, zero payment	Cap reporting is weakly dominant
Shared BAA	Gibbs with one model draw, critical payment	Universally truthful
Blockwise BAA	Gibbs with independent block draws, critical payment	Universally truthful
Singleton BAA	Gibbs with independent participant draws, critical payment	Universally truthful
Smoothed-bill Gibbs	Shared Gibbs, conditional mean fee	Truthful in expectation

G.5 EXPECTED OUTCOMES AND TAIL ESTIMATION

Test welfare is computed using summation with the chosen model distribution

$$\widehat{W}_{\text{test}}(q) = \frac{1}{m_{\text{test}}} \sum_j v_j \sum_h q(h) a_j(h).$$

Test-library gap is defined as $\max_h \widehat{W}_{\text{test}}(h) - \widehat{W}_{\text{test}}(q)$. The maximization is an evaluation quantity, and does not choose the deployed model. Effective library size is $\exp(-\sum_h q(h)\log q(h))$.

The default risk comparison uses label associated values with $n = 2048$ participants and temperature $T = 0.1n$. We use block sizes $n, 256, 32, 1$. The partition is chosen by randomly permuting participant identifiers independently of bids. Each split and condition has 4096 deployments. We analytically compute the expected payment and payer count using (9) and report means and standard deviations of five split-specific estimates. Repeated deployments are used to measure conditional risk for a given sample and do not provide additional independent training samples.

We use the left 0.99 quantile $Q_{0.99}(R) = \inf\{r : \Pr(R \leq r) \geq 0.99\}$. To estimate expected shortfall in (29), we use $\mathbb{E}[(R-t)_+] = \rho - \mathbb{E}[\min(R,t)]$ and minimize the resulting objective with the analytic mean and an empirical estimate of the truncated payment term. This preserves the known mean rather than estimating it from draws that may be dominated by a few large bills. The two risk measures describe different features of the distribution: a zero 0.99 quantile can coexist with positive expected shortfall.

G.6 Quantiles, expected shortfall, and a zero-bill atom

Let R be a nonnegative payment variable. Let $\rho = \mathbb{E}R$ and $u \in (0,1)$. The expected shortfall (Artzner et al., 1999; Acerbi & Tasche, 2002) is a coherent risk measure with a variational representation (Rockafellar & Uryasev, 2000)

$$\mathrm{ES}_u(R) = \inf_{t\in\mathbb{R}} \left\{ t + \frac{\mathbb{E}[(R-t)_+]}{1-u} \right\}. \tag{G.2}$$

A minimizer can be selected from among the u-quantiles. Nonnegativity allows one to restrict attention to $t \geq 0$, since for $t < 0$, the objective is $\rho/(1-u) - ut/(1-u)$, which decreases as t approaches zero.

If $\Pr(R > 0) \leq 1 - u$, then

$$Q_u(R) = 0, \qquad \mathrm{ES}_u(R) = \frac{\rho}{1-u}. \tag{G.3}$$

For $t \geq 0$, $\mathbb{E}[\min(R,t)] \leq t \Pr(R > 0)$. Re-writing the objective

$$\frac{\rho}{1-u} + t - \frac{\mathbb{E}[\min(R,t)]}{1-u}$$

shows that it is at least $\rho/(1-u)$, which is achieved at $t = 0$. Equation (G.3) demonstrates that for a sparse-payment distribution one can have a zero quantile while still having positive expected shortfall.

For the balanced shared-model family, the full tail distribution also leads to an explicit finite-sample expected-shortfall expression. Take $m = n/2$ and $\delta = 1 - u$. If $\delta \geq \tanh(V/(2T))$, use (G.3). If $\delta < \tanh(V/(2T))$, set

$$x_\delta = V - 2T \operatorname{artanh}(\delta), \qquad Q_u(R) = m x_\delta.$$

Integrating the tail distribution in (23) leads to

$$\mathrm{ES}_u(R) = m x_\delta + \frac{2mT}{\delta} \log\cosh\big(\operatorname{artanh}(\delta)\big). \tag{G.4}$$

Both expressions use the full payment distribution, including its zero atom. In the learned-model experiments, (G.2) uses the analytic conditional mean and empirical truncated payments. Means and expected-shortfall values are rounded following split-specific computations.

G.7 Implementation of payment draws

The correctness matrices are cached as Boolean arrays. We calculate $s = C^\top b$, $q = \mathrm{softmax}(s/T)$ and $\alpha = Cq$ once per report profile. Log-sum-exp, `expm1`, and `log1p` functions ensure a stable closed-form calculation of the payments. Rows of constant correctness are assigned zero payments.

For singleton payment simulations, the threshold law offers a scalar representation. Let

$$\gamma_i = \log \sum_{k:C_{ik}=1} e^{(s_k - b_i)/T} - \log \sum_{k:C_{ik}=0} e^{s_k/T}.$$

Simulate independent $U_i \sim \mathrm{Unif}(0,1)$, and set

$$c_i = T[\mathrm{logit}(U_i) - \gamma_i], \qquad a_i = \mathbf{1}\{c_i \leq b_i\}, \qquad p_i = a_i [c_i]_+.$$

This implements the individual allocation–payment law of the singleton mechanism. For shared and intermediate-block conditions, use Algorithm 1. Within a block, records sharing the same correctness vector and bid share the same threshold, so we compute a single threshold and then multiply by their count. Population-oriented metrics are evaluated directly under q.

G.8 Computational cost of allocation and pricing

With fixed $C \in \{0,1\}^{n\times K}$, the score function $s = C^\top b$ takes $O(nK)$ operations. The computation of q also takes $O(K)$ and of all correctness probabilities Cq takes $O(nK)$. Then computing all conditional mean payments and payer probabilities in Equation (9) takes $O(n)$ additional operations.

In a particular instantiation of Algorithm 1, every block requires K draws from a Gumbel distribution, plus K scores and one maximum computation. There are at most n blocks. On two correctness groups for each participant, this amounts to $O(K)$ work. This yields $O(nK)$ total work for allocation and pricing, even for singleton blocks. For a single block, the algorithm uses $O(K)$ temporary storage, in addition to the pre-computed matrix and output arrays.

The scalar singleton construction reduces a payment-only realization to $O(n)$ operations after its response offsets have been computed. Grouping identical pairs (C_i, b_i) reduces the number of threshold calculations on shared and intermediate blocks. The multiplicity of a group scales the payment and the payer's contribution. These optimizations maintain the law in Theorem 1 as they use the same response curve and the same randomization within a block.

Fitting, prediction and mechanism evaluation are separately treated. The first two give the fixed correctness matrices, while reporting, temperature, and deployment are based on those matrices with all fitted parameters remaining unchanged.

G.9 Full payment-risk results

Table 5 reports the underlying payment statistics and payer-count statistics for Adult. The variance is 5.960 for shared deployment, compared to 0.064 for singleton deployment. There is a reduction in the expected shortfall from 10.23 to 8.51 to 3.27 and then to 1.10 with intermediate block sizes. The 0.99 quantile of the payer count is zero in the shared condition, 12.4 at block size 256, 4.8 at size 32, and 2.0 for singleton blocks. The nonmonotone shape of these quantiles is consistent with a rare event distribution with a fixed mean.

Table 5: Adult payment risk under the same individual response and payment laws. Values are the means± standard deviations over five splits.

Block size	$Q_{.99}(R)$	$\mathrm{ES}_{.99}(R)$	$\mathrm{Var}(R)$	$Q_{.99}(N)$
n	0.00 ± 0.00	10.23 ± 0.47	5.960 ± 1.500	0.0 ± 0.0
256	2.91 ± 0.62	8.51 ± 0.25	0.777 ± 0.227	12.4 ± 2.3
32	2.05 ± 0.10	3.27 ± 0.49	0.158 ± 0.006	4.8 ± 0.4
1	1.04 ± 0.04	1.10 ± 0.28	0.064 ± 0.005	2.0 ± 0.0

Table 6 shows results for all value assignments on both datasets. Independent deployment always improves expected shortfall in all six cases. When values are uniform, expected shortfall goes from 24.21 to 1.99 on Adult and from 16.01 to 1.88 on Bank Marketing. When values are age-dependent, expected shortfall goes from 16.38 to 1.54 on Adult and from 10.37 to 1.11 on Bank Marketing. Mean payments do not change with block size.

Table 6: Welfare and payment statistics for all value assignments at $n = 2048$ and $T = 0.1n$. Entries are means and standard deviations over five splits. Shared and singleton deployments have the same welfare, mean payment ρ, and mean payer count μ. The last two columns are the 99% expected shortfall of the total payment.

Dataset	Values	Test welfare	ρ	μ	Shared ES	Singleton ES
Adult	Uniform	$.8302\pm.0016$	$.242\pm.007$	$.484\pm.013$	$24.21\pm.67$	$1.99\pm.76$
	Label	$.3289\pm.0024$	$.102\pm.005$	$.270\pm.010$	$10.23\pm.47$	$1.10\pm.28$
	Age	$.5140\pm.0062$	$.164\pm.003$	$.365\pm.008$	$16.38\pm.28$	$1.54\pm.77$
Bank	Uniform	$.8755\pm.0053$	$.160\pm.009$	$.320\pm.017$	$16.01\pm.87$	$1.88\pm.38$
	Label	$.2465\pm.0019$	$.066\pm.006$	$.180\pm.014$	$6.62\pm.64$	$1.02\pm.29$
	Age	$.5451\pm.0029$	$.104\pm.007$	$.234\pm.014$	$10.37\pm.71$	$1.11\pm.58$

G.10 Sensitivity analyses

The temperature comparison uses $T \in \{1, \sqrt{n}, 0.1n\}$ for unchanged samples. Comparing $T = 1$ to $0.1n$ decreases the average total payment of Adult by 4.56 and increases the empirical test-welfare

gap of Adult by 0.0142. The effective library size increases by 12.8. Both report sensitivity and the mass placed on alternative models are affected, and changes need not have monotone finite sample effects.

The logistic-only and tree-only ablations have eight declared classifiers each. The test welfare changes relative to the full Adult library by $+0.0019$ and -0.0020, and the mean total payment changes by -0.017 and $+0.011$. The model subsets are fixed independently of the auction outcome. The 0.99 quantile of payments made with smoothed-bill Gibbs is 0.21 on Adult and 0.16 on Bank Marketing. Expected payments agree with BAA. Realized payments follow the construction for truthfulness in expectation (see Appendix F).

G.11 Paired temperature and library comparisons

Table 7 details the numerical quantities that relate the Adult sensitivity results. The temperature contrast holds the library, values, observations, and test set fixed. The two library contrasts hold temperature and observations fixed and keep one estimator family. Signs are defined as the difference between the comparison condition and the full library or lower temperature condition shown in the table.

Table 7: Paired sensitivity contrasts on Adult. The temperature contrast compares $T = 0.1n$ to $T = 1$ and shows the change in the test-welfare gap. The library contrast compares a subset of eight classifiers to the full set of 16 classifiers and shows the change in test welfare. K_{eff} is the exponential of the Gibbs entropy. Dashes denote quantities that are excluded from the associated contrast.

Comparison	$\Delta\rho$	ΔW_{test}	Δ welfare gap	ΔK_{eff}
$T = 0.1n$ versus $T = 1$	-4.56	–	$+0.0142$	$+12.8$
Logistic-only versus full	-0.017	$+0.0019$	–	–
Tree-only versus full	$+0.011$	-0.0020	–	–

The effect of higher temperature is to distribute probability mass over more library members and to reduce the slope of each logistic response. The empirical welfare gap and conditional mean payment are statistics that capture the combined effect of changes in these two quantities. Model family removal affects the set of correctness vectors and their prior weights under the uniform prior. The welfare and payment effects of model family removal depend on which predictions remain in the library.

Correctness vectors, model probabilities and participant allocation-payment law are left unchanged by the deployment. Aggregated payment risk is a function of how the model draws are shared.