THE INFORMATION COST OF ASYMMETRIC RISK CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

When prediction errors have unequal consequences, overall accuracy may conceal failures that warrant separate risk limits. We ask how the granularity of these limits changes the information a representation must retain. We study finite-message representations fixed before numerical risk budgets are specified, with frontier error defined as the largest increase in optimal prediction error across all budgets. For a regular class of classification problems with a common default that avoids severe errors and one competing action, we prove that grouping the same severe error events into m groups yields optimal M-message frontier error $\Theta(M^{-2/m})$. With the source and error events held fixed, finer control thus changes the rate at which additional messages recover decision performance, even when each grouping receives its own optimal representation. Separate limits make distinctions between error types relevant to action; we identify cases where proportional risks or a default action that is optimal at every input eliminate the added cost. Preserving the frontier further bounds the loss in optimal objective value when risk preferences are specified after encoding. Synthetic experiments illustrate these effects, while experiments on frozen, calibrated WiCE verifier outputs show a corresponding cost of separate overclaim control over the tested penalty family.

1 INTRODUCTION

A prediction system can be accurate overall while making too many of the errors with the most serious consequences. Consider an *overclaim*: a system certifies a conclusion that goes beyond its evidence. In drug research, this could mean certifying a treatment's survival benefit based only on evidence that it suppresses an abnormal heart rhythm (CAST Investigators, 1989). In credit assessment, it could mean certifying a borrower's ability to repay based solely on the value of pledged collateral, without sufficient evidence about the borrower's cash flows (Basel Committee on Banking Supervision, 2025). Certifying the stronger claim can encourage treatment or investment on an inadequate evidential basis. Withholding certification can also be costly, but may leave an opportunity for further verification. When these consequences are asymmetric, false certification warrants a separate risk limit alongside the objective of accurate prediction.

Even within false certification, different errors may warrant different limits. In evidence verification, a claim may be supported, partially supported, or not supported by its cited evidence, as in WiCE (Kamoi et al., 2023). Predicting *supported* in either of the latter two cases overstates what the evidence establishes. A deployment concerned with overall overclaiming may cap their combined frequency. A common cap allows a reduction in one error to offset an increase in the other. Separate caps instead enforce an individual limit on each type when the two failures have different consequences. The source and error events are unchanged, but the second requirement calls for more detailed control. We ask how this choice changes the information that a representation must retain.

This distinction matters when a representation is fixed before the deployment specifies its risk budgets. The representation may be designed for the chosen error groups, while the decision rule adapts to the numerical budgets supplied later. Such reuse across downstream constraints is central to constrained omniprediction (Hu et al., 2023). To remain useful across these choices, the representation should support decisions close to the best achievable with the original information at every budget. The *risk frontier* records this best achievable prediction error as the budgets vary. Drawing on the decision-based view of quantization (Varshney & Varshney, 2008), we assess a finite-message rep-

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.

resentation by its largest increase in optimal prediction error across that frontier. This captures the information needed to keep future risk tradeoffs available.

Our central finding is that error grouping can change the information cost of retaining this flexibility. Under a shared budget, a representation needs to preserve the total severe-error risk relative to the benefit of acting. Separate budgets also require it to preserve how that risk is distributed across error types. Information that aggregate control can discard may therefore become essential under finer control. Under explicit source and action conditions, we quantify this change through matching approximation rates: finer control can require substantially more messages to achieve the same frontier accuracy, even when each grouping receives its own optimal representation. The same analysis explains why additional risk limits need not always demand more information. When conditional risks vary in fixed proportions, or the relevant distinctions leave the optimal action unchanged, finer control can impose no additional representation cost.

Preserving the frontier also supports preferences specified after encoding. When a deployment chooses a budget by balancing prediction error against an external risk cost, the frontier error bounds the loss in its optimal objective value. Synthetic experiments illustrate how grouping changes representation requirements, when redundancy removes that cost, and how using the grouping can improve on a posterior-quantization baseline. Experiments on frozen, calibrated WiCE verifier outputs show a corresponding cost of separate overclaim control over the tested penalty family.

2 Related Work

Risk-constrained classification accommodates unequal error consequences. Neyman–Pearson methods minimize one error subject to a bound on another (Scott & Nowak, 2005; Rigollet & Tong, 2011). Tong et al. (2018) develop score-thresholding algorithms with finite-sample guarantees and NP-ROC bands. Hierarchical and multiclass extensions prioritize severe mistakes and impose class-specific error constraints, including cost-sensitive formulations (Wang et al., 2024; Tian & Feng, 2025). More general confusion-matrix frameworks accommodate objectives and constraints involving multiple error types through randomized classifiers and cost-sensitive subproblems (Narasimhan, 2018; Narasimhan et al., 2024).

Risk control after prediction adapts predictive systems to specified performance requirements. Risk-controlling prediction sets provide high-probability population-risk guarantees, including losses with class-dependent error costs (Bates et al., 2021). Conformal risk control guarantees expected risk for bounded monotone losses and supports multiple risk requirements (Angelopoulos et al., 2024). Learn then Test uses hypothesis testing to select parameters satisfying simultaneous risk constraints (Angelopoulos et al., 2025). MultiRisk incorporates user-defined priorities through iterative score thresholding, with simultaneous risk control and objective guarantees under additional conditions (Joshi et al., 2025). Online methods regulate time-averaged losses, including multiple risks, as predictions and observations arrive (Feldman et al., 2023).

Reusable prediction separates predictor construction from subsequent decision requirements. Omnipredictors support loss-dependent post-processing that competes with a hypothesis class across a family of losses (Gopalan et al., 2022). Decision calibration links forecasts to downstream action choices (Zhao et al., 2021), while projected smooth calibration provides guarantees under label aggregation (Gopalan et al., 2024). Hu et al. (2023) extend omniprediction to objectives and constraints supplied after predictor construction. Their framework includes finite-range predictors and post-processing with specified group information, making it particularly relevant to fixed representations for subsequent constrained decisions. Online omniprediction with long-term constraints further addresses downstream regret and cumulative constraint violation (Bechavod et al., 2026).

Statistical experiment comparison evaluates information through attainable decision performance. Blackwell (1953) compares experiments across decision problems, while Le Cam (1964) develops approximate sufficiency and deficiency. Strasser (2000) studies finite approximations of statistical experiments from this perspective. Reid & Williamson (2011) connect Bayes risk, Neyman–Pearson trade-offs, divergences, and information in binary experiments; van Rooyen & Williamson (2014) apply deficiency to feature learning across downstream decision problems. We study the finite-message complexity of preserving an entire constrained risk frontier when the encoder is fixed before budget selection, comparing different groupings of the same error events under a fixed source.

3 PROBLEM SETUP

We study how the grouping of severe errors changes the information needed to preserve decision performance. For each grouping, a representation is chosen once, and the decision rule adapts to numerical risk budgets specified later.

3.1 PREDICTION AND RISK GROUPING

Let the labels and actions be $\{1, \dots, k, S\}$, with fixed $k \geq 1$, and write $[k] = \{1, \dots, k\}$. For an input X, let $W_i = \Pr(Y = i \mid X)$, $i \in [k]$, denote its conditional label probabilities. We assume the population distribution is known and take this posterior vector as the input to the representation problem, so that

$$P(Y = i \mid W = w) = w_i \quad (i \leq k), \qquad P(Y = S \mid W = w) = 1 - \sum_{i=1}^{k} w_i. \tag{1}$$

For an action A, the primary risk is $R_0 = \Pr(A \neq Y)$. The severe error events are fixed throughout the comparison: $R_i = \Pr(Y = i, A = S)$, $i \in [k]$. For example, predicting S on true labels 1 and 2 gives two distinct errors. A shared budget limits $R_1 + R_2$; separate budgets control their frequencies individually. More generally, a partition $\mathcal{G} = \{G_1, \dots, G_m\}$ of $[k]$ into nonempty groups specifies which error types share a budget. Its membership matrix $B \in \{0,1\}^{m \times k}$ has $B_{gi} = \mathbf{1}\{i \in G_g\}$, and

$$R_g^{\mathcal{G}} = \sum_{i \in G_g} R_i, \qquad q_{\mathcal{G}}(w) = Bw. \tag{2}$$

Here $q_g(w) = \Pr(Y \in G_g \mid W = w)$ is the conditional probability of a label in group g, which produces a severe error if the action is S. Only the partition changes; the source, actions, primary loss and underlying events remain the same.

3.2 REPRESENTATIONS FIXED BEFORE RISK BUDGETS

We compress the posterior vector W into a message Z with at most M possible values, allowing randomized encoding.

Definition 3.1 (Budget-independent M-message representation)**.** An encoder is a measurable Markov kernel $Q(z \mid w) = \Pr(Z = z \mid W = w)$ with $Y - W - Z$ and message alphabet size $|\mathcal{Z}| \leq M$. It may depend on the source and partition, but is fixed before the budget is supplied. The decoder observes only Z and the budget α; it may randomize and adapt its decision rule to α.

The full-information benchmark observes W. For $I \in \{W, Z\}$, a measurable randomized policy assigns action probabilities $\pi_a(I) \geq 0$ with $\sum_a \pi_a(I) = 1$. Its grouped risk is

$$R_g^{\mathcal{G}}(\pi) = \mathbb{E}[q_g(W)\pi_S(I)].$$

3.3 PRESERVING THE ENTIRE RISK FRONTIER

For a fixed grouping and budget $\alpha \in [0,1]^m$, we compare the smallest primary risk attainable from W with that attainable from Z. At information level I, this frontier is

$$F_I^{\mathcal{G}}(\alpha) = \inf_{\pi:\, R^{\mathcal{G}}(\pi) \leq \alpha} R_0(\pi). \tag{3}$$

The constraint is coordinatewise, and the policy is optimized separately for each budget. Every budget is feasible by choosing action 1, and $F_Z^{\mathcal{G}} \geq F_W^{\mathcal{G}}$. Their gap isolates the additional prediction error caused by compression, relative to full information under the same risk constraints.

Definition 3.2 (Full-frontier representation error)**.** For a fixed encoder Q, let $D_{\mathcal{G}}(Q)$ be its largest increase in optimal primary risk over all budgets. The optimal M-message error is

$$\begin{aligned} D_{\mathcal{G}}(Q) &= \sup_{\alpha \in [0,1]^m} \big(F_Z^{\mathcal{G}}(\alpha) - F_W^{\mathcal{G}}(\alpha)\big), \\ \mathcal{E}_M^{\mathcal{G}} &= \inf_{Q:\, |\mathcal{Z}| \leq M} D_{\mathcal{G}}(Q). \end{aligned} \tag{4}$$

A representation with $D_{\mathcal{G}}(Q) \le \varepsilon$ supports a feasible decision rule at every budget whose primary risk is at most ε above the full-information optimum. We ask how many messages are needed to provide this guarantee under different error groupings.

3.4 Source and Action Structure

Our source class combines a common tradeoff between two actions with variation in the severe-error probabilities.

Assumption 3.3 (Nondegenerate posterior source)**.** The law of W has density f with respect to k-dimensional Lebesgue measure, supported on a compact convex body $K \subset \mathcal{O}_k$ with nonempty k-dimensional interior, where

$$\mathcal{O}_k = \left\{ w : w_i > 0 \ (i \le k), \quad w_1 > w_i \ (i > 1), \quad \Delta(w) := 1 - \sum_i w_i - w_1 > 0 \right\}. \tag{5}$$

There are fixed constants $0 < f_- \le f_+ < \infty$ such that $f_- \le f(w) \le f_+$ almost everywhere on K. All source parameters are independent of M.

The conditional primary risks of actions 1 and S are $r_1(w) = 1 - w_1$ and $r_S(w) = \sum_i w_i$, so $\Delta(w) = r_1(w) - r_S(w)$. Action 1 is the best non-S action and incurs zero severe risk. Choosing S improves the primary risk by $\Delta(w) > 0$ while potentially making a severe error. Thus only actions 1 and S compete. The full-dimensional source lets the conditional label probabilities W_i vary in all k directions. These action and source conditions underpin the grouping comparison below.

4 Theory: The Information Cost of Risk Grouping

We first compare shared and separate budgets for two errors, then establish the general grouping law. Refinement and redundancy explain when finer control creates additional information requirements; the final result connects frontier preservation to later risk preferences. Detailed proofs are in Appendices A–C.

4.1 Merged versus Separate Error Control

We fix a source satisfying Assumption 3.3 and compare shared and separate budgets for its two severe errors. Consider labels N, P, S with posterior $(U, V, 1-U-V)$, where (U, V) has a fixed positive bounded density on $[0.19, 0.21] \times [0.29, 0.31]$. Here the default action 1 is P, action 2 is N, and $W = (V, U)$. The two severe errors are $R_{NS} = \Pr(Y = N, A = S)$ and $R_{PS} = \Pr(Y = P, A = S)$. We compare a budget on $R_{NS} + R_{PS}$ with separate budgets on (R_{NS}, R_{PS}). Action N is dominated by P, and the primary-risk improvement of S over P is $\Delta = 1 - U - 2V > 0$.

Corollary 4.1 (Two events, different information costs)**.** *For this same source and these same events,*

$$\mathcal{E}_M^{\mathrm{agg}} = \Theta(M^{-2}), \qquad \mathcal{E}_M^{\mathrm{split}} = \Theta(M^{-1}). \tag{6}$$

Both lower bounds allow arbitrary randomized M-message encoders; both upper bounds are attained in order by deterministic encoders.

These rates mean that preserving separate budgets to the same small frontier error requires more messages than preserving their sum. A shared budget uses the combined error probability relative to the improvement in primary risk; separate budgets also require distinguishing the two error types. The source, events and competing actions remain fixed. Appendix B.1 proves this specialization; the next theorem treats arbitrary groupings.

4.2 The Information Cost of General Risk Grouping

For k fixed severe error events, a partition specifies which errors share a budget. Assumption 3.3 gives a common default action with zero severe risk, a single risk-incurring action S with positive primary-risk improvement, and a full-dimensional posterior source with positive bounded density. Under these conditions, the number of groups determines the rate.

Theorem 4.2 (Information cost of error grouping)**.** *Under Assumption 3.3, for every partition $\mathcal{G}$ of $[k]$ into m nonempty groups, there exist $0 < c_{\mathcal{G}} \le C_{\mathcal{G}} < \infty$, independent of M, such that for every integer $M \ge 1$,*

$$c_{\mathcal{G}} M^{-2/m} \le \mathcal{E}_M^{\mathcal{G}} \le C_{\mathcal{G}} M^{-2/m}. \tag{7}$$

The lower bound holds for every measurable randomized encoder with at most M messages. A deterministic encoder achieves the upper bound. The same fixed source supports the conclusion for every partition.

The theorem quantifies the information needed to retain all budget choices. The relevant decision state is

$$a_{\mathcal{G}}(w) = \frac{Bw}{\Delta(w)}. \tag{8}$$

Each coordinate compares a group's conditional severe-error probability with the primary-risk improvement from choosing S. Under Assumption 3.3, these coordinates vary over an m-dimensional region and distinguish decisions across risk requirements. Grouping thus determines which distinctions the representation must retain. Appendix A establishes these properties and the matching full-frontier bounds.

In resource terms, keeping excess primary risk at most ε at every budget requires $\Theta(\varepsilon^{-m/2})$ messages as $\varepsilon \downarrow 0$, with k, K, f and the partition fixed. More groups therefore increase how rapidly the required representation size grows as the target error decreases.

4.3 When Finer Control Requires More Information

The theorem gives a direct comparison between coarser and finer control. A partition $\mathcal{H}$ refines $\mathcal{G}$ when each group in $\mathcal{H}$ is contained in a group of $\mathcal{G}$: some shared budgets are replaced by separate ones.

Corollary 4.3 (Refining a partition)**.** *Under Assumption 3.3, if $\mathcal{H}$ refines $\mathcal{G}$, then $D_{\mathcal{G}}(Q) \le D_{\mathcal{H}}(Q)$ for every fixed Q, and $\mathcal{E}_M^{\mathcal{G}} \le \mathcal{E}_M^{\mathcal{H}}$. If their group counts are $m < n$, respectively, then*

$$\frac{\mathcal{E}_M^{\mathcal{H}}}{\mathcal{E}_M^{\mathcal{G}}} \ge \frac{c_{\mathcal{H}}}{C_{\mathcal{G}}} M^{2/m - 2/n} \longrightarrow \infty. \tag{9}$$

Finer control imposes an additional cost even when each partition receives its own optimal encoder. If refinement increases the group count, the ratio of their optimal frontier errors grows without bound as M increases, although both errors tend to zero. This identifies a cost of the control requirements themselves. Appendix B.2 gives the detailed proof.

The role of source variation is visible in a proportional example. Consider instead a source with $W_i = c_i T$, where $c_i > 0$ and $\sum_i c_i = 1$ are fixed. Every policy then has $R_i = c_i R^{\mathrm{total}}$. All group budgets constrain the same scalar risk, so every partition has exactly the same fixed-encoder frontier error. Appendix B.3 gives the precise budget correspondence and a source for which every partition has rate $\Theta(M^{-2})$.

The action conditions play a complementary role. If the common zero-risk default is primary-risk optimal at every input, a constant message preserves every budget frontier. Together, these cases explain how finer grouping creates information demand through distinct risks and decisions that respond to them. Appendix B.4 also treats restricted families of weighted objectives, and Appendix B.5 gives an extension to bounded primary losses under the same severe-risk structure and explicit action conditions.

4.4 Choosing Risk Tradeoffs After Encoding

Once the representation is fixed, a deployment may specify a preference over risk tradeoffs and choose its budget accordingly. Fix a partition and suppress its superscript. For a continuous external cost $\Psi : [0,1]^m \to \mathbb{R}$, define

$$\begin{aligned} G_I(\alpha) &= F_I(\alpha) + \Psi(\alpha), \quad J_I(\Psi) = \min_{\alpha \in [0,1]^m} G_I(\alpha), \\ \mathcal{A}_I &= \arg\min_{\alpha \in [0,1]^m} G_I(\alpha). \end{aligned} \tag{10}$$

The frontier describes the available tradeoffs; Ψ determines which is preferred. Under Assumption 3.3, the frontiers are continuous and these minima exist.

Corollary 4.4 (Value of later risk preferences)**.** *If a fixed encoder has $D_{\mathcal{G}}(Q) \le \varepsilon$, then for every continuous Ψ,*

$$0 \le J_Z(\Psi) - J_W(\Psi) \le \varepsilon. \tag{11}$$

Every $\widehat{\alpha} \in \mathcal{A}_Z$ also satisfies $G_W(\widehat{\alpha}) - J_W \le \varepsilon$.

One representation therefore supports preferences supplied after encoding: the budget selected from its frontier remains near-optimal under full information. The representation from Theorem 4.2 gives $O(M^{-2/m})$ loss in this objective value for every such preference.

The selected budget is also close to the full-information optimal set when the objective has quadratic growth. If $G_W(\alpha) - J_W \ge (\kappa/2)\,\mathrm{dist}(\alpha, \mathcal{A}_W)^2$ for some $\kappa > 0$ throughout $[0,1]^m$, then every $\widehat{\alpha} \in \mathcal{A}_Z$ satisfies

$$\mathrm{dist}(\widehat{\alpha}, \mathcal{A}_W) \le \sqrt{2\varepsilon/\kappa}, \tag{12}$$

where dist denotes Euclidean distance to the optimal budget set. With κ fixed, the resulting distance bound is $O(M^{-1/m})$. Appendix C proves both guarantees and treats approximate budget optimization and worst-case loss over a family of risk costs. Section 5 examines these grouping and decision effects numerically.

5 Experiments

We examine how grouping the same severe errors changes representation cost, how an encoder can exploit the grouping, and what the resulting loss means for later decisions and evidence verification. In the synthetic experiments, frontier loss $D_{\mathcal{G}}(Q_M)$ measures the largest increase in optimal prediction error caused by compression, comparing compressed and full information under the same risk budgets. Each encoder uses at most M messages and remains fixed across budgets; smaller loss means better preservation of the frontier. We report numerical estimates for the constructed encoders, with protocols in Appendix E.

5.1 Grouping changes the rate; redundant risks do not

We fix three events $R_i = \Pr(Y = i, A = S)$ and a uniform three-dimensional posterior source satisfying Assumption 3.3. We compare the partitions $\{1,2,3\}$, $\{1\}\{2,3\}$ and $\{1\}\{2\}\{3\}$, using the state-grid construction from Theorem 4.2 for each partition. Figure 1(a) shows increasingly slow decay as control becomes finer. The fitted slopes are -2.00, -0.92 and -0.63, close to the respective -2, -1 and $-2/3$ rates of the constructive bound. Thus the separation appears while the source and all three severe events remain unchanged.

The difference is substantial at a common accuracy: a frontier loss of 10^{-3} corresponds to a worst-case increase of 0.1 percentage points in optimal prediction error. Interpolating these constructed-encoder curves at that target gives approximately 13, 180 and $8{,}700$ messages for one, two and three groups (Appendix E.1). Thus controlling the same events separately requires a much larger representation in this construction.

To isolate the role of independent risk directions, we also use a proportional source $W = cT$, with fixed c and scalar T. Every policy then has $R_i = c_i R^{\mathrm{total}}$. Figure 1(b) evaluates a single scalar encoder under both control rules: the errors coincide and retain the one-dimensional decay. This is the exact reduction in (45): separate budgets expose no additional decision coordinate when the risks are proportional.

5.2 Using the grouping to build the representation

We return to the full-dimensional uniform source to test whether a representation can exploit the lower-dimensional decision state created by coarse grouping. We compare two k-means encoders: one quantizes the grouping-specific state $a_{\mathcal{G}}(W)$, and the other quantizes the posterior W without using the grouping. The source, message budget and message-conditional risk decoding are the same.

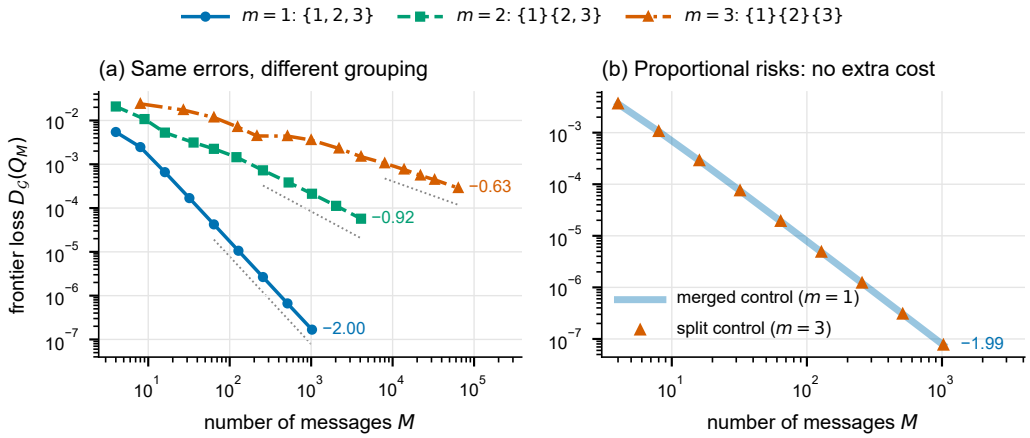

Figure 1: **Error grouping and finite-message frontier loss.** (a) The same uniform source and three severe events under one, two or three risk groups. Both axes are logarithmic; $M = n^m$ is the grid's message cap, including empty cells. Curves show estimated frontier loss. Endpoint annotations fit the five largest M; gray dotted segments indicate slopes $-2/m$, with arbitrary vertical offsets. (b) On the proportional-risk source, the same scalar encoder is evaluated under merged and split control. The blue line and orange triangles coincide, with fitted slope -1.99.

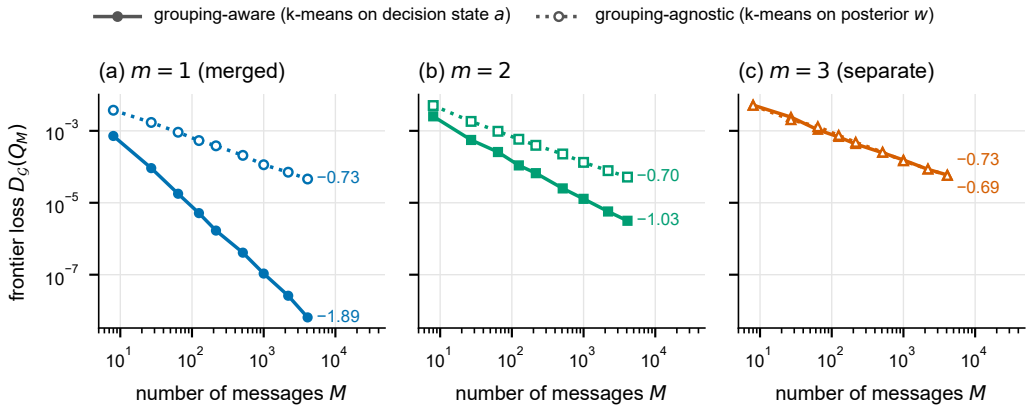

Figure 2: **Exploiting the deployment grouping.** (a–c) The same uniform source with one, two and three groups, using the partitions in Figure 1(a). Solid lines with filled markers use Δ-weighted k-means on the decision state; dotted lines with open markers use k-means on the posterior. Both use conditional-risk decoding. The axes show estimated frontier loss $D_{\mathcal{G}}(Q_M)$ versus M on logarithmic scales; annotations fit the five largest message counts. See Appendix E.1 for the common fitting and evaluation protocol.

Figure 2 shows a growing advantage for state-based quantization under coarse control. At $M = 4{,}096$, its frontier loss is approximately $7{,}000$ times smaller than this posterior baseline under merged control, and 16 times smaller with two groups. These ratios compare the additional prediction error caused by compression. With three separate groups, the curves nearly coincide: the decision state retains all three posterior dimensions. The comparison shows how encoding the grouping-specific state can realize the compression gains available under coarse control, while this posterior baseline has similar decay across groupings.

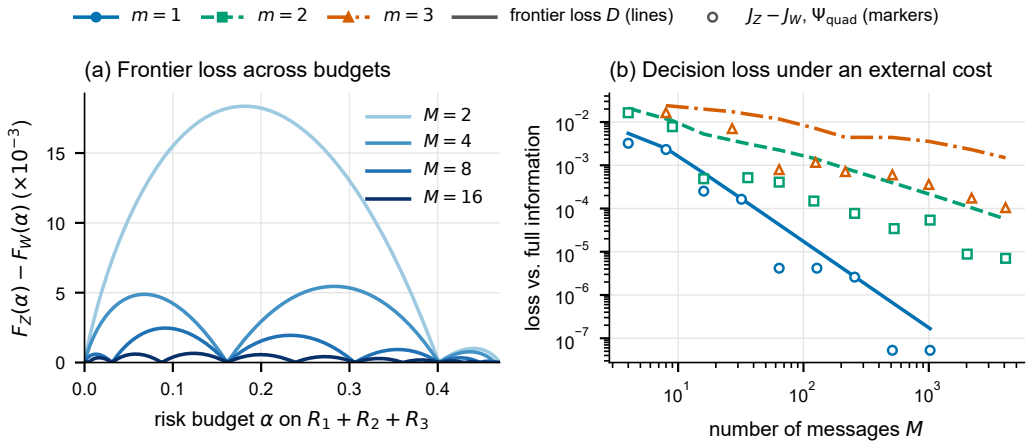

Figure 3: **From frontier approximation to risk selection.** (a) Merged-control gap $F_Z(\alpha) - F_W(\alpha)$ on the uniform source, evaluated on a budget grid from 0 to $\mathbb{E}\sum_i W_i = 0.47$; beyond this range both frontiers equal the unconstrained risk. The vertical scale is 10^{-3} and darker lines use more messages. (b) On logarithmic axes, lines show $D_{\mathcal{G}}(Q_M)$ and open markers show the optimal-value loss $J_Z - J_W$ for the quadratic external cost, with colors denoting the same partitions as Figure 1(a). Each encoder remains fixed during risk selection.

5.3 Frontier loss has a decision consequence

We next examine whether frontier accuracy protects the value of a risk tradeoff chosen after encoding. On the same uniform source, Figure 3(a) shows the merged frontier gap across budgets: increasing M from 2 to 16 lowers its maximum. To connect these gaps to a subsequent choice, we select a budget by minimizing $F_I(\alpha) + \Psi(\alpha)$, where Ψ is a quadratic cost on risk. For each grouping, this preference is held fixed as M varies. Figure 3(b) compares the resulting optimal-value loss $J_Z - J_W$ with $D_{\mathcal{G}}(Q_M)$. All evaluated losses lie below the frontier error, as in Corollary 4.4, and in some configurations the value loss approaches that error. At shared message counts, finer control incurs larger value loss in this comparison. The cost of finer control can therefore carry through to a decision made after encoding; uniform frontier accuracy bounds that cost. Budget-location error, which additionally uses curvature, is examined in Appendix E.3.

5.4 The grouping effect on evidence verification

We next use frozen calibrated outputs of a DeBERTa-based evidence verifier (He et al., 2023) on WiCE (Kamoi et al., 2023), with labels not supported (N), partially supported (P) and supported (S). Within each of two models, using full evidence or masked neighboring context, the outputs and overclaim events $N \to S$ and $P \to S$ are fixed. Only their control changes: a common penalty on both events, or two independently chosen penalties. Each control rule receives its own task-based codebook. On 949 development examples, we use the verifier's calibrated probabilities to evaluate the increase in expected penalized loss when decisions use fitted codewords rather than full outputs. We report the largest increase over the specified penalty range, denoted $\widehat{D}_{\mathrm{fit}}$.

Figure 4 shows larger errors under split control at every tested $M \geq 4$, for both models and all six codebook fits. On the original-fit curves, the split regret is about 3–6 times larger for $M \geq 8$. The difference also raises the message requirement: for the original full-model fit, split control needs 32 messages among the tested sizes to match the regret achieved by merged control with eight. These empirical posterior comparisons show the grouping effect even when each codebook is adapted to its control rule.

Appendix E gives additional source and partition comparisons, budget-location results, and a separate comparison of different joint posterior laws with matching marginals.

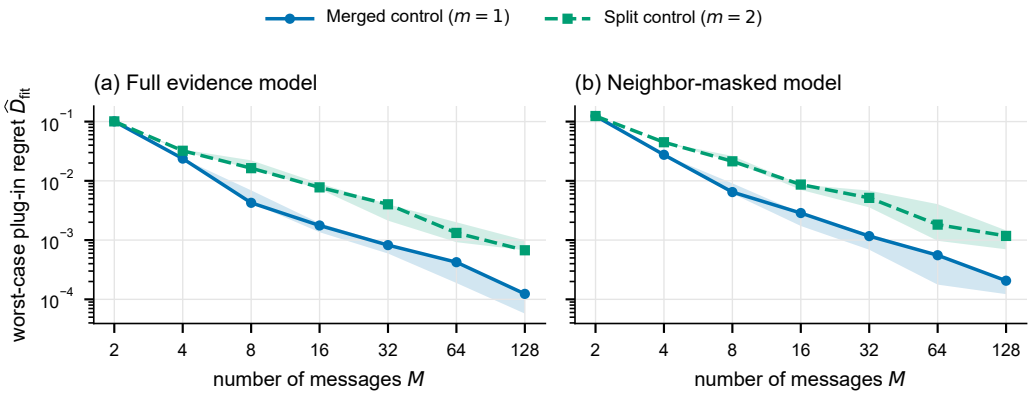

Figure 4: **Merged and split overclaim control on WiCE.** (a) Full evidence model; (b) neighbor-masked model. Each panel fixes the model outputs and compares separately fitted task-based codebooks. Both axes are logarithmic: the ordinate is the fitted decoder's estimated worst plug-in regret $\widehat{D}_{\text{fit}}$, maximized over penalties in $[0, 6]$ for merged control and $[0, 6]^2$ for split control. Decoding uses the fitted codewords. Lines use the original fit set; shading gives the minimum and maximum over that fit and five source-group bootstrap fits, with the verifier and development set held fixed. These bands describe codebook variation, not confidence intervals. See Appendix E.4.

6 Discussion and Conclusion

We have shown that the way severe errors are grouped can change the information required for risk control, even when the source and error events remain fixed. In the structured class studied here, preserving separate limits demands finer representation of the relation between risk and predictive benefit. The synthetic results illustrate this cost, and the evidence-verification study shows a corresponding separation over the tested penalty family. Together, these results make control granularity part of the representation problem.

The central issue is which distinctions must remain available after compression. Two inputs that admit similar decisions under a shared risk limit may call for different decisions when the errors are controlled separately. The decision state makes this dependence explicit by expressing each group's severe-error probability relative to the improvement from acting. The redundancy and invisible-boundary examples locate the source of complexity in risk variation that can affect a decision. This perspective connects the geometry of a source to the family of decisions it must support, identifying information that aggregate control permits an encoder to discard.

Preserving the frontier separates representation design from the choice of a particular risk tradeoff. Once the grouping is fixed, a deployment can specify numerical budgets or external risk costs later, with frontier error bounding the loss in optimal objective value. Representation quality therefore depends on the future decisions that compression leaves available. An encoder tailored to one operating point may discard distinctions that become essential when the preferred risk tradeoff changes. Choosing which errors warrant separate protection remains a deployment decision; our analysis quantifies its consequences for representation size. This suggests designing and evaluating compressed representations around the intended control requirements, alongside their performance at any one operating point.

The present analysis isolates this mechanism through a known posterior source and a common default competing with one risk-incurring action. With several such actions, different gains and risk profiles would create interacting decision boundaries. Determining which of these boundaries are visible to admissible risk preferences is a natural next step toward a broader theory of representation complexity. For learned representations, a complementary question is how posterior estimation and message construction jointly affect decision quality and budget feasibility. These directions extend the same organizing principle: the information a system must retain depends on the risk distinctions its future decisions must be able to make.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/f3549ef9b5ff520a7e41ff3cc306ab2b-Paper-Conference.pdf.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998. URL https://doi.org/10.1214/24-AOAS1998.

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6(58):1705–1749, 2005. URL https://jmlr.org/papers/v6/banerjee05b.html.

Basel Committee on Banking Supervision. Principles for the management of credit risk. Guidelines, Bank for International Settlements, April 2025. URL https://www.bis.org/publications/202504-guidelines-principles-management-credit-risk.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):43:1–43:34, 2021. doi: 10.1145/3478535. URL https://doi.org/10.1145/3478535.

Yahav Bechavod, Jiuyao Lu, and Aaron Roth. Omniprediction with long-term constraints. In *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 640–683. PMLR, 2026. URL https://proceedings.mlr.press/v336/bechavod26a.html.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953. doi: 10.1214/aoms/1177729032. URL https://doi.org/10.1214/aoms/1177729032.

CAST Investigators. Preliminary report: Effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *New England Journal of Medicine*, 321(6):406–412, 1989. doi: 10.1056/NEJM198908103210629. URL https://www.nejm.org/doi/abs/10.1056/NEJM198908103210629.

Shai Feldman, Liran Ringel, Stephen Bates, and Yaniv Romano. Achieving risk control in online learning settings. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=5Y04GWuoJu.

Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In *13th Innovations in Theoretical Computer Science Conference*, volume 215 of *Leibniz International Proceedings in Informatics*, pp. 79:1–79:21. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022. doi: 10.4230/LIPIcs.ITCS.2022.79. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ITCS.2022.79.

Parikshit Gopalan, Lunjia Hu, and Guy N. Rothblum. On computationally efficient multi-class calibration. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 1983–2026. PMLR, 2024. URL https://proceedings.mlr.press/v247/gopalan24a.html.

Siegfried Graf and Harald Luschgy. *Foundations of Quantization for Probability Distributions*, volume 1730 of *Lecture Notes in Mathematics*. Springer, 2000. doi: 10.1007/BFb0103945. URL https://link.springer.com/book/10.1007/BFb0103945.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2111.09543.

Lunjia Hu, Inbal Rachel Livni Navon, Omer Reingold, and Chutong Yang. Omnipredictors for constrained optimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13497–13527. PMLR, 2023. URL https://proceedings.mlr.press/v202/hu23b.html.

Sunay Joshi, Yan Sun, Hamed Hassani, and Edgar Dobriban. MultiRisk: Multiple risk control via iterative score thresholding, 2025. URL https://arxiv.org/abs/2512.24587. arXiv preprint arXiv:2512.24587.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. WiCE: Real-world entailment for claims in Wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7561–7583. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.470. URL https://aclanthology.org/2023.emnlp-main.470/.

Moritz Laurer. DeBERTa-v3-base-mnli-fever-anli. Hugging Face model card, n.d. URL https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli. Accessed 2026-09-26.

Lucien Le Cam. Sufficiency and approximate sufficiency. *The Annals of Mathematical Statistics*, 35 (4):1419–1455, 1964. doi: 10.1214/aoms/1177700372. URL https://doi.org/10.1214/aoms/1177700372.

Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28 (2):129–137, 1982. doi: 10.1109/TIT.1982.1056489.

Harikrishna Narasimhan. Learning with complex loss functions and constraints. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1646–1654. PMLR, 2018. URL https://proceedings.mlr.press/v84/narasimhan18a.html.

Harikrishna Narasimhan, Harish G. Ramaswamy, Shiv Kumar Tavker, Drona Khurana, Praneeth Netrapalli, and Shivani Agarwal. Consistent multiclass algorithms for complex metrics and constraints. *Journal of Machine Learning Research*, 25(367):1–81, 2024. URL https://jmlr.org/papers/v25/22-1137.html.

J. A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7 (4):308–313, 1965. doi: 10.1093/comjnl/7.4.308.

Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12(22):731–817, 2011. URL https://www.jmlr.org/papers/v12/reid11a.html.

Philippe Rigollet and Xin Tong. Neyman-Pearson classification, convexity and stochastic constraints. *Journal of Machine Learning Research*, 12:2831–2855, 2011. URL https://jmlr.org/papers/v12/rigollet11a.html.

SciPy community. scipy.stats.qmc.Sobol. SciPy 1.15.3 documentation, 2025. URL https://docs.scipy.org/doc/scipy-1.15.3/reference/generated/scipy.stats.qmc.Sobol.html. Accessed 2026-09-26.

Clayton Scott and Robert D. Nowak. A Neyman–Pearson approach to statistical learning. *IEEE Transactions on Information Theory*, 51(11):3806–3819, 2005. doi: 10.1109/TIT.2005.856955. URL https://doi.org/10.1109/TIT.2005.856955.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958. doi: 10.2140/pjm.1958.8.171. URL https://msp.org/pjm/1958/8-1/pjm-v8-n1-p14-s.pdf.

I. M. Sobol'. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics*, 7(4):86–112, 1967. doi: 10.1016/0041-5553(67)90144-9.

Helmut Strasser. Towards a statistical theory of optimal quantization. In *Data Analysis: Scientific Modeling and Practical Application*, pp. 369–383. Springer, 2000. doi: 10.1007/978-3-642-58250-9_30. URL https://www.math4tune.com/publications/Strasser%202000%2C%20quantization%20Bock%2C%20article.pdf.

Terence Tao. *An Introduction to Measure Theory*, volume 126 of *Graduate Studies in Mathematics*. American Mathematical Society, 2011. URL https://terrytao.wordpress.com/books/an-introduction-to-measure-theory/.

Ye Tian and Yang Feng. Neyman–Pearson multi-class classification via cost-sensitive learning. *Journal of the American Statistical Association*, 120(550):1164–1177, 2025. doi: 10.1080/01621459.2024.2402567. URL https://doi.org/10.1080/01621459.2024.2402567.

Xin Tong, Yang Feng, and Jingyi Jessica Li. Neyman-Pearson classification algorithms and NP receiver operating characteristics. *Science Advances*, 4(2):eaao1659, 2018. doi: 10.1126/sciadv.aao1659. URL https://doi.org/10.1126/sciadv.aao1659.

Brendan van Rooyen and Robert C. Williamson. Le Cam meets LeCun: Deficiency and generic feature learning, 2014. URL https://arxiv.org/abs/1402.4884. arXiv preprint arXiv:1402.4884.

Kush R. Varshney. Bayes risk error is a Bregman divergence. *IEEE Transactions on Signal Processing*, 59(9):4470–4472, 2011. doi: 10.1109/TSP.2011.2159500. URL https://krvarshney.github.io/pubs/Varshney_tsp2011.pdf.

Kush R. Varshney and Lav R. Varshney. Quantization of prior probabilities for hypothesis testing. *IEEE Transactions on Signal Processing*, 56(10):4553–4562, 2008. doi: 10.1109/TSP.2008.928164. URL https://arxiv.org/abs/0805.4338.

Lijia Wang, Y. X. Rachel Wang, Jingyi Jessica Li, and Xin Tong. Hierarchical Neyman-Pearson classification for prioritizing severe disease categories in COVID-19 patient data. *Journal of the American Statistical Association*, 119(545):39–51, 2024. doi: 10.1080/01621459.2023.2270657. URL https://doi.org/10.1080/01621459.2023.2270657.

Shengjia Zhao, Michael P. Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. Calibrating predictions to decisions: A novel approach to multi-class calibration. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22313–22324, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/bbc92a647199b832ec90d7cf57074e9e-Abstract.html.

AI USE STATEMENT

We used AI tools to assist with language polishing and grammar and spelling checks. They also helped identify potentially overlooked related work, check citation and reference formatting, implement code, and generate and visually refine figures. During manuscript preparation, AI tools assisted with drafting sections of the manuscript, focusing mainly on the presentation of our results and the formatting of figures, tables, and theorem environments. Further assistance included LaTeX typesetting and checking compliance with the conference template and formatting requirements. We have reviewed all content, including all AI-assisted work, and take responsibility for the final manuscript.

A Complete Proof of the Grouping Theorem

Throughout this appendix, fix a partition $\mathcal{G}$ into m groups and a source satisfying Assumption 3.3. We suppress $\mathcal{G}$ when unambiguous. All constants depend only on the fixed source and partition, unless specified otherwise. The argument applies to every measurable stochastic encoder $Q(z \mid w)$ of finite range.

We use $\|\cdot\|$ for the Euclidean norm and $|E|$ for Lebesgue volume. The source assumption requires neither continuity of f nor smoothness of ∂K.

A.1 Two competing actions and weighted message centers

For nonnegative risk weights $\lambda \in \mathbb{R}^m_+$, define the weighted value

$$V_I^{\mathcal{G}}(\lambda) = \inf_{\pi \text{ based on } I} \{R_0(\pi) + \lambda^\top R^{\mathcal{G}}(\pi)\}, \qquad I \in \{W, Z\}. \tag{13}$$

The weighted conditional risk of S is $r_S^\lambda(w) = r_S(w) + \lambda^\top B w$, whereas action 1 has $r_1^\lambda(w) = r_1(w)$. For the state $a_{\mathcal{G}}(w) = Bw/\Delta(w)$,

$$r_S^\lambda(w) - r_1^\lambda(w) = \lambda^\top B w - \Delta(w) = \Delta(w)\big(\lambda^\top a_{\mathcal{G}}(w) - 1\big). \tag{14}$$

Thus the rule chooses S when $\lambda^\top a_{\mathcal{G}}(w) < 1$ and chooses 1 otherwise, including equality. We write $a = a_{\mathcal{G}}$ when the partition is fixed. The next lemma verifies the reduction to these two actions at both information levels and identifies the state of each actual message.

Lemma A.1 (Action reduction and message posteriors)**.** *Under Assumption 3.3, at either information level $I = W$ or $I = Z$, restricting policies to actions 1 and S preserves every constrained and weighted optimum. For each message with positive probability, $w_z = \mathbb{E}[W \mid Z = z] \in K$, its improvement is $\Delta(w_z) = \mathbb{E}[\Delta(W) \mid Z = z] > 0$, and its decision state is*

$$a_z = \frac{B\mathbb{E}[W \mid Z = z]}{\mathbb{E}[\Delta(W) \mid Z = z]} = \frac{\mathbb{E}[\Delta(W) a_{\mathcal{G}}(W) \mid Z = z]}{\mathbb{E}[\Delta(W) \mid Z = z]}. \tag{15}$$

Proof. For a posterior w, the primary conditional risks are $r_i(w) = 1 - w_i$ and $r_S(w) = \sum_i w_i$. Every non-S action has zero group risk. Since $w_1 > w_i$ for $i > 1$, replacing any non-S action by 1 cannot increase primary risk or change any group risk. Thus all infima can be restricted to policies mixing 1 and S. Their primary-risk gap is $r_1 - r_S = \Delta > 0$.

The Markov condition and tower property imply $P(Y = i \mid Z) = \mathbb{E}[W_i \mid Z]$. Indeed, $Y - W - Z$ gives $\mathbb{E}[\mathbf{1}\{Y = i\} \mid W, Z] = W_i$; conditioning again on Z proves the claim. The same argument gives the posterior probability of label S. For every positive-probability message z, its posterior w_z belongs to K, because K is closed and convex. In particular the same dominance and positive improvement hold at every message. Since Δ is affine, $\Delta(w_z) = \mathbb{E}[\Delta(W) \mid z] > 0$, proving (15). Zero-probability messages may be deleted. For a concrete integral expression, put $p_z = \int_K Q(z \mid w) f(w)\, dw$. For $p_z > 0$,

$$w_z = \frac{\int_K w Q(z \mid w) f(w)\, dw}{p_z}, \qquad a_z = \frac{\int_K \Delta(w) a(w) Q(z \mid w) f(w)\, dw}{\int_K \Delta(w) Q(z \mid w) f(w)\, dw}.$$

Both are averages with nonnegative weights. The first belongs to K because averages of points in a closed convex set remain in that set. The second expression displays the additional Δ weighting directly. In particular, the risk difference at a message is $\lambda^\top B w_z - \Delta(w_z)$, so minimizing over $1, S$ uses exactly the state a_z from (15). This argument holds for any stochastic kernel Q. □

For any fixed encoder, a W-policy can simulate a Z-policy by sampling Z from $Q(\cdot \mid W)$ and applying the decoder. All risks are preserved, so $F_W^{\mathcal{G}}(\alpha) \le F_Z^{\mathcal{G}}(\alpha)$ at every budget.

If a single budget were known before designing the encoder, two messages would suffice to transmit the action sampled from any near-optimal policy using 1 and S. The decoder could simply output the transmitted action, preserving the policy's risks. The requirement in (4) is instead to fix one encoder for all subsequent budgets.

For weights $\lambda \geq 0$, conditional minimization gives

$$V_I(\lambda) = \mathbb{E}[1 - (W_I)_1] - \mathbb{E}\big[(\Delta(W_I) - \lambda^\top B W_I)_+\big], \qquad W_W = W, \quad W_Z = \mathbb{E}[W \mid Z]. \tag{16}$$

For completeness, writing $q_I = BW_I$, the conditional minimum is

$$\min\{1 - (W_I)_1,\ 1 - (W_I)_1 - \Delta(W_I) + \lambda^\top q_I\} = 1 - (W_I)_1 - (\Delta(W_I) - \lambda^\top q_I)_+.$$

The threshold rule is measurable, so integrating these pointwise minima attains the weighted infimum and proves (16). Weighted minimization supplies the dual values. The constrained decoder selects a feasible policy at the specified budget, as connected by the duality argument below.

A.2 Exact penalty and the frontier–value identity

Lemma A.2 (A common bounded dual domain)**.** *Under Assumption 3.3, let $b = \min_{w \in K, i \leq k} w_i > 0$ and $\Lambda = 2/b$. For every admissible Q, every partition, and $I \in \{W, Z\}$,*

$$F_I^{\mathcal{G}}(\alpha) = \max_{\lambda \in [0,\Lambda]^m} \{V_I^{\mathcal{G}}(\lambda) - \lambda^\top \alpha\}, \tag{17}$$

$$D_{\mathcal{G}}(Q) = \sup_{\lambda \in [0,\Lambda]^m} \{V_Z^{\mathcal{G}}(\lambda) - V_W^{\mathcal{G}}(\lambda)\}. \tag{18}$$

These identities hold for every $\alpha \in [0,1]^m$, including budgets with zero coordinates, and fix Q throughout.

Proof of Lemma A.2. Let $b = \min_{w \in K, i} w_i$ and $\Lambda = 2/b$. For a policy at information level I, let $s = \Pr(A = S)$ and let $r_g = R_g^{\mathcal{G}}(\pi)$. Each nonempty group has conditional probability at least b, at both information levels, so $r_g \geq bs$. Define

$$\theta = \min\left\{1,\ \min_{g: r_g > 0} \frac{\alpha_g}{r_g}\right\}, \tag{19}$$

with an empty inner minimum interpreted as $+\infty$. Multiply the policy's probability of S by θ at every input and transfer the remainder to action 1. The resulting risks are θr_g and are feasible. If there is a violation, a minimizing index g_* gives

$$(1 - \theta)s = \frac{r_{g_*} - \alpha_{g_*}}{r_{g_*}} s \leq b^{-1}(r_{g_*} - \alpha_{g_*})_+ \leq b^{-1} \sum_g (r_g - \alpha_g)_+. \tag{20}$$

Without a violation no mass is removed. The increase in primary risk is at most the removed mass, hence for every $\alpha \in [0,1]^m$,

$$F_I(\alpha) = \inf_\pi \left\{ R_0(\pi) + \Lambda \sum_g (R_g^{\mathcal{G}}(\pi) - \alpha_g)_+ \right\}. \tag{21}$$

More explicitly, if π^{rep} denotes the repaired policy, then

$$F_I(\alpha) \leq R_0(\pi^{\mathrm{rep}}) \leq R_0(\pi) + b^{-1} \sum_g (r_g - \alpha_g)_+ \leq R_0(\pi) + \Lambda \sum_g (r_g - \alpha_g)_+.$$

Taking the infimum over π proves one direction of (21). For the other, use feasible policies with primary risk at most $F_I(\alpha) + \epsilon$ and let $\epsilon \downarrow 0$. This also covers $s = 0$, vanishing coordinates of α, and optima defined only as infima.

The penalty is the maximum of $\lambda^\top(R^{\mathcal{G}}(\pi) - \alpha)$ over $[0,\Lambda]^m$. Write $\mathcal{P}_I$ for the measurable randomized policies and set $L(\lambda, \pi) = R_0(\pi) + \lambda^\top(R^{\mathcal{G}}(\pi) - \alpha)$. We use the compact-maximizing-side version of Sion (1958, Theorem 4.2, p. 175): a concave–convexlike function on a compact maximizing space and an arbitrary minimizing space has equal max–inf and inf–sup values when it is upper semicontinuous in the first variable. Here the maximizing space is $[0,\Lambda]^m$ and the minimizing space is $\mathcal{P}_I$. Convex combinations of multipliers remain in the box. For policies π_0, π_1, the pointwise mixture $\pi_t(\cdot \mid I) = (1-t)\pi_0(\cdot \mid I) + t\pi_1(\cdot \mid I)$ is measurable and has the corresponding mixture

of all risks. Consequently L is affine in both arguments, which verifies both concave–convexlike requirements. Its dependence on λ is continuous. The theorem therefore gives

$$\inf_{\pi\in\mathcal{P}_I}\max_{\lambda\in[0,\Lambda]^m} L(\lambda,\pi) = \sup_{\lambda\in[0,\Lambda]^m}\inf_{\pi\in\mathcal{P}_I} L(\lambda,\pi) = \sup_{\lambda\in[0,\Lambda]^m}\{V_I(\lambda)-\lambda^\top\alpha\}.$$

The exact penalty identifies the left side with $F_I(\alpha)$. The maximum is attained because V_I is Lipschitz in λ, as all group risks lie in $[0,1]$. In fact $|V_I(\lambda)-V_I(\lambda')|\le\|\lambda-\lambda'\|_1$: each policy's two weighted objectives differ by at most that amount, and taking infima preserves the bound. Thus the supremum is a maximum, proving (17) without a strict-feasibility requirement.

For every nonnegative λ one also has

$$V_I(\lambda) = \inf_{\alpha\in[0,1]^m}\{F_I(\alpha)+\lambda^\top\alpha\}. \tag{22}$$

For a fixed budget, a near-optimal feasible policy gives $V_I(\lambda)\le F_I(\alpha)+\lambda^\top\alpha$. Conversely, for any policy choose $\alpha = R^{\mathcal{G}}(\pi)$, which lies in the budget cube. Then $F_I(\alpha)+\lambda^\top\alpha$ is at most that policy's weighted objective. Taking infima proves the identity.

For a fixed Q, let $D = D_{\mathcal{G}}(Q)$. Equation (17) bounds each frontier difference by $\sup_{\lambda\in[0,\Lambda]^m}(V_Z - V_W)$. Conversely, $F_Z\le F_W + D$ and (22) imply $V_Z(\lambda)\le V_W(\lambda)+D$. This proves (18), without interchanging $\inf_Q$ with either supremum. It also gives

$$|F_I(\alpha)-F_I(\beta)|\le\Lambda\|\alpha-\beta\|_1. \tag{23}$$

As a maximum of affine functions of α, F_I is convex and coordinatewise nonincreasing. □

A.3 Weighted pushforward density and its dimension

Definition A.3 (Pushforward of a finite measure)**.** For a finite measure ξ and a measurable map T, its pushforward $T_\#\xi$ is the measure defined by $(T_\#\xi)(E) = \xi(T^{-1}(E))$ for every measurable set E. Equivalently, for a nonnegative measurable function g,

$$\int g(x)\,(T_\#\xi)(dx) = \int g(T(w))\,\xi(dw).$$

This is the usual change-of-variables identity for image measures (Tao, 2011, Section 1.4).

Here define $\xi_\Delta(dw) = \Delta(w)f(w)\,dw$ on K and $\mathcal{A}_{\mathcal{G}} = a_{\mathcal{G}}(K)$. Its weighted image measure is

$$\mu_\Delta(E) = \mathbb{E}[\Delta(W)\mathbf{1}\{a_{\mathcal{G}}(W)\in E\}] = (a_{\mathcal{G}})_\#\xi_\Delta(E). \tag{24}$$

Consequently,

$$\int g(a)\,\mu_\Delta(da) = \int_K g(a_{\mathcal{G}}(w))\Delta(w)f(w)\,dw.$$

The total mass is $\mathbb{E}\Delta$; no normalization to a probability measure is made. The next lemma computes its density and dimension.

Lemma A.4 (Density of the derived decision state)**.** *Under Assumption 3.3, $\mathcal{A}_{\mathcal{G}} = a_{\mathcal{G}}(K)$ is a compact convex set with nonempty m-dimensional interior and positive coordinate bounds. The finite measure μ_Δ in* (24) *has a Lebesgue density h bounded above on $\mathbb{R}^m$ and bounded below by a positive constant on a cube of positive volume. Every actual message center belongs to $\mathcal{A}_{\mathcal{G}}$.*

Proof. Set $c = (2,1,\dots,1)^\top$, so $\Delta(w) = 1-c^\top w$. The fractional linear transformation

$$x = T(w) = \frac{w}{1-c^\top w}, \qquad T^{-1}(x) = \frac{x}{1+c^\top x} \tag{25}$$

is one-to-one on the positive-denominator domain. Differentiation gives

$$DT(w) = \Delta^{-1}I + \Delta^{-2}wc^\top, \qquad \det DT(w) = \Delta^{-(k+1)}. \tag{26}$$

To check the determinant, use the elementary rank-one identity $\det(I+uv^\top) = 1+v^\top u$:

$$\det DT(w) = \Delta^{-k}\det(I+\Delta^{-1}wc^\top) = \Delta^{-k}\left(1+\frac{c^\top w}{\Delta}\right) = \Delta^{-(k+1)}.$$

The identity itself follows by multilinearity in the columns: terms containing two columns proportional to u vanish, leaving the identity term and the sum $\sum_i v_i u_i$. On K, both transformations have bounded derivatives and nonvanishing Jacobians. The image $T(K)$ is compact and has nonempty k-dimensional interior. It is also convex: for $w_t = (1-t)w_0 + tw_1$,

$$T(w_t) = \frac{(1-t)\Delta(w_0)}{\Delta(w_t)} T(w_0) + \frac{t\Delta(w_1)}{\Delta(w_t)} T(w_1), \tag{27}$$

The coefficients sum to one. The second coefficient is $t\Delta(w_1)/[(1-t)\Delta(w_0) + t\Delta(w_1)]$, a continuous strictly increasing function from 0 to 1. Hence every point on the segment between $T(w_0)$ and $T(w_1)$ occurs, proving convexity. The explicit smooth inverse in (25) maps an open neighborhood of each interior point back into the source domain; thus $T(K)$ has nonempty interior as well.

Push the measure $\Delta(w)f(w)\,dw$ forward by T. Its density is

$$\widetilde{f}(x) = f(T^{-1}x)\Delta(T^{-1}x)^{k+2}\mathbf{1}\{x \in T(K)\}. \tag{28}$$

Indeed, $|\det DT^{-1}(x)| = \Delta(T^{-1}x)^{k+1}$, and the extra factor Δ comes from the definition of the weighted measure. With $\Delta_- := \min_K \Delta > 0$ and $\Delta_+ := \max_K \Delta < \infty$, one may bound this density on $T(K)$ by $f_-\Delta_-^{k+2}$ and $f_+\Delta_+^{k+2}$. It has finite positive upper and lower bounds on $T(K)$. The state is $a = Bx$, and B has rank m because its rows correspond to nonempty disjoint groups. Extend these rows to an invertible matrix J and write $Jx = (a, v)$ with $v \in \mathbb{R}^{k-m}$. The joint density in these coordinates is $\widetilde{f}(J^{-1}(a,v))/|\det J|$. Integrating out v yields a density h of μ_Δ with

$$0 \le h(a) \le H < \infty \quad \text{a.e.}, \qquad h(a) \ge h_0 > 0 \quad \text{a.e. on a cube } C \text{ with } |C| > 0. \tag{29}$$

Here is the integration explicitly. For $m < k$, choose a bounded box $D_v \subset \mathbb{R}^{k-m}$ containing all fiber coordinates of $JT(K)$. Writing $\widetilde{f} = 0$ outside $T(K)$, Tonelli's theorem gives

$$h(a) = \frac{1}{|\det J|} \int_{\mathbb{R}^{k-m}} \widetilde{f}(J^{-1}(a,v))\,dv \le \frac{f_+\Delta_+^{k+2}|D_v|}{|\det J|} =: H.$$

Because $JT(K)$ has nonempty interior, it contains $C \times D_0$, where C is an m-dimensional cube and D_0 a $(k-m)$-dimensional box, both with positive volume. For almost every $a \in C$,

$$h(a) \ge \frac{1}{|\det J|} \int_{D_0} f_-\Delta_-^{k+2}\,dv = \frac{f_-\Delta_-^{k+2}|D_0|}{|\det J|} =: h_0 > 0.$$

Null sets remain null under the smooth coordinate changes, and Fubini's theorem gives the stated almost-everywhere fiber bounds. When $m = k$, $J = B$ is already invertible and the transformed density has the same upper and lower bounds with no fiber integral.

Consequently $\mathcal{A}_\mathcal{G} = BT(K)$ is compact, convex, and full dimensional in $\mathbb{R}^m$. Fix bounds

$$0 < \underline{a} \le a_g \le \overline{a} < \infty \quad (a \in \mathcal{A}_\mathcal{G},\ g \le m). \tag{30}$$

We may take $\underline{a} = b$, since $q_g \ge b$ and $\Delta \le 1$. All actual message centers are in $\mathcal{A}_\mathcal{G}$; alternatively this follows directly from their positive-weight mean representation. This weighted measure has total mass $\mathbb{E}\Delta$. □

A.4 Exact boundary integral and global distinguishability

For states $a, a' \in \mathcal{A}_\mathcal{G}$ and weights $\lambda \ge 0$, define

$$e_\lambda(a, a') = (1 - \lambda^\top a)_+ - (1 - \lambda^\top a)\mathbf{1}\{\lambda^\top a' < 1\}. \tag{31}$$

This is the normalized regret of using the action selected at a' when the true state is a. Multiplication by the true improvement Δ recovers the conditional weighted-risk regret.

Definition A.5 (Global quadratic lower bound)**.** A nonnegative distortion d on $\mathcal{A}_\mathcal{G} \times \mathcal{A}_\mathcal{G}$ has a global quadratic lower bound if a constant $c > 0$, independent of the pair, satisfies $d(a, a') \ge c\|a - a'\|^2$ for every $a, a' \in \mathcal{A}_\mathcal{G}$.

We establish this property for the averaged decision regret $\overline{e}(a, a') = \int e_\lambda(a, a')\,\nu(d\lambda)$ with one fixed weight distribution ν. Its relation to the Bregman form of Bayes risk error (Varshney, 2011) is given after the proof.

Lemma A.6 (Visible boundaries give global quadratic control)**.** *Let $\mathcal{A}_\mathcal{G}$ have coordinate bounds $0 < \underline{a} \le a_g \le \overline{a} < \infty$, and take $\Lambda > \max\{1, 1/\underline{a}\}$. There exist a probability measure ν supported on $[0,\Lambda]^m$ and $c_e > 0$, depending only on $m, \underline{a}, \overline{a}, \Lambda$, such that*

$$\int e_\lambda(a,a')\,\nu(d\lambda) \ge c_e\|a-a'\|^2 \qquad (a,a' \in \mathcal{A}_\mathcal{G}).$$

For any admissible encoder, its weighted-value gap is the expectation in (32)*.*

Proof. The quantity e_λ in (31) is nonnegative. It equals the absolute normalized risk gap if the action selected at a' is suboptimal at a, and zero otherwise. For a real message center, the selected action is Bayes-optimal for that message. Affinity of the conditional risks and the tower property give

$$V_Z(\lambda) - V_W(\lambda) = \mathbb{E}[\Delta(W) e_\lambda(a(W), a_Z)]. \tag{32}$$

To expand this step, put $b_Z = \mathbf{1}\{\lambda^\top a_Z < 1\}$ and use $\Delta_Z = \mathbb{E}[\Delta \mid Z]$. Then

$$\begin{aligned}\mathbb{E}[\Delta(1-\lambda^\top a)b_Z] &= \mathbb{E}[(\Delta_Z - \lambda^\top B\mathbb{E}[W \mid Z])b_Z] \\ &= \mathbb{E}[(\Delta_Z - \lambda^\top B\mathbb{E}[W \mid Z])_+].\end{aligned}$$

Also $\Delta(1-\lambda^\top a)_+ = (\Delta - \lambda^\top BW)_+$. Subtracting the two expressions in (16), whose default terms agree by $\mathbb{E}[\mathbb{E}[W_1 \mid Z]] = \mathbb{E}W_1$, gives (32). The fixed tie rule selecting 1 matches the strict inequality in (31). Other message-level Bayes tie choices have the same expected weighted value.

Choose $0 < \tau \le 1$ with $(m-1)\tau\overline{a} \le 1/2$, and let $\mathcal{B} = [0,\Lambda] \times [0,\tau]^{m-1}$. For $m = 1$, omit t and all empty vectors below. Write

$$s = 1/a_1, \qquad r = (a_2,\dots,a_m)/a_1, \qquad \beta(a,t) = s - t^\top r. \tag{33}$$

For every $a \in \mathcal{A}_\mathcal{G}$ and t in the small box,

$$\frac{1}{2\overline{a}} \le \beta(a,t) \le \frac{1}{\underline{a}} < \Lambda. \tag{34}$$

The same is true of a'. The two actions disagree only between $\beta(a,t)$ and $\beta(a',t)$ along the ℓ coordinate. There the true normalized gap is $a_1|\ell - \beta(a,t)|$. Integration gives the exact identity

$$\int_0^\Lambda e_{(\ell,t)}(a,a')\,d\ell = \frac{a_1}{2}\big(\beta(a,t) - \beta(a',t)\big)^2. \tag{35}$$

Write $\beta = \beta(a,t)$ and $\beta' = \beta(a',t)$. For example, if $\beta' \ge \beta$, the surrogate selects S while the true optimum selects 1 on $[\beta,\beta')$, and

$$\int_0^\Lambda e_{(\ell,t)}(a,a')\,d\ell = a_1\int_\beta^{\beta'}(\ell-\beta)\,d\ell = \frac{a_1}{2}(\beta'-\beta)^2.$$

If $\beta' < \beta$, the corresponding integral is $a_1\int_{\beta'}^{\beta}(\beta-\ell)\,d\ell$, with the same result. When the thresholds agree it is zero. Endpoint tie choices affect a set of one-dimensional measure zero.

Let (s',r') be the coordinates of a', and put $\delta s = s - s'$ and $\delta r = r - r'$. Under uniform $t \in [0,\tau]^{m-1}$, independent coordinates give

$$\mathbb{E}_t[(\delta s - t^\top \delta r)^2] = \left(\delta s - \frac{\tau}{2}\sum_j \delta r_j\right)^2 + \frac{\tau^2}{12}\|\delta r\|^2 \ge \zeta(\delta s^2 + \|\delta r\|^2), \tag{36}$$

where for $m > 1$ one may choose

$$\zeta = \min\left\{\frac{1}{2}, \frac{\tau^2}{12[1+(m-1)\tau^2/2]}\right\} > 0. \tag{37}$$

To verify this, let $u = \delta s - (\tau/2)\sum_j \delta r_j$ and use $\delta s^2 + \|\delta r\|^2 \le 2u^2 + [1+(m-1)\tau^2/2]\|\delta r\|^2$. The variance term in (36) uses $\mathbb{E}t_j = \tau/2$, $\mathrm{Var}(t_j) = \tau^2/12$, and zero cross-covariances. The

choice (37) makes the coefficients of u^2 and $\|\delta r\|^2$ large enough for the displayed lower bound. For $m = 1$, take $\zeta = 1$.

The inverse map is $(s, r) \mapsto (1/s, r/s)$. Here $s, s' \geq 1/\overline{a}$ and $\|r'\| \leq R_* := \sqrt{m-1}\,\overline{a}/\underline{a}$. Consequently

$$|s^{-1} - (s')^{-1}| \leq \overline{a}^2|s - s'|, \qquad \left\|\frac{r}{s} - \frac{r'}{s'}\right\| \leq \overline{a}\|r - r'\| + R_*\overline{a}^2|s - s'|.$$

Squaring and using $(x + y)^2 \leq 2x^2 + 2y^2$ gives

$$\|a - a'\|^2 \leq \overline{a}^4(1 + 2R_*^2)|s - s'|^2 + 2\overline{a}^2\|r - r'\|^2 \leq L_{\mathrm{inv}}^2(|s - s'|^2 + \|r - r'\|^2),$$

where $L_{\mathrm{inv}}^2 = \max\{\overline{a}^4(1 + 2R_*^2), 2\overline{a}^2\}$. For $m = 1$, the first scalar inequality suffices with $L_{\mathrm{inv}} = \overline{a}^2$. Let ν be the uniform probability measure on $\mathcal{B}$. Dividing (35) by Λ and averaging in t gives

$$\overline{e}(a, a') := \int e_\lambda(a, a')\,\nu(d\lambda) \geq \frac{\underline{a}\zeta}{2\Lambda L_{\mathrm{inv}}^2}\|a - a'\|^2 =: c_e\|a - a'\|^2. \tag{38}$$

This is a global inequality for all pairs in $\mathcal{A}_\mathcal{G}$. Equal states have identical action families by (8); distinct states have strictly positive averaged regret and hence disagree on a positive-measure set of weights. Thus $a_\mathcal{G}$ identifies the action family up to null sets of weights. The improvement Δ retains the regret scale in (32). □

Bregman form of the averaged regret. For the uniform measure ν just constructed, define $\phi(a) = \int(1 - \lambda^\top a)_+\,\nu(d\lambda)$. For every positive a, the hyperplane $\lambda^\top a = 1$ has zero ν-measure. Difference quotients are bounded by $\|\lambda\|$, which is bounded on the support of ν, so differentiation under the integral gives

$$\nabla\phi(a') = -\int \lambda\,\mathbf{1}\{\lambda^\top a' < 1\}\,\nu(d\lambda).$$

Since $(1 - \lambda^\top a')_+ = (1 - \lambda^\top a')\mathbf{1}\{\lambda^\top a' < 1\}$, substitution yields

$$\overline{e}(a, a') = \phi(a) - \phi(a') - \nabla\phi(a')^\top(a - a').$$

This is the Bregman divergence generated by ϕ (Banerjee et al., 2005, Definition 1). Lemma A.6 supplies its positive quadratic lower bound on $\mathcal{A}_\mathcal{G}$ directly. In particular it also verifies strict convexity there: applying the bound with second argument $a_t = (1 - t)a + ta'$ to both endpoints and taking weights $1 - t, t$ gives

$$(1 - t)\phi(a) + t\phi(a') - \phi(a_t) \geq c_e t(1 - t)\|a - a'\|^2 \qquad (0 < t < 1).$$

A.5 Lower bound for arbitrary randomized encoders

The finite-M bound below has the exponent of classical squared-error quantization (Graf & Luschgy, 2000, Theorem 6.2, with $r = 2$).

Lemma A.7 (Finite-codebook volume bound)**.** *Let a finite measure on $\mathbb{R}^m$ have density $h \geq h_0 > 0$ almost everywhere on a cube C of volume $v > 0$. For every integer $M \geq 1$ and any collection of at most M centers $a_z \in \mathbb{R}^m$,*

$$\int \min_z \|a - a_z\|^2 h(a)\,da \geq \frac{h_0 v}{2}\left(\frac{v}{2\omega_m}\right)^{2/m} M^{-2/m},$$

where ω_m is the volume of the Euclidean unit ball.

Proof. Set $\rho = (v/(2M\omega_m))^{1/m}$. The union of the M radius-ρ balls has Lebesgue measure at most $M\omega_m\rho^m = v/2$. Its complement inside C therefore has volume at least $v/2$. On that complement every squared distance is at least ρ^2. Integrating there gives $h_0(v/2)\rho^2$, which is the claimed expression. Centers outside C and overlapping balls satisfy the same union bound. □

Lower bound in Theorem 4.2. Fix any Q with at most M positive-probability messages. By (18), nonnegativity, Tonelli's theorem, and (38),

$$D_{\mathcal{G}}(Q) \geq \int \big(V_Z(\lambda) - V_W(\lambda)\big)\, \nu(d\lambda) \geq c_e \mathbb{E}[\Delta \|a - a_Z\|^2]$$
$$\geq c_e \int \min_z \|a - a_z\|^2 \, \mu_\Delta(da). \tag{39}$$

For the first inequality, a supremum is at least the average under any probability measure supported on its domain. Next, insert (32) and exchange expectation with integration in λ. The integrand is measurable and nonnegative, so Tonelli's theorem permits this exchange. Lemma A.6 then applies separately to each realized pair $(a(W), a_Z)$. The last inequality holds pointwise after conditioning on W: the stochastic kernel averages distances to its fixed collection of message centers, and this average is at least the smallest distance. The argument applies even when Q depends on all of W.

Explicitly, for almost every w,

$$\sum_z Q(z \mid w)\Delta(w)\|a(w) - a_z\|^2 \geq \Delta(w) \min_z \|a(w) - a_z\|^2.$$

Integrating over w produces the last term of (39). The centers may depend on Q; Lemma A.7 is uniform over all choices of at most M centers.

Taking $v = |C|$ from Lemma A.4, the volume lemma yields

$$\int_C \min_z \|a - a_z\|^2 h(a)\, da \geq \frac{h_0 v}{2} \left(\frac{v}{2\omega_m}\right)^{2/m} M^{-2/m}. \tag{40}$$

Combining this with (39) and taking the infimum over Q proves the lower bound in Theorem 4.2 for all $M \geq 1$. In particular, a valid constant is $c_{\mathcal{G}} = c_e(h_0 v/2)(v/(2\omega_m))^{2/m}$. □

A.6 Deterministic construction and uniform upper bound

Lemma A.8 (Uniform regret of the state-grid encoder)**.** *Suppose μ_Δ has density at most H and $\mathcal{A}_{\mathcal{G}}$ lies in a cube of side L_A. For $M \geq 1$, set $n = \lfloor M^{1/m} \rfloor$. The deterministic encoder using n^m equal grid cells has, with its actual message centers and $h_M = \sqrt{m} L_A/n$,*

$$\sup_{\lambda \in [0,\Lambda]^m} (V_Z(\lambda) - V_W(\lambda)) \leq 2Hm\Lambda L_A^{m-1} h_M^2.$$

Proof. Enclose $\mathcal{A}_{\mathcal{G}}$ in a cube of side length $L_A > 0$ and divide it into $n^m \leq M$ equal cells with $n = \lfloor M^{1/m} \rfloor$. The encoder sends the index of the cell containing $a(W)$; use a fixed assignment on cell boundaries. The true center a_z is a positive-weight mean of points from that cell, and therefore belongs to its closure. Consequently,

$$\|a - a_Z\| \leq h_M := \sqrt{m} L_A/n. \tag{41}$$

For any nonzero $\lambda \in [0, \Lambda]^m$, an action mismatch implies

$$|1 - \lambda^\top a| \leq |\lambda^\top (a - a_Z)| \leq \|\lambda\| h_M, \qquad e_\lambda(a, a_Z) \leq \|\lambda\| h_M. \tag{42}$$

Choose j with $|\lambda_j| \geq \|\lambda\|/\sqrt{m}$. Holding the other coordinates fixed, the slab $|1 - \lambda^\top a| \leq \|\lambda\| h_M$ has length at most $2\sqrt{m} h_M$ along coordinate j. Indeed, if $v = 1 - \sum_{i \neq j} \lambda_i a_i$, its j-th slice is contained in $[(v - \|\lambda\| h_M)/\lambda_j, (v + \|\lambda\| h_M)/\lambda_j]$, of length $2\|\lambda\| h_M/\lambda_j \leq 2\sqrt{m} h_M$. Integrating these lengths over the other $m-1$ coordinates of the enclosing cube bounds its volume by $2\sqrt{m} h_M L_A^{m-1}$. Using the density upper bound and the enclosing cube gives

$$\mu_\Delta\{a : |1 - \lambda^\top a| \leq \|\lambda\| h_M\} \leq 2H\sqrt{m} L_A^{m-1} h_M. \tag{43}$$

Thus, by (32), uniformly in the entire weight box,

$$V_Z(\lambda) - V_W(\lambda) \leq 2Hm\Lambda L_A^{m-1} h_M^2. \tag{44}$$

At $\lambda = 0$ the difference is zero because both information levels select S. The slice argument remains uniform as λ approaches zero: its norm cancels in the ratio defining the slice length. For large h_M, the mass bound may exceed total mass but remains valid. Finally $\lfloor M^{1/m} \rfloor \geq M^{1/m}/2$, so $h_M^2 \leq 4mL_A^2 M^{-2/m}$. □

Taking $n = \lfloor M^{1/m} \rfloor$ uses at most M messages, and a valid constant in the upper bound is $C_{\mathcal{G}} = 8Hm^2\Lambda L_A^{m+1}$. Equation (18) gives the desired upper bound, completing the proof of Theorem 4.2.

For a budget decoder, use the constrained optimum based on the message posteriors, allowing randomization to meet the budget. With finitely many messages and actions this is a finite linear program over a nonempty compact polytope, so its optimum exists. The weighted values above certify its performance through duality. Both stages use population information. More explicitly, if $p_z = \Pr(Z = z)$, $q_z = Bw_z$ and $t_z \in [0, 1]$ is the decoder's probability of S, the program is

$$\min_{0 \le t_z \le 1} \sum_z p_z\{1 - (w_z)_1 - \Delta(w_z)t_z\} \quad \text{subject to} \quad \sum_z p_z(q_z)_g t_z \le \alpha_g \quad (g \le m).$$

The zero vector is feasible for every budget. Compactness of the finite box and closedness of the constraints give attainment. Thus the construction specifies an actual feasible budget decoder as well as its weighted-value certificate.

B Mechanism Boundaries and a Bounded-Loss Extension

B.1 Checking the two-event specialization

Proof of Corollary 4.1. For Corollary 4.1, set $W_1 = V$, $W_2 = U$ and use

$$K = [0.29, 0.31] \times [0.19, 0.21].$$

Every point has $W_1 - W_2 \ge 0.08$, positive coordinates, and $0.17 \le \Delta = 1 - 2W_1 - W_2 \le 0.23$. The rectangle is compact, convex and has nonempty two-dimensional interior. Swapping the coordinates of the original source preserves its positive density bounds. Thus all conditions of Assumption 3.3 hold, with $b = 0.19$.

For the aggregate partition, $B = (1, 1)$ and the state is $a_{\text{agg}} = (U + V)/(1 - U - 2V)$. For the split partition, ordering the coordinates by events NS, PS gives $a_{\text{split}} = (U, V)/(1 - U - 2V)$. If its coordinates are (x, y), the inverse is

$$U = \frac{x}{1 + x + 2y}, \qquad V = \frac{y}{1 + x + 2y}.$$

Both are the states derived by (8). Lemma A.4 gives a one-dimensional weighted density for the first partition and a two-dimensional weighted density for the second, with the bounds used in Lemmas A.7 and A.8. Theorem 4.2 therefore gives the rates in (6) on the same source. At both information levels, P dominates N, leaving P and S as the competing actions. Both partitions concern the same two severe events. □

B.2 Refinement at a fixed encoder

Proof of Corollary 4.3. Suppose $\mathcal{H}$ refines $\mathcal{G}$, with n and m groups respectively. Let $C \in \{0, 1\}^{m \times n}$ record which fine groups form each coarse group, so $R^{\mathcal{G}}(\pi) = CR^{\mathcal{H}}(\pi)$ for every policy. For each coarse weight $\lambda \in [0, \Lambda]^m$, set $\widetilde{\lambda} = C^\top \lambda$. Each fine group belongs to exactly one coarse group, so every coordinate of $\widetilde{\lambda}$ is a coordinate of λ and belongs to $[0, \Lambda]$. For either information level and the same fixed encoder,

$$V_I^{\mathcal{G}}(\lambda) = \inf_\pi\{R_0(\pi) + \lambda^\top CR^{\mathcal{H}}(\pi)\} = V_I^{\mathcal{H}}(C^\top \lambda).$$

The multiplier bound $\Lambda = 2/b$ in Lemma A.2 is common to the two partitions. Consequently,

$$\begin{aligned} D_{\mathcal{G}}(Q) &= \sup_{\lambda \in [0,\Lambda]^m} \{V_Z^{\mathcal{H}}(C^\top \lambda) - V_W^{\mathcal{H}}(C^\top \lambda)\} \\ &\le \sup_{\widetilde{\lambda} \in [0,\Lambda]^n} \{V_Z^{\mathcal{H}}(\widetilde{\lambda}) - V_W^{\mathcal{H}}(\widetilde{\lambda})\} = D_{\mathcal{H}}(Q). \end{aligned}$$

Taking infima over the same class of M-message encoders gives $\mathcal{E}_M^{\mathcal{G}} \le \mathcal{E}_M^{\mathcal{H}}$. If $n > m$, Theorem 4.2 gives $\mathcal{E}_M^{\mathcal{H}} \ge c_{\mathcal{H}} M^{-2/n}$ and $0 < \mathcal{E}_M^{\mathcal{G}} \le C_{\mathcal{G}} M^{-2/m}$. Dividing yields (9); its exponent $2/m - 2/n$ is positive. □

B.3 EXACT REDUNDANCY AMONG DISTINCT EVENTS

Proposition B.1 (Proportional risks preserve every frontier error)**.** *If $W_i = c_i T$ with fixed $c_i > 0$ and $\sum_i c_i = 1$, then, for every admissible encoder and $I \in \{W, Z\}$, the split frontier satisfies*

$$F_I^{\mathrm{split}}(\alpha) = F_I^{\mathrm{total}}\left(\min\left\{1, \min_i \frac{\alpha_i}{c_i}\right\}\right). \tag{45}$$

Hence $D_{\mathrm{split}}(Q) = D_{\mathrm{total}}(Q)$. The same equality of frontier errors holds between any partition and total risk.

Proof. Suppose the posterior source instead has the form $W_i = c_i T$, where $c_i > 0$ and $\sum_i c_i = 1$. For any full-information or encoded policy,

$$R_i = \mathbb{E}[c_i T \pi_S(I)] = c_i R^{\mathrm{total}}, \qquad R^{\mathrm{total}} = \mathbb{E}[T \pi_S(I)]. \tag{46}$$

For encoded policies this expression is evaluated under the joint law of (T, Z); no deterministic-encoder assumption is needed. The split constraints are equivalent to $R^{\mathrm{total}} \le \min_i \alpha_i / c_i$. Since $R^{\mathrm{total}} \le 1$, truncating that bound at 1 has no effect, proving (45). Conversely, every $\gamma \in [0, 1]$ is represented by $\alpha_i = c_i \gamma$, so for every fixed Q,

$$D_{\mathrm{split}}(Q) = D_{\mathrm{total}}(Q). \tag{47}$$

The same argument works for every partition, replacing c_i by its group sum. Thus all partitions have exactly the same optimal representation error. The equality follows at the level of feasible policy sets: for each vector budget, its scalar effective budget imposes precisely the same restriction on $\mathbb{E}[T\pi_S(I)]$. Both information levels use this identical reparameterization, and the map from vector to effective scalar budget is surjective. Taking the supremum of their frontier difference thus gives equality, not merely a one-sided comparison. Taking the infimum over the same class of encoders preserves that equality. □

For a concrete example, take $k = 3$, $c = (1/2, 1/3, 1/6)$ and let T have a fixed positive bounded density on $[2/5, 1/2]$. Here $\Delta = 1 - 3T/2 \in [1/4, 2/5]$. The scalar state for total risk is $a = T/(1 - 3T/2)$, whose derivative is $(1 - 3T/2)^{-2}$. Its image is $[1, 2]$ and its inverse is $T = a/(1 + 3a/2)$. Writing f_T for the density of T, the weighted state density is exactly

$$h(a) = \frac{f_T(a/(1 + 3a/2))}{(1 + 3a/2)^3} \mathbf{1}\{1 \le a \le 2\}. \tag{48}$$

This follows by multiplying f_T by $\Delta = 1/(1 + 3a/2)$ and the inverse Jacobian $(1 + 3a/2)^{-2}$. Thus h has positive lower and finite upper bounds on $[1, 2]$.

The common posterior lower bound is $b = 1/15$, so the repair argument gives the bounded dual identity with $\Lambda = 30$. On this interval, for all states a, a', the exact scalar boundary integral is

$$\frac{1}{\Lambda} \int_0^{\Lambda} e_\lambda(a, a')\, d\lambda = \frac{a}{2\Lambda} \left(\frac{1}{a} - \frac{1}{a'}\right)^2 \ge \frac{1}{32\Lambda} |a - a'|^2.$$

The thresholds lie in $[1/2, 1] \subset (0, \Lambda)$, and the inequality uses $a \ge 1$ and $aa' \le 4$. Applying the randomized-center argument and Lemma A.7 with $m = 1$ gives $\Omega(M^{-2})$ for total risk. Conversely, divide $[1, 2]$ into M equal intervals. The true weighted center remains in its interval, so the scalar mismatch slab has length at most $2/M$ and its regret is at most Λ/M. The bounded density gives an $O(M^{-2})$ weighted gap uniformly in λ, and duality transfers it to all scalar budgets. Proposition B.1 transfers both bounds to every partition. Hence every partition has error $\Theta(M^{-2})$. This one-dimensional source illustrates the role of source dimension in Theorem 4.2.

B.4 INVISIBLE DECISIONS

Suppose the task family is restricted to a weight set $\mathcal{L}$ such that $\lambda^\top B w < \Delta(w)$ for every $w \in K$ and $\lambda \in \mathcal{L}$. Then S minimizes conditional weighted risk everywhere for every task. A constant message followed by S has zero weighted-value loss on $\mathcal{L}$. The full-budget objective instead uses the complete multiplier box in Lemma A.2.

Alternatively, suppose action 1 still dominates the other non-S actions but $r_1(w) < r_S(w)$ everywhere. Action 1 then minimizes the primary risk and incurs zero severe risk. Consequently $F_I^{\mathcal{G}}(\alpha) = \mathbb{E}[r_1(W)]$ for every budget and both information levels; one message suffices. The strict positive improvement in Assumption 3.3 excludes this case. To see equality at both information levels, every policy has conditional primary risk at least $r_1(W)$ before averaging; a policy based on a message is still such a randomized policy when viewed from W. The constant policy choosing 1 attains this lower bound and is feasible at every budget. In the restricted-weight example, the same argument applies to each weighted risk with the constant policy choosing S.

B.5 Changing the primary loss while retaining the risk structure

Retain the source of Assumption 3.3 and the severe events $\Pr(Y = i, A = S)$, but replace the primary loss by a matrix $L_0(y, a) \in [0, L]$. Let $r_a(w) = \mathbb{E}[L_0(Y, a) \mid W = w]$. Assume that action 1 strictly dominates every other non-S action on K and that

$$\Delta_L(w) = r_1(w) - r_S(w) = d_0 + d^\top w > 0 \quad (w \in K), \qquad d_0 \neq 0. \tag{49}$$

The new frontier is defined using $R_0 = \mathbb{E}[L_0(Y, A)]$.

Proposition B.2 (Bounded primary-loss extension)**.** *Under these conditions, for every partition into m groups the optimal M-message frontier error is $\Theta(M^{-2/m})$, with a lower bound for arbitrary randomized encoders and an upper bound from a deterministic construction.*

Proof. The dominance reduction is unchanged. Repairing a policy by moving mass from S to 1 increases primary risk by at most L times that mass. Thus take

$$\Lambda > \max\{1, L/b, \Delta_{L,\max}/b\}, \qquad \Delta_{L,\max} = \max_K \Delta_L, \tag{50}$$

to obtain the same bounded dual and frontier–value identities. The state is $a = Bw/\Delta_L(w)$ and the message center uses Δ_L in both numerator and denominator of (15). Indeed, the difference of weighted conditional risks is

$$r_S(w) + \lambda^\top Bw - r_1(w) = \Delta_L(w)(\lambda^\top a(w) - 1).$$

The affine risks and convexity of K preserve the dominance assumptions after conditioning on a message. Since Δ_L is affine, the actual message state is $\mathbb{E}[\Delta_L a \mid Z]/\mathbb{E}[\Delta_L \mid Z]$.

For $T_L(w) = w/\Delta_L(w)$, direct differentiation gives

$$T_L^{-1}(x) = \frac{d_0 x}{1 - d^\top x}, \qquad \det DT_L(w) = \frac{d_0}{\Delta_L(w)^{k+1}} \neq 0. \tag{51}$$

Here $DT_L = \Delta_L^{-1} I - \Delta_L^{-2} w d^\top$, and the rank-one determinant identity yields

$$\det DT_L = \Delta_L^{-k}\left(1 - \frac{d^\top w}{\Delta_L}\right) = \frac{d_0}{\Delta_L^{k+1}}.$$

Solving $x = w/(d_0 + d^\top w)$ for w gives the displayed inverse; its denominator is nonzero because $1 - d^\top x = d_0/\Delta_L(w)$. The inverse denominator has the same sign as d_0 on the image; both signs of d_0 are allowed. Positivity and affinity of Δ_L give the same line-segment reparameterization as (27). The weighted pushforward density is now

$$f(T_L^{-1}x)\frac{\Delta_L(T_L^{-1}x)^{k+2}}{|d_0|}\mathbf{1}\{x \in T_L(K)\}. \tag{52}$$

It has positive finite bounds on the compact image. The rank-m projection therefore again has a globally bounded density and a positive lower bound on an interior cube.

Use $\underline{a} = b/\Delta_{L,\max}$ and the corresponding finite upper bound $\overline{a}$. Equation (50) ensures $1/\underline{a} < \Lambda$. Choose $\tau \leq 1$ with $(m-1)\tau\overline{a} \leq 1/2$. Because $\Lambda > 1$, the small box lies inside $[0, \Lambda]^m$, and because $\Lambda > 1/\underline{a}$, both thresholds are strictly internal. Lemma A.6 therefore applies to the unchanged normalized regret e_λ. Multiplying by Δ_L gives the true regret identity, and Lemma A.7 gives the lower bound using the new density's interior cube. For the upper bound, the Δ_L-weighted center stays in its grid cell. The same slice-length calculation in Lemma A.8 uses only the density upper bound, the cube side length and the multiplier cap. It gives $O(M^{-2/m})$ uniformly in λ. Finally, the repaired bounded dual identity transfers this to all budgets. Each constant is fixed by K, f, L_0 and the partition, independently of M. □

The nonzero intercept ensures invertibility; $d_0 = 0$ makes the Jacobian degenerate. The extension preserves the common zero-risk default and the unique risk-incurring action S, while permitting changes to the primary loss and its improvement scale.

B.6 Information interfaces and the scope of the bounds

For a more general input X, let $W_i = P(Y = i \mid X)$ and suppose its law satisfies Assumption 3.3. Set $W_S = 1 - \sum_{i=1}^{k} W_i$. Every finite encoding with $Y - X - Z$ induces a kernel $Q(z \mid w) = P(Z = z \mid W = w)$ and obeys $Y - W - Z$. Indeed, conditional on X, the label and the encoder's private randomness are independent, and the label probabilities depend on X through W. Thus $P(Y = i, Z = z \mid W) = W_i P(Z = z \mid W)$. In detail, conditioning first on X and then on W gives

$$\mathbb{E}[\mathbf{1}\{Y = i\}\mathbf{1}\{Z = z\} \mid W] = \mathbb{E}[W_i P(Z = z \mid X) \mid W] = W_i \mathbb{E}[P(Z = z \mid X) \mid W].$$

Moreover any policy based on X has a policy based on W, obtained by conditional averaging, with exactly the same risks. Therefore the full-information frontier is unchanged and the lower bound applies. For an X-policy, define $\bar{\pi}_a(W) = \mathbb{E}[\pi_a(X) \mid W]$. These conditional expectations form a probability vector. For each label i and action a,

$$\Pr(Y = i, A = a) = \mathbb{E}[W_i \pi_a(X)] = \mathbb{E}[W_i \bar{\pi}_a(W)].$$

Thus the entire joint label–action distribution, and hence every risk used here, is preserved. In the reverse direction, a W-policy is an X-policy by composition. This proves equality of the full-information frontiers in both directions.

For the upper bound, a representation H from which $a_G(W)$ can be recovered measurably can support the deterministic grid encoder. The budget decoder also needs the population message probabilities and conditional risks. These population quantities suffice to implement the budget decoder and attain the constructive bound.

C Risk Selection: Values, Policies, and Optimal Sets

Under Assumption 3.3, fix a partition, an encoder Q, and the budget cube $\mathcal{D} = [0, 1]^m$. Let $0 \le F_Z - F_W \le \varepsilon$ throughout $\mathcal{D}$. By (23), both frontiers are continuous. For a continuous cost Ψ, $G_I = F_I + \Psi$ attains its minimum on the compact domain, and $\mathcal{A}_I$ is nonempty and compact.

C.1 Value and budget guarantees

Definition C.1 (Growth relative to the optimal set)**.** An objective G with minimum J and nonempty optimal set $\mathcal{A}$ has p-order growth with constant $c > 0$ on $\mathcal{D}$ if $G(\alpha) - J \ge c \operatorname{dist}(\alpha, \mathcal{A})^p$ for every $\alpha \in \mathcal{D}$. Here $\operatorname{dist}(\alpha, \mathcal{A}) = \inf_{\beta \in \mathcal{A}} \|\alpha - \beta\|$. Quadratic growth uses $p = 2$ and the convention $c = \kappa/2$.

Proposition C.2 (Approximate budget optimization)**.** *Under the conditions of Corollary 4.4, if $G_Z(\widehat{\alpha}) \le J_Z + \eta$ for $\eta \ge 0$, then $G_W(\widehat{\alpha}) - J_W \le \varepsilon + \eta$. If G_W satisfies the quadratic growth condition stated before* (12)*, then*

$$\operatorname{dist}(\widehat{\alpha}, \mathcal{A}_W) \le \sqrt{2(\varepsilon + \eta)/\kappa}. \tag{53}$$

Proof of Corollary 4.4 and Proposition C.2. Pointwise, $G_W \le G_Z \le G_W + \varepsilon$. For any $\alpha_W \in \mathcal{A}_W$, this gives

$$J_W = \min_{\alpha \in \mathcal{D}} G_W(\alpha) \le \min_{\alpha \in \mathcal{D}} G_Z(\alpha) = J_Z \le G_Z(\alpha_W) \le G_W(\alpha_W) + \varepsilon = J_W + \varepsilon.$$

This proves (11). If $G_Z(\widehat{\alpha}) \le J_Z + \eta$, then

$$J_W \le G_W(\widehat{\alpha}) \le G_Z(\widehat{\alpha}) \le J_Z + \eta \le J_W + \varepsilon + \eta. \tag{54}$$

Under $G_W(\alpha) - J_W \ge (\kappa/2) \operatorname{dist}(\alpha, \mathcal{A}_W)^2$, this proves (53). Setting $\eta = 0$ gives the objective guarantee in Corollary 4.4 and the location bound in (12). More generally, $G_W(\alpha) - J_W \ge c \operatorname{dist}(\alpha, \mathcal{A}_W)^p$ with $c, p > 0$ gives distance at most $((\varepsilon + \eta)/c)^{1/p}$. □

We use the following convention: g is κ-strongly convex on $\mathcal{D}$, with $\kappa > 0$, if for $x, y \in \mathcal{D}$ and $t \in [0,1]$,

$$g((1-t)x + ty) \le (1-t)g(x) + tg(y) - \frac{\kappa}{2}t(1-t)\|x-y\|^2.$$

Thus strong convexity of Ψ, together with convexity of F_W from Appendix A.2, makes G_W strongly convex with the same constant. If two distinct points minimized G_W, the inequality with $t = 1/2$ would put their midpoint strictly below the minimum. Hence the optimum α_* is unique. For $t \in (0,1)$,

$$\begin{aligned} G_W(\alpha_*) &\le G_W((1-t)\alpha_* + t\alpha) \\ &\le (1-t)G_W(\alpha_*) + tG_W(\alpha) - \frac{\kappa}{2}t(1-t)\|\alpha - \alpha_*\|^2. \end{aligned} \tag{55}$$

Subtracting $(1-t)G_W(\alpha_*)$ and dividing by $t > 0$ gives

$$G_W(\alpha) - J_W \ge \frac{\kappa}{2}(1-t)\|\alpha - \alpha_*\|^2.$$

Letting $t \downarrow 0$ proves the growth bound, including boundary optima. For example, $\Psi(\alpha) = u^\top \alpha + (\kappa/2)\|\alpha\|^2$ with $u \ge 0$ is nondecreasing and strongly convex on $\mathcal{D}$.

For a fixed problem there is also qualitative convergence to the optimal set without a specified growth rate. For $r > 0$, define

$$\gamma(r) = \inf_{\alpha \in \mathcal{D}:\operatorname{dist}(\alpha, \mathcal{A}_W) \ge r} (G_W(\alpha) - J_W), \tag{56}$$

with value $+\infty$ if the set is empty. Otherwise the set is compact, and continuity gives a point α_r at which the infimum is attained. If $\gamma(r) = 0$, then $G_W(\alpha_r) = J_W$, so $\alpha_r \in \mathcal{A}_W$, contradicting $\operatorname{dist}(\alpha_r, \mathcal{A}_W) \ge r > 0$. Therefore $\gamma(r) > 0$. Equation (54) then gives $\operatorname{dist}(\widehat{\alpha}, \mathcal{A}_W) < r$ whenever $\varepsilon + \eta < \gamma(r)$. Thus $\varepsilon + \eta \to 0$ implies convergence to $\mathcal{A}_W$ for a fixed source and a fixed cost. The modulus quantifies the additional sensitivity relevant when the preference itself varies.

C.2 Costs of actual risks and the preference-family identity

Proposition C.3 (Budget cost equals optimal actual-risk cost)**.** *For continuous coordinatewise nondecreasing Ψ on $\mathcal{D} = [0,1]^m$,*

$$J_I(\Psi) = \inf_{\pi \text{ based on } I} \{R_0(\pi) + \Psi(R^{\mathcal{G}}(\pi))\}. \tag{57}$$

Proof. Write H_I for the infimum on the right. Every policy has $R^{\mathcal{G}}(\pi) \in \mathcal{D}$, and choosing this vector as the budget gives

$$J_I \le F_I(R^{\mathcal{G}}(\pi)) + \Psi(R^{\mathcal{G}}(\pi)) \le R_0(\pi) + \Psi(R^{\mathcal{G}}(\pi)).$$

Taking the infimum over π proves $J_I \le H_I$. Conversely, fix $\alpha \in \mathcal{D}$ and $\epsilon > 0$. A feasible near-minimizer π_ϵ satisfies $R_0(\pi_\epsilon) \le F_I(\alpha) + \epsilon$. Monotonicity gives

$$H_I \le R_0(\pi_\epsilon) + \Psi(R^{\mathcal{G}}(\pi_\epsilon)) \le F_I(\alpha) + \Psi(\alpha) + \epsilon.$$

Letting $\epsilon \downarrow 0$ and taking the infimum over α proves $H_I \le J_I$. □

In particular, if a selected budget satisfies the preceding η tolerance and a policy $\widehat{\pi}$ based on Z satisfies $R^{\mathcal{G}}(\widehat{\pi}) \le \widehat{\alpha}$ and $R_0(\widehat{\pi}) \le F_Z(\widehat{\alpha}) + \eta_p$, then

$$0 \le R_0(\widehat{\pi}) + \Psi(R^{\mathcal{G}}(\widehat{\pi})) - J_W \le \varepsilon + \eta + \eta_p. \tag{58}$$

For the upper bound, the complete chain is

$$R_0(\widehat{\pi}) + \Psi(R^{\mathcal{G}}(\widehat{\pi})) \le F_Z(\widehat{\alpha}) + \eta_p + \Psi(\widehat{\alpha}) = G_Z(\widehat{\alpha}) + \eta_p \le J_W + \varepsilon + \eta + \eta_p.$$

For the lower bound, a W-policy simulates $\widehat{\pi}$ by first sampling Z using Q. Its actual-risk cost is therefore at least the infimum in (57) for $I = W$. This evaluates the actual policy risks even when its chosen budget has slack.

Proposition C.4 (Worst-case value loss over risk preferences)**.** *Let $\mathcal{C}$ be any class of continuous coordinatewise nondecreasing costs on $\mathcal{D}$ containing $\Psi_\lambda(\alpha) = \lambda^\top \alpha$ for every $\lambda \in [0, \Lambda]^m$, with Λ as in Lemma A.2. For every fixed encoder,*

$$\sup_{\Psi \in \mathcal{C}} \{J_Z(\Psi) - J_W(\Psi)\} = D_{\mathcal{G}}(Q). \tag{59}$$

Proof. The value bound gives $\sup_{\Psi \in \mathcal{C}}(J_Z - J_W) \le D_{\mathcal{G}}(Q)$. For $\Psi_\lambda(\alpha) = \lambda^\top \alpha$, (22) gives $J_I(\Psi_\lambda) = V_I(\lambda)$. Equation (18) yields the reverse inequality and hence the identity. In full, for the same fixed encoder throughout,

$$D_{\mathcal{G}}(Q) \;\ge\; \sup_{\Psi \in \mathcal{C}}(J_Z(\Psi) - J_W(\Psi)) \;\ge\; \sup_{\lambda \in [0,\Lambda]^m} (V_Z(\lambda) - V_W(\lambda)) \;=\; D_{\mathcal{G}}(Q).$$

Taking the infimum over the same encoder class gives

$$\inf_{Q:\, |\mathcal{Z}| \le M} \sup_{\Psi \in \mathcal{C}} \{J_Z(\Psi) - J_W(\Psi)\} = \mathcal{E}_M^{\mathcal{G}} = \Theta(M^{-2/m}). \tag{60}$$

The cost witnessing the worst-case loss may depend on Q. □

Individual costs can have smaller losses. For example, when $\Psi = 0$, always choosing S is primary-risk optimal on K, and a constant representation attains the same optimal value as full information.

C.3 An exact example separating value and location

The following two-atom example isolates the distinction between value and location. It uses labels and actions N, P, S and total severe risk $\Pr(Y \ne S, A = S)$. Let $X \in \{1, 2\}$ be equiprobable, with posteriors

$$p_1 = (1/5, 3/10, 1/2), \qquad p_2 = (19/100, 29/100, 13/25). \tag{61}$$

At both points P dominates N. Writing $q_i = 1 - p_{i,S}$, $\Delta_i = p_{i,S} - p_{i,P}$ and $T_i = \Delta_i / q_i$ gives

$$(q_1, \Delta_1, T_1) = (1/2, 1/5, 2/5), \qquad (q_2, \Delta_2, T_2) = (12/25, 23/100, 23/48). \tag{62}$$

Starting from action P everywhere, use S first at point 2, whose improvement per unit risk is larger, and then at point 1. Randomization at a point fills each segment. Explicitly, let $t_i \in [0, 1]$ be the probability of selecting S at atom i. The two risks are

$$R_0 = \frac{141}{200} - \frac{t_1}{10} - \frac{23 t_2}{200}, \qquad R = \frac{t_1}{4} + \frac{6 t_2}{25}.$$

For $0 \le \alpha \le 6/25$, set $t_1 = 0$ and $t_2 = 25\alpha/6$. For $6/25 \le \alpha \le 49/100$, set $t_2 = 1$ and $t_1 = 4(\alpha - 6/25)$. Substitution in the second case gives

$$R_0 = \frac{59}{100} - \frac{2}{5}\left(\alpha - \frac{6}{25}\right) = \frac{343}{500} - \frac{2}{5}\alpha.$$

For $\alpha \ge 49/100$, both probabilities equal one, giving

$$F_X(\alpha) = \begin{cases} 141/200 - (23/48)\alpha, & 0 \le \alpha \le 6/25, \\ 343/500 - (2/5)\alpha, & 6/25 \le \alpha \le 49/100, \\ 49/100, & 49/100 \le \alpha \le 1. \end{cases} \tag{63}$$

The threshold ordering proves optimality: exchanging any selected lower-ratio mass for available higher-ratio mass increases improvement at fixed risk. Indeed, transferring an amount h of risk from atom 1 to atom 2 decreases primary risk by $h(T_2 - T_1) > 0$. Thus an optimum never uses $t_1 > 0$ while $t_2 < 1$, which gives exactly the three cases above.

For $0 < \delta < 1/2$, pass X through a binary channel with crossover probability δ. Its equiprobable message posteriors are $(1 - \delta)p_1 + \delta p_2$ and $\delta p_1 + (1 - \delta)p_2$. Their severe probabilities and improvements are

$$q_{Z=1} = \frac{1}{2} - \frac{\delta}{50}, \quad \Delta_{Z=1} = \frac{1}{5} + \frac{3\delta}{100}, \qquad q_{Z=2} = \frac{12}{25} + \frac{\delta}{50}, \quad \Delta_{Z=2} = \frac{23}{100} - \frac{3\delta}{100}.$$

Uniformly over budgets,

$$0 \le F_Z(\alpha) - F_X(\alpha) \le (37/12)\delta. \tag{64}$$

To prove this, apply any full-information optimal policy at message z as though the input were z. Couple the policy randomization so that actions agree when $Z = X$. The primary and severe risks each change by at most δ. If the resulting severe risk exceeds the budget, scale down all S probabilities. Since every message has $q_Z \ge 12/25$, this removes at most $(25/12)\delta$ action mass. More explicitly, write s for its total S-mass and r for its severe risk. Then $r \ge (12/25)s$ and $r - \alpha \le \delta$. Scaling by α/r removes

$$(1 - \alpha/r)s = \frac{r-\alpha}{r}s \le \frac{25}{12}(r-\alpha) \le \frac{25}{12}\delta.$$

The total increase is at most $\delta + (25/12)\delta$, proving the bound.

With the fixed cost $\Psi(\alpha) = (2/5)\alpha$, the full-information optimal set is $[6/25, 49/100]$ and $J_X = 343/500$. At weight $2/5$, the weighted S-minus-P gap is zero at p_1 and $-19/500$ at p_2. At the two messages these gaps are respectively $-19\delta/500$ and $-19(1-\delta)/500$. Both are strictly negative, so the compressed optimum selects S everywhere. Since the cost is strictly increasing, its unique optimal budget is $49/100$, and $J_Z = J_X$. It lies in the full optimal set but is distance $1/4$ from its other endpoint. The natural location target is therefore the full optimal set.

For a varying preference, take $\Psi_\delta(\alpha) = \lambda_\delta \alpha$ with $\lambda_\delta = 2/5 + \delta/100$. Now the full-information optimal budget is uniquely $6/25$. The smaller message threshold satisfies

$$T_{Z=1} - 2/5 = \frac{(19/500)\delta}{1/2 - \delta/50} > \delta/100, \tag{65}$$

and the other threshold is larger. Thus the compressed optimal budget is still uniquely $49/100$. The ordering can be checked without approximation:

$$T_{Z=2} - T_{Z=1} = \frac{\Delta_{Z=2} q_{Z=1} - \Delta_{Z=1} q_{Z=2}}{q_{Z=1} q_{Z=2}} = \frac{(19/1000)(1-2\delta)}{q_{Z=1} q_{Z=2}} > 0.$$

Here $T_1 < \lambda_\delta < T_2$, so the full-information policy selects S only at atom 2. Its primary risk is $59/100$ and its severe risk is $6/25$. The compressed policy selects S at both messages, with both risks equal to $49/100$. Therefore

$$J_Z(\Psi_\delta) = \frac{49}{100} + \frac{49}{100}\lambda_\delta, \qquad J_X(\Psi_\delta) = \frac{59}{100} + \frac{6}{25}\lambda_\delta,$$
$$J_Z(\Psi_\delta) - J_X(\Psi_\delta) = -\frac{1}{10} + \frac{1}{4}\left(\frac{2}{5} + \frac{\delta}{100}\right) = \frac{\delta}{400}.$$

Together with $49/100 - 6/25 = 1/4$, this gives

$$|\alpha_{Z,\delta} - \alpha_{X,\delta}| = 1/4, \qquad J_Z(\Psi_\delta) - J_X(\Psi_\delta) = \delta/400. \tag{66}$$

The vanishing value loss coexists with a fixed location difference as the preferences approach indifference. This gives a concrete role to the growth modulus in location guarantees. The example is atomic and is used for selection sensitivity; the rate theorem concerns the continuous sources of Assumption 3.3.

D The Averaged-Regret Connection to Prior Quantization

Bayes-risk quantization supplies a useful interpretation of the averaged regret in our lower bound. We give an explicit mapping to the prior quantization framework of Varshney & Varshney (2008) and Varshney (2011).

D.1 The testing experiment and its regret

Fix a probability measure ν on nonnegative weights with finite positive means $u_j = \mathbb{E}_\nu[\lambda_j]$, $j \le m$. For $a = a_{\mathcal{G}}(w)$, set

$$s(a) = 1 + \sum_{j=1}^{m} u_j a_j, \qquad \rho_0(a) = \frac{1}{s(a)}, \qquad \rho_j(a) = \frac{u_j a_j}{s(a)}. \tag{67}$$

The vector $\rho(a)$ is a prior on $m+1$ hypotheses. In a fixed testing experiment with observation λ, take likelihood densities relative to ν to be $h_0(\lambda) = 1$ and $h_j(\lambda) = \lambda_j/u_j$. They each integrate to one. There are two decisions: selecting S incurs unit loss on hypotheses $1, \dots, m$ and zero on hypothesis 0; selecting the default incurs the complementary loss. Multiplying the two unnormalized conditional risks by $s(a)$ gives $\lambda^\top a$ and 1, respectively. Hence the testing rule using the surrogate prior $\rho(a')$ agrees with the original rule using state a'. In detail, $\rho_j \geq 0$ and $\sum_{j=0}^m \rho_j = (1 + \sum_j u_j a_j)/s(a) = 1$. The likelihood normalizations are

$$\int h_0 \, d\nu = 1, \qquad \int h_j \, d\nu = \frac{\mathbb{E}_\nu \lambda_j}{u_j} = 1.$$

Write $b(\lambda) \in \{0, 1\}$ for a testing decision, where 1 selects S. At prior ρ, its Bayes risk is

$$\mathcal{R}(\rho, b) = \int \left\{ \rho_0 h_0(\lambda)(1 - b(\lambda)) + \sum_{j=1}^m \rho_j h_j(\lambda) b(\lambda) \right\} \nu(d\lambda).$$

For the prior in (67), this becomes

$$\mathcal{R}(\rho(a), b) = \frac{1}{s(a)} \int \{1 + (\lambda^\top a - 1) b(\lambda)\} \, \nu(d\lambda).$$

The Bayes decision using a surrogate a' is $b_{a'}(\lambda) = \mathbf{1}\{\lambda^\top a' < 1\}$, with the same default tie rule as in Appendix A.1.

Definition D.1 (Bayes risk error in the testing experiment)**.** For a prior ρ' and its Bayes decision $b_{\rho'}$, set $d_{\mathrm{test}}(\rho, \rho') = \mathcal{R}(\rho, b_{\rho'}) - \inf_b \mathcal{R}(\rho, b)$. Thus the rule is selected at ρ' and its loss is evaluated at ρ, with the observation integrated out (Varshney & Varshney, 2008; Varshney, 2011).

For every pair of states,

$$\int e_\lambda(a, a') \, \nu(d\lambda) = s(a) d_{\mathrm{test}}(\rho(a), \rho(a')). \tag{68}$$

Indeed, pointwise minimization gives

$$\inf_b \mathcal{R}(\rho(a), b) = \frac{1}{s(a)} \int \min\{1, \lambda^\top a\} \, \nu(d\lambda).$$

Subtract this expression from $\mathcal{R}(\rho(a), b_{a'})$ and use $\min\{1, x\} = 1 - (1 - x)_+$. The integrand after multiplication by $s(a)$ is

$$(1 - \lambda^\top a)_+ - (1 - \lambda^\top a)\mathbf{1}\{\lambda^\top a' < 1\} = e_\lambda(a, a'),$$

which proves (68).

D.2 Source tilting and conditional-mean representatives

Tilt the original source by the positive factor $\Delta(W)s(a(W))$:

$$d\widetilde{P}(W) = \frac{\Delta(W)s(a(W))}{C} \, dP(W), \qquad C = \mathbb{E}[\Delta(W)s(a(W))]. \tag{69}$$

Compactness bounds Δ and a, and the u_j are finite and positive, so $0 < C < \infty$. The tilted joint law is

$$\widetilde{P}(dw, z) = C^{-1} \Delta(w) s(a(w)) Q(z \mid w) P(dw).$$

In particular it keeps the same stochastic encoder and message alphabet. For a positive-probability message z and an integrable function g,

$$\mathbb{E}_{\widetilde{P}}[g(W) \mid z] = \frac{\mathbb{E}[\Delta(W)s(a(W))g(W) \mid z]}{\mathbb{E}[\Delta(W)s(a(W)) \mid z]}.$$

Substitute $g = \rho_j(a)$ and cancel the factor $s(a)$. The conditional mean prior then satisfies

$$\mathbb{E}_{\widetilde{P}}[\rho_0(a) \mid z] = \frac{\mathbb{E}[\Delta \mid z]}{\mathbb{E}[\Delta s(a) \mid z]} = \rho_0(a_z),$$

$$\mathbb{E}_{\widetilde{P}}[\rho_j(a) \mid z] = \frac{u_j \mathbb{E}[\Delta a_j \mid z]}{\mathbb{E}[\Delta s(a) \mid z]} = \rho_j(a_z). \tag{70}$$

Here $\mathbb{E}[\Delta s(a) \mid z] = \mathbb{E}[\Delta \mid z] s(a_z)$ by (15). Explicitly,

$$\mathbb{E}[\Delta s(a) \mid z] = \mathbb{E}[\Delta \mid z] + \sum_j u_j \mathbb{E}[\Delta a_j \mid z] = \mathbb{E}[\Delta \mid z] \left(1 + \sum_j u_j (a_z)_j\right).$$

The resulting representative is exactly $\rho(a_z)$. This matches the conditional-mean representative for Bayes-risk quantization (Varshney, 2011, Section IV-A). Integrating (32) and substituting (68) gives

$$\begin{aligned}\int (V_Z - V_W)\, d\nu &= \mathbb{E}_P \left[\Delta(W) \int e_\lambda(a(W), a_Z)\, \nu(d\lambda)\right] \\ &= \mathbb{E}_P[\Delta(W) s(a(W)) d_{\mathrm{test}}(\rho(a(W)), \rho(a_Z))] \\ &= C\mathbb{E}_{\widetilde{P}}[d_{\mathrm{test}}(\rho(a(W)), \rho(a_Z))].\end{aligned}$$

The first equality uses Tonelli's theorem on nonnegative regret. Thus

$$\int (V_Z(\lambda) - V_W(\lambda))\, \nu(d\lambda) = C\mathbb{E}_{\widetilde{P}} d_{\mathrm{test}}\big(\rho(a), \mathbb{E}_{\widetilde{P}}[\rho(a) \mid Z]\big). \tag{71}$$

This mapping preserves the number of messages and the conditional-mean representatives. If Q also depends on W beyond a, conditional averaging under $\widetilde{P}$ gives a kernel $\overline{Q}(z \mid a)$ preserving the joint law of (a, Z). The map $a \mapsto \rho(a)$ is invertible on its image, so this is an encoder of the testing prior with the same alphabet. Specifically, take $\overline{Q}(z \mid a) = \mathbb{E}_{\widetilde{P}}[Q(z \mid W) \mid a(W) = a]$. For any bounded measurable g,

$$\mathbb{E}_{\widetilde{P}}[g(a(W))\mathbf{1}\{Z = z\}] = \mathbb{E}_{\widetilde{P}}[g(a(W))Q(z \mid W)] = \mathbb{E}_{\widetilde{P}}[g(a(W))\overline{Q}(z \mid a(W))].$$

This verifies preservation of the joint law. The inverse is $a_j = \rho_j/(u_j\rho_0)$, since $\rho_0 > 0$, so the conditional-mean prior and the distortion in (71) are also preserved.

The connection identifies the averaged component with mean Bayes risk error. Our frontier objective takes a supremum over subsequent budgets, equivalently over weights by (18). The boundary-layer argument in Appendix A.6 provides the corresponding uniform control. The grouping theorem further derives the state and its weighted source density directly from the error partition, making comparisons across partitions possible on one fixed source.

E Experimental Details and Supplementary Illustrations

E.1 Synthetic sources, encoders and frontier evaluation

A common population problem. The main source is uniform on $K = [0.20, 0.26] \times [0.08, 0.14] \times [0.10, 0.16]$. The conditional label law is (1), the primary loss is classification error, and the three severe risks are $R_i = \Pr(Y = i, A = S)$. The posterior is known, so population risks are integrated directly without sampling labels or training a predictor. Here $w_1 > \max(w_2, w_3)$, $\Delta(w) \geq 0.18$ and $b = \min_{w,i} w_i = 0.08$. Thus the common multiplier cap in Lemma A.2 is $\Lambda = 25$. The one-, two- and three-group comparisons use $\{1, 2, 3\}$, $\{1\}\{2, 3\}$ and $\{1\}\{2\}\{3\}$, respectively.

Finite-message construction. For each partition, an enclosing cube of the decision-state support $a_{\mathcal{G}}(K)$ is divided into n equal intervals per coordinate. The message identifies a cell; its decoder uses the conditional risks, equivalently the Δ-weighted state center in (15). The resource axis is the available alphabet size $M = n^m$; empty cells do not transmit messages. The grid is fixed before evaluating any weight or budget. Resolutions on the main source are

$$\begin{aligned} m = 1: &\quad n \in \{4, 8, 16, 32, 64, 128, 256, 512, 1024\}, \\ m = 2: &\quad n \in \{2, 3, 4, 6, 8, 11, 16, 23, 32, 45, 64\}, \\ m = 3: &\quad n \in \{2, 3, 4, 5, 6, 8, 10, 13, 16, 20, 24, 28, 32, 40\}. \end{aligned}$$

The last five three-group resolutions extend the curve to $M = 64{,}000$. The decision-cost comparison and fixed-encoder refinement comparison use the three-group resolutions only through $n = 16$.

Integration and maximization. For a fixed encoder, we evaluate

$$D_{\mathcal{G}}(Q) = \sup_{\lambda \in [0,\Lambda]^m} \mathbb{E}\big[\Delta(W) e_\lambda\big(a_{\mathcal{G}}(W), a_{Q(W)}\big)\big], \tag{72}$$

where e_λ is defined in (31). The identity with the full-budget frontier error follows from (18); the plotted values are numerical estimates of this supremum. In grid-encoder comparisons, expectations and cell centers use 2^{23} Sobol points (Sobol', 1967) per replicate, with SciPy 1.15.3 LMS+shift scrambling (SciPy community, 2025) and seeds $0, 1, 2$. For one group, maximization uses a grid followed by golden-section refinement. For multiple groups, it uses a coarse polar grid followed by Nelder–Mead refinement (Nelder & Mead, 1965). Curves average three replicates of numerical integration, with no encoder-training randomness. On the main source, relative standard errors are at most 0.35% and slopes agree across replicates to three decimal places. A maximization check using one million random directions and 20,000 points per axis found no larger three-group peak.

Using the grouping in quantizer construction. Figure 2 holds the uniform source, message budget and k-means algorithm fixed, and changes the quantization geometry. The grouping-aware representation clusters $a_{\mathcal{G}}(W)$ with sample weights $\Delta(W)$; the grouping-agnostic representation clusters W with equal weights. Both use Lloyd's assignment-and-centroid procedure (Lloyd, 1982), with M clusters, one initialization, at most 100 iterations and random seed zero. The fitting subset contains 2^{18} points from the 2^{23} scrambled Sobol points with integration seed zero. Nearest-center assignment of all integration points is performed on GPU in the respective coordinate space. For each induced cell, both decoders use the conditional center in (15), estimated on the integration points. The message counts are $M \in \{8, 27, 64, 125, 216, 512, 1000, 2197, 4096\}$. Frontier loss uses the same dual maximization protocol as above; the curves use this single integration sequence rather than a three-sequence average. Fits to the five largest M give state-based slopes $-1.89, -1.03, -0.69$ and posterior-based slopes $-0.73, -0.70, -0.73$ for one, two and three groups. For $m = k = 3$, the map $W \mapsto a_{\mathcal{G}}(W)$ is invertible and both quantization spaces remain three-dimensional.

Proportional-risk control. For Figure 1(b), take $W_i = c_i T$, $c = (1/2, 1/3, 1/6)$ and $T \sim \mathrm{Unif}[2/5, 1/2]$. Then $\Delta = 1 - 3T/2$ and the scalar state is $a_{\mathrm{total}} = T/\Delta$. An equal-interval scalar grid with $n = M$ uses the same resolutions as the one-group main comparison. We evaluate this same encoder with merged and fully split risks. The split weighted gap depends on λ only through $c^\top \lambda$, so the equality of the supremum errors follows from (45). The plotted merged and split errors agree to numerical precision.

Table 1: Interpolated message caps for a target frontier loss on the uniform source. Parentheses give $\log_2 M$. The first one-group value is an upper bound because the smallest tested alphabet already meets the target.

Target loss	One group	Two groups	Three groups
10^{-2}	≤ 4 (≤ 2.0)	10 (3.3)	79 (6.3)
3×10^{-3}	7 (2.8)	39 (5.3)	1,366 (10.4)
10^{-3}	13 (3.7)	181 (7.5)	8,716 (13.1)
5×10^{-4}	18 (4.2)	392 (8.6)	26,372 (14.7)
3×10^{-4}	24 (4.6)	696 (9.4)	59,632 (15.9)

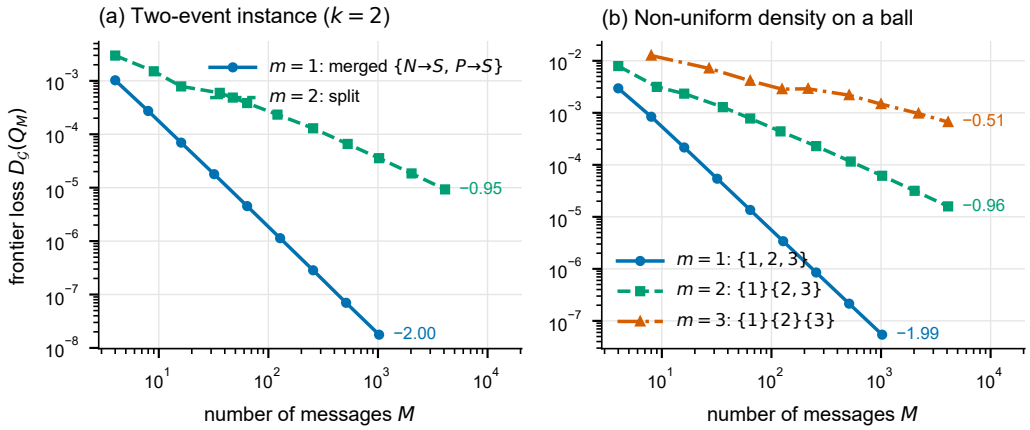

Figure A1: **The separation across sources.** (a) Merged and separate control of the two overclaim events on the uniform rectangular source of Section 4.1. (b) Three groupings of the same three events on a ball with truncated Gaussian density. Both panels show the state-grid encoder's estimated frontier loss against $M = n^m$ on logarithmic axes, averaged over three integration sequences. Endpoint annotations are log–log fits to the five largest message counts in each curve.

Message requirements. Table 1 uses log–log interpolation between adjacent message counts to estimate where each grid-encoder curve first reaches the target error. Log-alphabet size $\log_2 M$ corresponds to $\lceil \log_2 M \rceil$ binary fixed-length bits, rather than mutual information or average code length.

E.2 Other sources, partitions and fixed-encoder refinement

Two events and a nonuniform source. Figure A1(a) instantiates Corollary 4.1 with a uniform posterior source $(U, V) \in [0.19, 0.21] \times [0.29, 0.31]$ and severe events $N \to S$ and $P \to S$. Equivalently, $W = (V, U)$ and action 1 is P. The same state-grid construction, with the one- and two-group resolutions above, gives tail slopes -2.00 and -0.95. Figure A1(b) instead uses a three-dimensional ball of radius 0.028 centered at $(0.23, 0.11, 0.13)$, with density proportional to $\exp[-\|w - (0.23, 0.11, 0.13)\|^2/(2 \times 0.015^2)]$ on that ball. This compact source has a positive bounded nonuniform density and remains inside $\mathcal{O}_3$. The one- and two-group tail slopes are -1.99 and -0.96. The three-group curve, evaluated only through $M = 16^3$, decreases more slowly, with slope -0.51 over its five largest message counts. It shows the same ordering of information costs, but has not reached a tail slope as close to $-2/3$ as the longer main-source curve.

Group membership and refinement. On the main uniform source, Figure A2(a) compares all three two-group partitions. Their fitted slopes, -0.92, -0.90 and -0.92, are similar even though the finite-message errors differ. This separates the group-count exponent from the partition-dependent constants in Theorem 4.2. Panel (b) asks a different question: for each $M = n^3$, we hold the three-group encoder fixed and evaluate it under successively coarser risk families. The errors obey

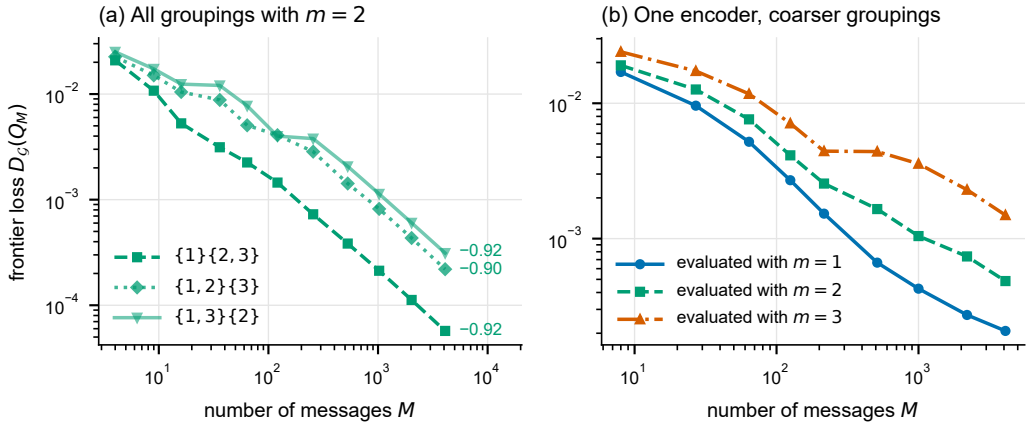

Figure A2: **Changing group membership and holding the encoder fixed.** (a) All two-group partitions of the main uniform source, each with its own two-dimensional state grid. Annotations fit the five largest M. (b) A single three-group encoder at each $M = n^3$ is evaluated with one, two and three groups. The representation is identical across the three curves at a given M. Both panels report mean frontier loss over three integration sequences on logarithmic axes.

$D_{\{1,2,3\}}(Q) \le D_{\{1\}\{2,3\}}(Q) \le D_{\{1\}\{2\}\{3\}}(Q)$ at all nine resolutions and all three integration seeds. Unlike the separately constructed encoders in Figure 1(a), this comparison isolates the fixed-Q task-family inclusion in Corollary 4.3.

E.3 Optimal values and selected budget locations

Frontier profiles. For merged control, write $q(w) = \sum_i w_i$ and $T(w) = \Delta(w)/q(w) = 1/a_{\text{total}}(w)$. The primal frontier is obtained by ordering T from largest to smallest and accumulating risk mass q, with fractional selection at the boundary atom. The finite-message frontier applies the same operation to message conditional risks and improvements. Figure 3(a) displays 2,001 equally spaced budgets from zero to 0.47, for $M = 2, 4, 8, 16$. The primal and dual supremum estimates agree within 6.4×10^{-7}. The largest gap for $M = 2$ is below 0.02; increasing M reduces that gap and the gaps at the other plotted budgets.

An external cost fixed across message sizes. For each grouping we use

$$\Psi(\alpha) = \eta^\top \alpha + \frac{\kappa}{2}\|\alpha\|^2, \qquad \kappa = 1, \qquad \eta = s\mathbf{1}. \tag{73}$$

The scalar s is calibrated using full information to a 50% selection rate for action S, then held fixed over M. We evaluate the optimal value using the conjugate form

$$J_I = \max_{\lambda \ge 0} \left\{ V_I^{\mathcal{G}}(\lambda) - \sum_{g=1}^{m} \phi_\kappa(\lambda_g - \eta_g) \right\}, \qquad \phi_\kappa(t) = \max_{0 \le r \le 1} \{tr - \kappa r^2/2\}. \tag{74}$$

The bounded-budget conjugate $\phi_\kappa(t)$ equals 0 for $t \le 0$, $t^2/(2\kappa)$ for $0 < t < \kappa$, and $t - \kappa/2$ for $t \ge \kappa$. Figure 3(b) uses nine one-group, eleven two-group and nine three-group resolutions, each with three integration replicates. The inequality $0 \le J_Z - J_W \le D_{\mathcal{G}}(Q_M)$ holds for all 87 grouping–resolution–replicate configurations. The ratio $(J_Z - J_W)/D_{\mathcal{G}}(Q_M)$ ranges from 0.07 to 0.98.

Budget location uses curvature. The frontier $F_W^{\mathcal{G}}$ is convex by (17), so the quadratic term makes $F_W^{\mathcal{G}} + \Psi$ κ-strongly convex. Its minimizing budget α_W is consequently unique and satisfies the quadratic-growth condition stated before (12). For exact minimization, the corresponding compressed optimum obeys

$$\|\alpha_Z - \alpha_W\| \le \sqrt{2D_{\mathcal{G}}(Q_M)/\kappa}. \tag{75}$$

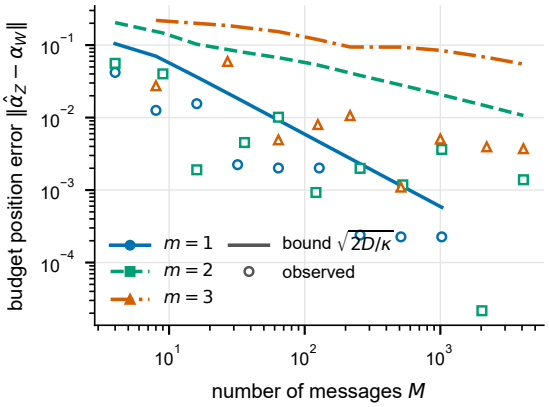

Figure A3: **Selected budgets under a quadratic risk cost.** Open markers show the Euclidean distance between compressed and full-information minimizing budgets for (73); lines show $\sqrt{2D_{\mathcal{G}}(Q_M)/\kappa}$ with $\kappa = 1$. Colors identify the three main-source partitions. Both axes are logarithmic and the displayed quantities average three integration sequences. This is a location comparison under quadratic growth, separate from the optimal-value loss in Figure 3(b).

Figure A3 compares the numerically selected budgets with this bound. All recorded configurations lie below it. The observed position error need not decrease at every successive resolution: the bound controls its size and uses the common positive growth constant.

E.4 Evidence-verification outputs and task-based codebooks

Data and frozen inputs. WiCE consists of claims and cited evidence from Wikipedia with fine-grained entailment annotations (Kamoi et al., 2023). We use a DeBERTaV3 backbone (He et al., 2023), `MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli` (Laurer, n.d.), with a separately trained evidence-aggregation head. Both are frozen; outputs use full evidence or masked neighboring context. The data comprise 3,125 codebook-fit examples, 172 validation examples, 171 probability-calibration examples and 949 development examples. The four splits have disjoint source groups; the development set contains 308 groups. Within each model, both control rules use the same examples and calibrated probability vectors.

Temperature calibration. Before codebook fitting, single-temperature scaling (Guo et al., 2017) minimizes negative log likelihood on the 171 calibration examples from 62 source groups for each model. Writing p_{raw} for the uncalibrated probability vector, the input to the quantizer is $p = \mathrm{softmax}(\log p_{\mathrm{raw}}/T_{\mathrm{cal}})$. The scalar search is bounded to $[0.05, 20]$ with absolute tolerance 10^{-9}. The fitted temperatures are 1.306936 for the full model and 1.293740 for the masked model. Calibration examples were selected from the pre-existing calibration pool by shuffling source groups with seed 30013 and taking the first 62 groups; the remaining groups form the 172-example validation split. Neither the codebook-fit set nor the development set enters the temperature fit.

Loss family and evaluation domain. The primary loss is zero–one error. With the true label as the first argument and the action as the second, the only modified entries are

$$L_{a,t}(N, S) = 1 + a, \qquad L_{a,t}(P, S) = 1 + t. \tag{76}$$

Merged control restricts (a, t) to (λ, λ) for $\lambda \in [0, 6]$; split control uses $(a, t) \in [0, 6]^2$. For a probability vector p, let $r^{\lambda}_{A}(p) = \sum_y p_y L_\lambda(y, A)$ and $\delta_\lambda(p) \in \arg\min_{A \in \{N,P,S\}} r^{\lambda}_{A}(p)$, with tie order N, P, S. Let c_z be a fitted codeword and $Q(p_i)$ the routed message for development example i. The metric in Figure 4 is

$$\widehat{D}_{\mathrm{fit}} = \sup_{\lambda \in \mathcal{L}} \frac{1}{949} \sum_{i=1}^{949} \left[r^{\lambda}_{\delta_\lambda(c_{Q(p_i)})}(p_i) - \min_{A} r^{\lambda}_{A}(p_i) \right], \tag{77}$$

where $\mathcal{L}$ is the diagonal or the square above. The decoder uses the fit codeword c_z directly; it does not replace it with the development-set conditional mean. Thus (77) evaluates the fitted decoder against the full-posterior Bayes rule. It includes any mismatch of the fitted codewords to their development cells, whereas $D_{\mathcal{G}}(Q)$ uses optimal message conditional risks. Development labels do not enter this plug-in metric. Merged evaluation uses 6,001 grid points and golden-section refinement of the ten leading peaks, with tolerance 10^{-9}. Split evaluation uses a 301×301 grid and Nelder–Mead refinement (Nelder & Mead, 1965) from the twenty leading candidates, clipping weights to $[0, 6]^2$.

Fitting and repeated codebooks. We fit a separate codebook for each control rule and each alphabet size $M \in \{2, 4, 8, 16, 32, 64, 128\}$. The fitting distortion is

$$d(p, c) = \operatorname*{average}_{\lambda \in \mathcal{L}} \left[r^{\lambda}_{\delta_\lambda(c)}(p) - \min_A r^{\lambda}_A(p) \right],$$

approximated with 401 midpoint weights for merged control and a 61×61 midpoint grid for split control. Fitting alternates minimum-distortion assignments and updates to the mean fit posterior in each cell, following the assignment-and-centroid structure of Lloyd quantization (Lloyd, 1982). Initializations use posterior k-means centers and four random subsets of fit posteriors, retaining the best fit objective. The k-means initialization uses ten starts, at most 300 iterations and tolerance 10^{-6}. Development routing minimizes the same $d(p, c)$ over the fitted codewords. The average-distortion fitting objective is distinct from the worst-case evaluation metric. There are six fits: one on the original fit set with seed 13, and five on source-group bootstrap samples with seeds $13, 37, 73, 101, 137$. The development set is unchanged. Accordingly, the bands in Figure 4 show the minimum and maximum across fitted codebooks, rather than a confidence interval or variability from training six different verifiers.

At $M \geq 4$, split control has larger error in all 36 size–fit combinations for each model. On the original full-model fit, matching the merged-control regret at $M = 8, 16, 32$ requires split-control alphabets of $32, 64, 128$ among the tested sizes. Each comparison uses four times as many messages, or two additional fixed-length bits. The masked model has the same $8 \to 32$ and $16 \to 64$ comparisons, but its merged $M = 32$ regret is not matched by split control at $M = 128$.

E.5 Identical marginals can conceal different frontiers

Here we fix merged overclaim control and change the joint posterior law, comparing full-information frontiers for risk $\Pr(Y \neq S, A = S)$ on $K_2 = [0.19, 0.21] \times [0.29, 0.31]$. Write $(p_N, p_P, p_S) = (u, v, 1 - u - v)$, $q = u + v$ and $\Delta = 1 - u - 2v$. The question is whether posterior marginals determine the frontier.

Computing the frontier from a weighted exchange ratio. Action P dominates N on K_2, and choosing S instead reduces primary risk by Δ at severe-risk cost q. Define $T = \Delta/q$ and the finite measure $\mu(E) = \mathbb{E}[q\mathbf{1}\{T \in E\}]$. Let $Q^{\downarrow}_{\mu}(s)$ denote its decreasing quantile, parameterized by cumulative mass $0 < s < \mathbb{E}q$. Then

$$F(\alpha) = \mathbb{E}(1 - V) - \int_0^{\min(\alpha, \mathbb{E}q)} Q^{\downarrow}_{\mu}(s)\, ds. \tag{78}$$

Indeed, a policy selecting S with probability $h(u, v)$ has risk $\mathbb{E}(qh)$ and improvement $\mathbb{E}(\Delta h) = \mathbb{E}(qTh)$. Allocating the available q-mass in decreasing order of T, with fractional allocation at a threshold atom, maximizes this improvement. This gives (78): the relevant distribution is the q-weighted law of T.

Two three-point distributions. For the first pair, each row below gives the three equally likely (u, v) values for one distribution:

$$\mathcal{P}_+ = \{(0.195, 0.300), (0.200, 0.305), (0.205, 0.295)\},$$
$$\mathcal{P}_- = \{(0.195, 0.305), (0.200, 0.295), (0.205, 0.300)\}.$$

The marginals of U, V and $1 - U - V$ agree exactly. Both distributions have priors $(0.2, 0.3, 0.5)$, Bayes error 0.5, $F(0) = 0.7$ and $F(0.5) = 0.5$. Nevertheless, the frontiers cross at $\alpha = 149/600$:

$$F_+(33/200) - F_-(33/200) = 1/600, \qquad F_-(199/600) - F_+(199/600) = 1/600. \tag{79}$$

Figure A4(a) subtracts their common chord $0.7 - 0.4\alpha$ to display these exact differences.

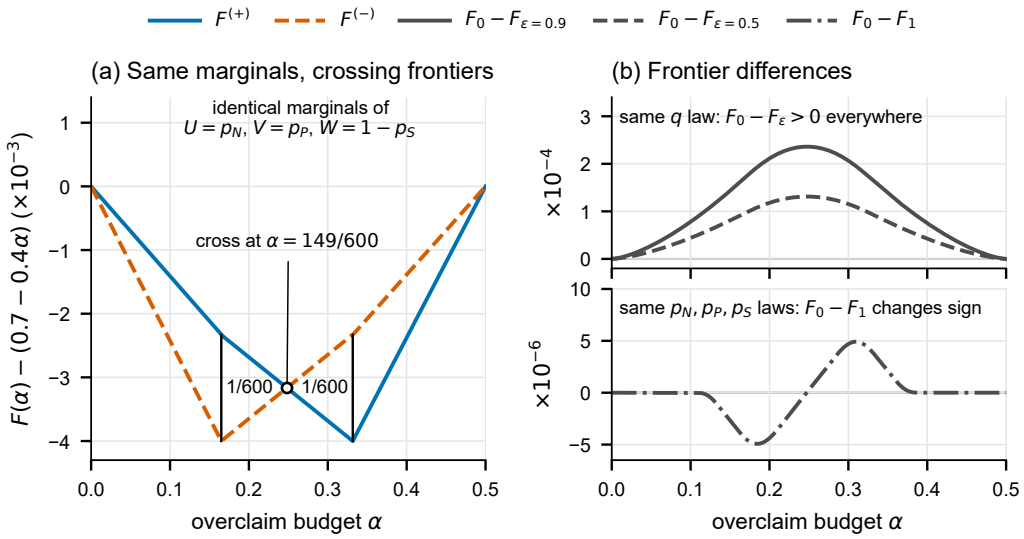

Figure A4: **Joint posterior structure changes the frontier.** All panels fix merged overclaim control and vary the source law. (a) Two three-point laws preserve all three posterior marginals. The common chord $0.7 - 0.4\alpha$ is subtracted; units are 10^{-3}. Vertical segments mark gaps of $1/600$; the scalar W denotes $q = 1 - p_S$. (b, top) Same-q density laws give differences $F_0 - F_\varepsilon$ for $\varepsilon = 0.9$ (solid) and 0.5 (dashed), in units of 10^{-4}: positive inside $(0, 0.5)$ and zero at the endpoints. (b, bottom) Laws preserving all posterior marginals give a sign-changing $F_0 - F_1$, in units of 10^{-6}. Horizontal lines mark zero. All frontiers use full information.

Positive densities with the same overclaim-score law. Let $f_0 = 2500$ on K_2. For $q \in (0.48, 0.52)$, define

$$d(q) = \min(q - 0.48, 0.52 - q), \qquad v_0(q) = q/2 + 0.05, \qquad z = \frac{2(v - v_0(q))}{d(q)}.$$

The perturbed density is $f_\varepsilon(u, v) = 2500[1 + \varepsilon(3z^2 - 1)]$, with $\varepsilon = 0.5$ or 0.9; values at the corner points do not matter. It is positive, preserves the full law of q and $\mathbb{E}[V \mid q] = v_0(q)$, and retains the same priors and frontier endpoints. The resulting $F_0 - F_\varepsilon$ is positive at every interior budget. Figure A4(b, top) shows peaks of approximately 1.31×10^{-4} and 2.36×10^{-4}. This pair preserves the law of q, whereas the next pair also preserves the separate laws of p_N and p_P.

Positive densities preserving every coordinate marginal. Let $r = 1/2000$ and $g_r(x) = \mathbf{1}\{x \in [-r, r]^2\}/(4r^2)$. Using the six centers above, set

$$f_1(x) = 2500 + 10^{-3}\left[\sum_{p \in \mathcal{P}_+} g_r(x - p) - \sum_{p \in \mathcal{P}_-} g_r(x - p)\right], \qquad x \in K_2. \tag{80}$$

The six squares are disjoint and contained in K_2, so $1500 \le f_1 \le 3500$ and its integral is one. The positive and negative centers have identical multisets of u, v and $u + v$ values. The projected perturbations therefore cancel for each coordinate marginal. Figure A4(b, bottom) nevertheless shows both signs of $F_0 - F_1$, with extrema of roughly $\pm 4.9 \times 10^{-6}$. The same priors, Bayes error and separate posterior histograms can thus accompany different, crossing tradeoffs.

Numerical evaluation. The discrete example uses exact rational arithmetic. The density examples use (78) on a 4000×4000 midpoint grid, with perturbation squares aligned to grid cells. Refinement to 8000×8000 changes the frontier differences by at most 5.8×10^{-11}; relative marginal discrepancies are at most 2.3×10^{-14}. No multiplier optimization is used.