Good Enough is Better: Uncertainty-Aware Aggregation for Multi-LLM Systems

Ahmet Faruk Cetinkaya [*1] **Enes Ağırman** [*1] **Cem Tekin** [1]

Abstract

Recent advances in Large Language Models (LLMs) have driven the deployment of complex, multi-stage agentic pipelines. While standard aggregation and multi-agent consensus methods can improve system reliability, sub-agents frequently disagree, and their performance degrades unpredictably under distribution shifts. In this work, we propose the Robust Satisficing Ensemble (RSE), a probabilistic aggregation framework designed to secure multi-agent systems against unpredictable out-of-distribution (OOD) data. By bridging a robust satisficing objective with a quadratic risk surrogate, RSE derives a closed-form analytical inner adversary. This fundamental reformulation enables highly efficient, mathematically guaranteed weight updates without the computational bottleneck of standard minimax optimization. We evaluate RSE across diverse settings, ranging from generative agentic workflows, such as reward modeling under heuristic shifts and selective copying, to standard distribution shift benchmarks (SST-5, TREC, CivilComments, and HuffPost). Our results demonstrate that RSE prevents catastrophic failure and consistently outperforms strong robust baselines, including GroupDRO, IRM, and SRM. These results establish RSE as a practical, theoretically grounded safeguard for reliable agentic routing in dynamic environments.

1. Introduction

Recent advancements in deep learning architectures, particularly the development of self-attention mechanisms and transformer architectures, enabled scalable training of LLMs on large corpora (Vaswani et al., 2017; Shah et al., 2024; Gao et al., 2025). LLMs such as Gemma-3, Phi-4 and Qwen3 have demonstrated strong performance across diverse natural language processing tasks, including text classification, reasoning and sentiment analysis (Abdin et al., 2024b; Team et al., 2025; Yang et al., 2025). As these models continue to improve in scale and capability, they are increasingly deployed as core components in long-horizon agentic workflows and high-stakes applications where reliability, uncertainty quantification, and robustness are crucial (Weidinger et al., 2022; Wang et al., 2025).

Despite their impressive performance, individual LLMs often exhibit sensitivity to prompt phrasing (Razavi et al., 2025), inherent stochasticity (Zhang & Martinez, 2025), or spurious correlations in the training data (Li et al., 2025). To mitigate these issues, modern agentic pipelines increasingly rely on hierarchical compositions and multi-agent routing to stabilize LLM predictions (Jiang et al., 2023b; Turkmen et al., 2025). Architectures formed via diversity-optimized selection (Tekin et al., 2024), prompt-based ensembling (Arora et al., 2023), or weighted majority voting (Taubenfeld et al., 2025) aim to improve predictive reliability, reduce variance, and mitigate the effect of individual sub-agent failures. However, in complex agentic systems, sub-agents or reward models often disagree. Despite empirical gains, these routing systems frequently fail to generalize beyond their calibration data (Ovadia et al., 2019; Oh et al., 2025). Particularly, prompt variations, changes in input characteristics, or distribution shifts can cause unpredictable and pronounced performance degradation (Koh et al., 2021; Białek et al., 2025). Recent studies propose methods to detect or correct distributional changes in the data (Lipton et al., 2018; Miao et al., 2024; Zhang et al., 2024; Zhu et al., 2025). Yet standard aggregation strategies do not explicitly account for such variability, leaving multi-stage pipelines vulnerable to OOD data. Moreover, even under the standard i.i.d. assumption, optimizing routing weights on a finite training set can lead to overly optimistic performance estimates. This occurs because the aggregation configuration tends to exploit sampling noise in the training metric, creating a selection-induced gap between training performance and true out-of-sample reliability (the optimizer's curse) (Smith & Winkler, 2006).

In this paper, we introduce RSE, a principled probabilistic aggregation framework grounded in a novel robust satis-

[*]Equal contribution [1]Department of Electrical and Electronics Engineering, Bilkent University, Ankara, Türkiye. Correspondence to: Enes Agirman <enes.agirman@bilkent.edu.tr>, Ahmet Faruk Cetinkaya <faruk.cetinkaya@ug.bilkent.edu.tr>, Cem Tekin <cemtekin@ee.bilkent.edu.tr>.

Accepted to the ICML 2026 Workshop on Statistical Frameworks for Uncertainty in Agentic Systems, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

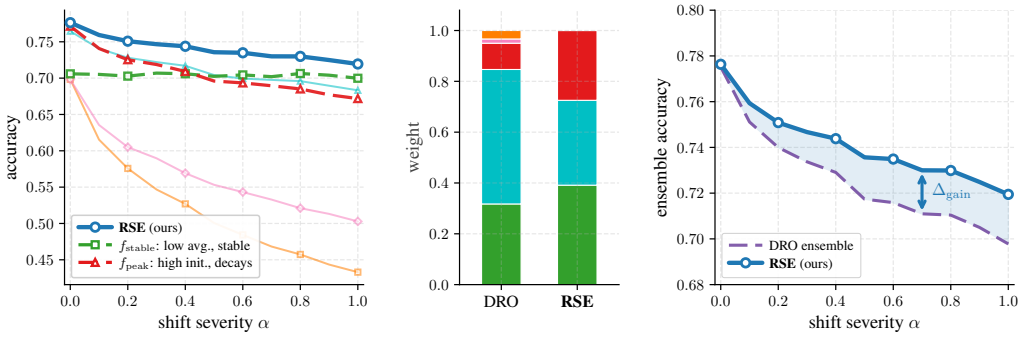


Figure 1. **RSE combines stable and high-peak models to prevent catastrophic failure. (a)** Constituent models face a trade-off between high initial accuracy that decays rapidly (f_{peak}, dashed red triangles) and lower, stable performance (f_{stable}, dashed green squares). The faded solid lines represent the remaining active constituent models participating in the ensemble. The resulting RSE ensemble (thick blue with circles) consistently outperforms individual models across the entire shift trajectory. **(b)** Colors match (a). While DRO yields a suboptimal weight allocation, RSE explicitly minimizes fragility to strategically balance highly stable and high-peak predictors. **(c)** This integration allows RSE to maintain superior, stable accuracy across all shift severities α, achieving a significant gain (Δ_{gain}) over DRO, which collapses alongside its fragile components. *(Note: Visualizations reflect actual empirical results on the TREC dataset. See Section 5 for detailed setups and benchmarks.)*

ficing (RS) formulation. We define a notion of ensemble risk and derive a closed-form solution that minimizes the RS objective, *fragility*, providing a mathematically guaranteed aggregation mechanism to confidently decide which sub-agent's output to trust. Our formulation enables efficient, computationally lightweight aggregation for pre-trained LLMs without requiring model fine-tuning. Figure 1 demonstrates the superior robustness of RSE compared to individual LLMs under severe shift. Through experiments, we demonstrate that RSE consistently outperforms standard baselines in scenarios with OOD data, while maintaining competitive in-distribution performance.

2. Related Works

Aggregation and Routing in LLMs. Classical ensemble methods improve reliability by aggregating diverse models (Breiman, 1996; Dietterich, 2000; Tekin et al., 2016; Lakshminarayanan et al., 2017), with deep learning extensions leveraging optimization trajectory diversity (Huang et al., 2017; Izmailov et al., 2018; Wortsman et al., 2022). However, these traditional techniques remain vulnerable to correlated errors when constituent models share training data. In modern LLM pipelines, diversity is instead induced through prompts, reasoning paths, or sub-agent generations (Naik et al., 2023; Hu et al., 2025). Inference-time aggregation strategies such as self-consistency (Wang et al., 2023), boosted prompting (Pitis et al., 2023), and multi-prompt decoding (Guo et al., 2025) combine multiple reasoning chains to stabilize predictions. Internalized routing, such as mixture-of-experts (Cai et al., 2025), largely targets computational efficiency rather than OOD generalization. Recent multi-agent frameworks have explored agreement-based consensus (Li et al., 2024a), Bayesian-optimal weighting derived from agent correlations (Ai et al., 2025), and contextual bandits for cost-aware routing (Elumar et al., 2025). Yet none of these methods explicitly account for distributional shift at test time. Most directly related to our setting, SRM dynamically adjusts aggregation weights using a distributional graph over training environments (Qiao et al., 2025), but relies on an iterative alternating gradient procedure that introduces computational overhead.

Distributionally Robust Optimization (DRO). A parallel line of work focuses on improving model reliability under distribution shifts through DRO. DRO methods optimize for the worst-case performance across a set of plausible test distributions, often referred to as the *ambiguity set*. This framework provides strong theoretical guarantees and has inspired applications in language modeling and fairness-aware training (Namkoong & Duchi, 2016; Mohajerin Esfahani & Kuhn, 2018; Kim et al., 2025), with works extending these ideas to ensemble learning (Sagawa et al., 2020). However, specifying an appropriate ambiguity set is challenging in practice: too small a set fails to cover the true test distribution, while too large a set leads to overly pessimistic solutions. Furthermore, the resulting min–max objectives introduce significant computational overhead, as the inner maximization must be solved or approximated at each training step. These instabilities are particularly pronounced in multi-agent LLM settings, where the prediction space is high-dimensional and the constituent models share correlated errors from overlapping pretraining corpora, making worst-case optimization simultaneously more difficult and more conservative than necessary.

Robust Satisficing (RS). To address these limitations, RS has been proposed as a more interpretable and flexible alter-

native (Long et al., 2023; Li et al., 2024b; Sim et al., 2025). Rather than optimizing for the worst-case distribution within a predefined ambiguity set, RS specifies a target performance level τ (the aspired in-distribution risk) and minimizes the rate at which performance deteriorates under distributional change, a quantity termed *fragility*. This formulation yields models that meet the target in-distribution performance while controlling the robustness–accuracy trade-off through a single interpretable parameter τ. Unlike DRO, RS does not rely on a fixed ambiguity set but implicitly accounts for a broader class of data distributions, avoiding worst-case over-conservatism while maintaining protection against distributional changes. Early studies have demonstrated its effectiveness in sequential decision making and Bayesian optimization, supported by theoretical guarantees and empirical results (Ruan et al., 2023; Saday et al., 2023; 2025). However, its application to ensemble learning remains largely unexplored, and no existing formulation is directly transferable to the inference-only LLM setting: existing formulations either assume access to a parametric model that can be fine-tuned (Yan et al., 2024), or rely on iterative bisection to solve the inner maximization, which scales poorly when the number of sub-agents M is large and predictions must be aggregated at inference time without gradient access to the underlying models.

To the best of our knowledge, ours is the first work to adapt the robust satisficing framework for the aggregation of pre-trained models. Crucially, our contribution is not merely an application of RS to a new domain, but a fundamental reformulation of its optimization mechanics. Standard RS implementations (such as Yan et al. (2024)) are bottlenecked by intractable minimax dual problems that require slow and unstable iterative bisection searches. By explicitly coupling the RS objective with a quadratic risk surrogate, RSE derives an exact, closed-form analytical inner adversary.

3. Problem Formulation

We consider supervised prediction with an ensemble of M pre-trained LLMs on a K-class task with the label set $\mathcal{Y} := \{y_1, ..., y_K\}$. Let $\mathcal{F} = \{f_1, \ldots, f_M\}$ denote the set of LLMs and let $\Delta_d := \{\mathbf{z} \in \mathbb{R}^d_{\geq 0} : \mathbf{1}^\top \mathbf{z} = 1\}$ denote the probability simplex in d dimensions. Given an input x, model f_m outputs a probability vector $\mathbf{p}_m(x) \in \Delta_K$. The ensemble is parameterized by weights $\mathbf{w} = (w_1, \ldots, w_M)^\top \in \Delta_M$ and outputs the aggregated probability distribution

$$\widehat{\mathbf{p}}_{\mathbf{w}}(x) = \sum_{m=1}^{M} w_m \, \mathbf{p}_m(x). \tag{1}$$

Let $L\colon \mathcal{Y} \times \Delta_K \to \mathbb{R}_{\geq 0}$ be a loss function. For each class $y_j \in \mathcal{Y}$ we define the class-conditional risk

$$r(\mathbf{w}, y_j) := \mathbb{E}_{X \sim P(X|Y=y_j)}\Big[L\big(y_j, \widehat{\mathbf{p}}_{\mathbf{w}}(X)\big)\Big]. \tag{2}$$

Collecting these into a vector gives

$$\mathbf{r}(\mathbf{w}) := \big(r(\mathbf{w}, y_1), \ldots, r(\mathbf{w}, y_K)\big)^\top \in [0, \infty)^K. \tag{3}$$

A label distribution is a vector $\mathbf{q} := (q_1, \ldots, q_K)^\top \in \Delta_K$ with $q_j := P(Y = y_j)$. We fix a *reference* label distribution $\mathbf{q}_{\text{ref}} \in \Delta_K$, representing the nominal operating regime (e.g., a held-out calibration set, the training set). For any $\mathbf{q} \in \Delta_K$, the corresponding expected risk of the ensemble is

$$R(\mathbf{w}, \mathbf{q}) := \sum_{j=1}^{K} q_j \, r(\mathbf{w}, y_j) = \mathbf{q}^\top \mathbf{r}(\mathbf{w}). \tag{4}$$

To quantify how different a prospective deployment distribution $\mathbf{q}$ is from the reference $\mathbf{q}_{\text{ref}}$, we use a nonnegative discrepancy

$$\Delta(\mathbf{q}, \mathbf{q}_{\text{ref}}) \geq 0, \qquad \Delta(\mathbf{q}, \mathbf{q}_{\text{ref}}) = 0 \Longleftrightarrow \mathbf{q} = \mathbf{q}_{\text{ref}}.$$

The framework below is stated for a general Δ; in our algorithm we instantiate Δ with KL divergence, which leads to a closed-form inner adversary.

Aspiration level. Instead of minimizing $R(\mathbf{w}, \mathbf{q}_{\text{ref}})$ directly, we adopt a robust satisficing view. We fix an *aspiration level* $\tau > 0$ and require the ensemble to be satisfactory at the reference distribution:

$$R(\mathbf{w}, \mathbf{q}_{\text{ref}}) \leq \tau. \tag{5}$$

The parameter τ sets the target in-distribution performance: smaller τ demands a stronger baseline at $\mathbf{q}_{\text{ref}}$, while larger τ allows more slack in exchange for stability under changes in data distribution.

Robust Satisficing Objective. The goal is to find ensemble weights $\mathbf{w}$ that are robustly satisfactory. We fix an aspiration level $\tau > 0$, representing the maximum acceptable risk on the source distribution. We then seek to minimize the "fragility" of the ensemble against distribution shifts.

Definition 3.1 (Fragility)**.** Fix a reference distribution $\mathbf{q}_{\text{ref}} \in \Delta_K$ and an aspiration level $\tau > 0$. For a weight vector $\mathbf{w} \in \Delta_M$, the fragility $\kappa_\tau(\mathbf{w})$ is defined as the minimum $\kappa \geq 0$ such that for all $\mathbf{q} \in \Delta_K$:

$$R(\mathbf{w}, \mathbf{q}) \leq \tau + \kappa \, \Delta(\mathbf{q}, \mathbf{q}_{\text{ref}}). \tag{6}$$

Well-posed regime and extended definition. Fix $\tau > 0$ and $\mathbf{q}_{\text{ref}} \in \Delta_K$ with $(q_{\text{ref}})_j > 0$. Define

$$\mathcal{W}_\tau := \Big\{\mathbf{w} \in \Delta_M : \; R(\mathbf{w}, \mathbf{q}_{\text{ref}}) < \tau < \max_{j \in \{1, \ldots, K\}} r(\mathbf{w}, y_j)\Big\}.$$

Outside $\mathcal{W}_\tau$, following the extended-value convention of Saday et al. (2025), we set

$$\kappa_\tau(\mathbf{w}) = \begin{cases} 0, & \text{if } \tau \geq \max_j r(\mathbf{w}, y_j), \\ +\infty, & \text{if } R(\mathbf{w}, \mathbf{q}_{\text{ref}}) \geq \tau. \end{cases}$$

Remark 3.2. While the theoretical formulation relies on the restricted domain $\mathcal{W}_\tau$, in our experiments we found it sufficient to constrain $\mathbf{w} \in \Delta_M$. Across all datasets evaluated, the entire optimization trajectory naturally remained within $\mathcal{W}_\tau$ without explicit enforcement.

Remark 3.3 (Intuition behind the three regimes). The extended definition of $\kappa_\tau(\mathbf{w})$ captures three qualitatively distinct behaviors of the ensemble. When $\tau \geq \max_j r(\mathbf{w}, y_j)$, the aspiration level is so permissive that even the worst single-class risk is acceptable; the ensemble is trivially robust and no reweighting is needed ($\kappa = 0$). When $R(\mathbf{w}, \mathbf{q}_{\text{ref}}) \geq \tau$, the ensemble already fails to meet the aspiration level on the reference distribution itself, making robustness impossible regardless of how the label distribution shifts ($\kappa = +\infty$). The practically relevant interior regime $\mathcal{W}_\tau$ lies between these extremes: the ensemble is satisfactory in-distribution, but some classes are harder than the aspiration level permits under shift. Fragility then measures precisely how fast the worst-case label shift can push the ensemble beyond τ, and minimizing it yields the most shift-resistant aggregation weights.

Small $\kappa_\tau(\mathbf{w})$ means that even adverse changes in label proportions can only increase the risk slowly. As shown in (Saday et al., 2023), this definition implies that fragility can be expressed as:

$$\kappa_\tau(\mathbf{w}) = \sup_{\mathbf{q} \in \Delta_K,\ \mathbf{q} \neq \mathbf{q}_{\text{ref}}} \frac{R(\mathbf{w}, \mathbf{q}) - \tau}{\Delta(\mathbf{q}, \mathbf{q}_{\text{ref}})}. \tag{7}$$

Our final design problem becomes

$$\min_{\mathbf{w} \in \mathcal{W}_\tau} \kappa_\tau(\mathbf{w}). \tag{8}$$

Equivalently, we may view the optimization over Δ_M under the extended definition, which assigns $\kappa_\tau(\mathbf{w}) = +\infty$ when $R(\mathbf{w}, \mathbf{q}_{\text{ref}}) \geq \tau$.

4. The RSE Algorithm

By adopting a squared-error loss and a KL-divergence discrepancy, we show that the RSE can be solved via an efficient alternating procedure that iterates between an analytical inner adversary and a convex quadratic weight update.

4.1. Quadratic Risk Model for LLM Ensembles

We take L to be the *Brier Score* between the one-hot label and the ensemble predicted label distribution: $L\big(y_j, \widehat{\mathbf{p}}_{\mathbf{w}}(x)\big) = \big\|\mathbf{e}_j - \widehat{\mathbf{p}}_{\mathbf{w}}(x)\big\|_2^2$. Using (1) and expanding the square in (2) gives, exactly,

$$\begin{aligned} r(\mathbf{w}, y_j) &:= \mathbb{E}_{X|y_j}\left[\left\|\mathbf{e}_j - \sum_{m=1}^{M} w_m \mathbf{p}_m(X)\right\|_2^2\right] \\ &= \mathbb{E}_{X|y_j}\Bigg[\mathbf{e}_j^\top \mathbf{e}_j - 2\sum_a w_a \mathbf{e}_j^\top \mathbf{p}_a(X) \\ &\qquad + \sum_{a,b} w_a w_b \mathbf{p}_a(X)^\top \mathbf{p}_b(X)\Bigg]. \end{aligned} \tag{9}$$

By linearity of expectation, and noting $\mathbf{e}_j^\top \mathbf{e}_j = 1$, we can bring the expectation inside the summations:

$$\begin{aligned} r(\mathbf{w}, y_j) &= 1 - 2\sum_a w_a \mathbb{E}_{X|y_j}[p_{a,j}(X)] \\ &\quad + \sum_{a,b} w_a w_b \mathbb{E}_{X|y_j}\big[\mathbf{p}_a(X)^\top \mathbf{p}_b(X)\big]. \end{aligned} \tag{10}$$

Define the per-class first and second moments $\bar{\mathbf{p}}_m^{(j)} := \mathbb{E}_{X|Y=y_j}[\mathbf{p}_m(X)]$, $P_{ab}^{(j)} := \mathbb{E}_{X|Y=y_j}\big[\mathbf{p}_a(X)^\top \mathbf{p}_b(X)\big]$ and collect them into $\mathbf{A}_j \in \mathbb{R}^{M \times M}$, $(\mathbf{A}_j)_{ab} = P_{ab}^{(j)}$, $\mathbf{b}_j \in \mathbb{R}^M$, $(\mathbf{b}_j)_m = -2(\bar{\mathbf{p}}_m^{(j)})_j$. Then (10) can be written as the convex quadratic form

$$r(\mathbf{w}, y_j) = \mathbf{w}^\top \mathbf{A}_j \mathbf{w} + \mathbf{b}_j^\top \mathbf{w} + 1. \tag{11}$$

Each $\mathbf{A}_j$ is positive semidefinite (see the derivation and proof in Appendix C). In practice, the expectations in $\bar{\mathbf{p}}_m^{(j)}$ and $P_{ab}^{(j)}$ are estimated from the training data. The details are described in Appendix C.

4.2. KL-Based Fragility and Inner Adversary

We instantiate the discrepancy in Definition 3.1 with the Kullback–Leibler divergence between label distributions: $\Delta(\mathbf{q}, \mathbf{q}_{\text{ref}}) = D_{\text{KL}}(\mathbf{q} \,\|\, \mathbf{q}_{\text{ref}}) = \sum_{j=1}^{K} q_j \log \frac{q_j}{(q_{\text{ref}})_j}$.

For fixed $\mathbf{w}$ and $\kappa > 0$, it is convenient to work with the KL-penalized inner problem where we want to find the label shift that maximizes the penalized expected risk.

$$\Phi(\kappa; \mathbf{w}) = \sup_{\mathbf{q} \in \Delta_K} \Big\{\mathbf{q}^\top \mathbf{r}(\mathbf{w}) - \kappa\, D_{\text{KL}}(\mathbf{q} \,\|\, \mathbf{q}_{\text{ref}})\Big\}, \tag{12}$$

where $\mathbf{r}(\mathbf{w})$ is built from (3). See Appendix D for the derivation. The form of the optimal adversary and the value of (12) can be made explicit.

Lemma 4.1 (KL inner problem)**.** *For any fixed* $\mathbf{w}$ *and* $\kappa > 0$*, the problem* (12) *has a unique maximizer* $\mathbf{q}^\star \in \Delta_K$ *given by*

$$(q^\star)_j = \frac{(q_{\text{ref}})_j \exp\big(r(\mathbf{w}, y_j)/\kappa\big)}{\sum_{k=1}^{K} (q_{\text{ref}})_k \exp\big(r(\mathbf{w}, y_k)/\kappa\big)}, \tag{13}$$

and the corresponding optimal value is

$$\Phi(\kappa;\mathbf{w}) = \kappa \log\Big(\sum_{j=1}^{K}(q_{\mathrm{ref}})_j \exp\big(r(\mathbf{w},y_j)/\kappa\big)\Big). \quad (14)$$

The proof is given in Appendix E. Thus, under a KL penalty, the adversary corresponds to an exponential tilt of the reference label distribution and $\Phi(\kappa;\mathbf{w})$ plays the role of a log-moment generating function of the class-wise risks.

Combining this explicit inner solution with the fragility ratio definition in (7) yields a one-dimensional characterization of the fragility index in the KL case.

Lemma 4.2 (Fragility for KL discrepancy)**.** *Assume* $(q_{\mathrm{ref}})_j > 0$ *for all* j *and let* $\mathbf{w} \in \mathcal{W}_\tau$*, i.e.,*

$$R(\mathbf{w},\mathbf{q}_{\mathrm{ref}}) < \tau < \max_j r(\mathbf{w},y_j).$$

Then there exists a unique $\kappa_\tau(\mathbf{w}) \in (0,\infty)$ *such that*

$$\begin{aligned}\tau &= \Phi\big(\kappa_\tau(\mathbf{w});\mathbf{w}\big)\\ &= \kappa_\tau(\mathbf{w}) \log\Big(\sum_{j=1}^{K}(q_{\mathrm{ref}})_j \exp\big(r(\mathbf{w},y_j)/\kappa_\tau(\mathbf{w})\big)\Big), \quad (15)\end{aligned}$$

and the associated worst-case label distribution is obtained by substituting $\kappa = \kappa_\tau(\mathbf{w})$ *into* (13)*.*

Derivation of (15) follows from Appendix D together with the closed form of Φ in Appendix E. Existence and uniqueness of $\kappa_\tau(\mathbf{w})$, as well as the Lagrange–Bürmann expansion used for computation, are established in Appendix F.

4.3. Robust Satisficing Ensemble Algorithm

We now combine the quadratic risk model and the KL-based fragility into an alternating procedure that approximately solves (8) by reweighting the ensemble on the simplex.

Given current weights $\mathbf{w}_t$, we first compute the per-class risks $r(\mathbf{w}_t, y_j)$ using (11). We then evaluate κ_t using the truncated Lagrange–Bürmann series (Derivation in App. F) and form the associated worst-case label distribution

$$(q_t^\star)_j = \frac{(q_{\mathrm{ref}})_j \exp\big(r_j^{(t)}/\kappa_t\big)}{\sum_{k=1}^{K}(q_{\mathrm{ref}})_k \exp\big(r_k^{(t)}/\kappa_t\big)}.$$

For a candidate $\mathbf{w}$ we define the adversarially tilted quadratic risk

$$\begin{aligned}\tilde{r}_t(\mathbf{w}) &= \sum_{j=1}^{K}(q_t^\star)_j\, r(\mathbf{w},y_j)\\ &= \mathbf{w}^\top \mathbf{A}_{\mathrm{adv},t}\mathbf{w} + \mathbf{b}_{\mathrm{adv},t}^\top \mathbf{w} + 1, \quad (16)\end{aligned}$$

where $\mathbf{A}_{\mathrm{adv},t} = \sum_{j=1}^{K}(q_t^\star)_j \mathbf{A}_j$, $\mathbf{b}_{\mathrm{adv},t} = \sum_{j=1}^{K}(q_t^\star)_j \mathbf{b}_j$. Since each $\mathbf{A}_j$ is PSD and $(q_t^\star)_j \geq 0$, $\mathbf{A}_{\mathrm{adv},t}$ is PSD as well.

Proximal stabilization and stopping. The quadratic subproblem induced by $\mathbf{A}_{\mathrm{adv},t} \succeq 0$ can be ill-conditioned (or admit multiple minimizers on the simplex) when the base models are highly correlated. We therefore perform a *proximal* update by adding a stabilizer $\frac{\rho}{2}\|\mathbf{w}-\mathbf{w}_t\|_2^2$ with $\rho > 0$. This keeps the step a simplex-constrained convex QP while making it ρ-strongly convex, hence ensuring a unique update $\mathbf{w}_{t+1}$.

Beyond well-posedness, the proximal term also provides a principled optimization guarantee: Appendix H shows that, under mild regularity conditions on the KL fragility root, the update yields a *uniform sufficient decrease* in the fragility objective and converges to a stationary point of $\kappa_\tau(\cdot)$ over the simplex. Moreover, the commonly used stopping criterion $\|\mathbf{w}_{t+1}-\mathbf{w}_t\|_2 \leq \varepsilon_{\mathrm{stop}}$ admits an explicit certificate in terms of the KKT residual of the QP subproblem.

The weight update solves the convex quadratic program

$$\begin{aligned}\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} \quad & \mathbf{w}^\top \mathbf{A}_{\mathrm{adv},t}\mathbf{w} + \mathbf{b}_{\mathrm{adv},t}^\top \mathbf{w} + \frac{\rho}{2}\|\mathbf{w}-\mathbf{w}_t\|_2^2\\ \text{s.t.} \quad & \mathbf{w} \in \Delta_M.\end{aligned} \quad (17)$$

Any off-the-shelf convex QP solver (e.g., via `cvxpy`) can be used to compute $\mathbf{w}_{t+1}$. See Algorithm 1 for the overall procedure for implementing RSE.

Proposition 4.3 (Descent, stationarity, and a stopping certificate for proximal RSE)**.** *Assume* $(q_{\mathrm{ref}})_j > 0$ *for all* j *and that the iterates* $\{\mathbf{w}_t\}$ *produced by Algorithm 1 remain in a compact subset* $\mathcal{K} \subset \mathcal{W}_\tau$*. Let* L *be the Lipschitz constant of* $\nabla\kappa_\tau$ *on* $\mathcal{K}$ *(Lemma H.7) and let* $c_{\min} > 0$ *be the uniform lower bound from Lemma H.5. If* $\rho > \frac{L}{2c_{\min}}$*, then the sequence* $\{\kappa_\tau(\mathbf{w}_t)\}_{t\geq 0}$ *is nonincreasing and convergent,* $\|\mathbf{w}_{t+1}-\mathbf{w}_t\|_2 \to 0$*, and every limit point of* $\{\mathbf{w}_t\}$ *is stationary for* $\min_{\mathbf{w}\in\Delta_M}\kappa_\tau(\mathbf{w})$ *in the sense of Definition H.14. Furthermore, the stopping rule* $\|\mathbf{w}_{t+1}-\mathbf{w}_t\|_2 \leq \varepsilon_{\mathrm{stop}}$ *guarantees that the KKT residual of the QP at* $\mathbf{w}_{t+1}$ *is at most* $\rho\,\varepsilon_{\mathrm{stop}}$ *(Corollary H.12).*

Proof sketch. Differentiate the KL root $\Phi(\kappa;\mathbf{w}) = \tau$ to express $\nabla\kappa_\tau(\mathbf{w})$ via the tilted risk gradient, then use ρ-strong convexity of the proximal QP to obtain sufficient decrease and the KKT-based stopping certificate. Full proofs are given in Appendix H.

5. Experiments

We comprehensively evaluate RSE across six diverse datasets designed to expose the core vulnerabilities of modern LLM ensembles: generative pipeline fragility, adversarial label shifts, and natural distribution shifts. For generative workflows, we adapt the Anthropic HH-RLHF dataset (Bai et al., 2022) to evaluate robust reward modeling under heuristic shifts. We also evaluate resilience to positional

Algorithm 1 Robust Satisficing Ensemble (RSE)

Require: reference label distribution $\mathbf{q}_{\text{ref}}$, aspiration level τ, moments $\{\mathbf{A}_j, \mathbf{b}_j\}_{j=1}^K$, proximal weight $\rho > 0$, tolerance $\varepsilon_{\text{stop}}$

Initialize $\mathbf{w}_0 \in \Delta_M$
$\Delta_w \leftarrow +\infty$, $t \leftarrow 0$
while $\Delta_w > \varepsilon_{\text{stop}}$ **do**
  for $j = 1, \ldots, K$ **do**
    $r_j \leftarrow r(\mathbf{w}_t, y_j)$ using (11)
  end for
  $\kappa_t \leftarrow \kappa_\tau(\mathbf{w}_t)$ using the truncated Lagrange–Bürmann series (Appendix F).
  for $j = 1, \ldots, K$ **do**
    $(q_t^\star)_j \leftarrow \frac{(q_{\text{ref}})_j \exp(r_j/\kappa_t)}{\sum_{k=1}^K (q_{\text{ref}})_k \exp(r_k/\kappa_t)}$ (Eq. (13))
  end for
  $\mathbf{A}_{\text{adv},t} \leftarrow \sum_{j=1}^K (q_t^\star)_j \, \mathbf{A}_j$
  $\mathbf{b}_{\text{adv},t} \leftarrow \sum_{j=1}^K (q_t^\star)_j \, \mathbf{b}_j$
  $\mathbf{w}_{t+1} \leftarrow$ solution of (17)
  $\Delta_w \leftarrow \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2$
  $t \leftarrow t + 1$
end while
return $\mathbf{w}_t$ and $\kappa_\tau(\mathbf{w}_t)$

bias and prompt variation using a custom selective-copying task, detailed in Appendix B.5. To assess performance under adversarial conditions, we apply a custom severe-shift protocol to SST-5 (Socher et al., 2013) and TREC (Li & Roth, 2002). Finally, we evaluate natural distribution shifts using CivilComments (Koh et al., 2021) for subpopulation robustness and HuffPost (Yao et al., 2022) for temporal drift. Implementation details, including baseline hyperparameters, aspiration level τ selection, and prompt templates, are in Appendix A; individual-model and extended results are in Appendix B.

Baselines. We compare RSE against Non-optimization baselines (Uniform, Accuracy-Weighted, Majority Vote) and Optimization-based baselines: Empirical Risk Minimization (ERM) (Vapnik, 1998), Group Distributionally Robust Optimization (DRO) (Sagawa et al., 2020), Invariant Risk Minimization (IRM) (Arjovsky et al., 2019), SRM (Qiao et al., 2025), and KL-Divergence-based Robust Satisficing (KL-RS) (Yan et al., 2024).

Models and Setup. To construct the ensemble, we employ a diverse set of $M = 8$ LLMs with varying architectures and parameter counts: Mistral-Small-24B-Instruct (Mistral AI, 2025), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023a), Phi-4 (Abdin et al., 2024b), Phi-3.5 mini (Abdin et al., 2024a), Qwen3-4B (Yang et al., 2025), Gemma-3-12B-Instruct (Team et al., 2025), GPT-oss-20B (Agarwal et al., 2025), and Falcon-3-7B-Instruct (Team, 2024). We extract token-level confidence scores to yield true probability distributions rather than relying on brittle verbalized confidence.

5.1. Uncertainty in Generative Pipelines

Mitigating Heuristic Overfitting in Reward Modeling. In agentic workflows, steering open-ended generation (e.g., LLM-as-a-Judge) heavily relies on pairwise preference ranking. However, standard reward models are susceptible to spurious correlations, such as equating verbosity with quality. To evaluate robustness against such heuristic overfitting, we deployed the $M = 8$ ensemble as a composite reward model on the Anthropic/hh-rlhf dataset. The complete training set comprises over 161,000 pairwise comparisons, performing full inference and extracting token-level probabilities across all eight LLMs is computationally prohibitive. Consequently, we calibrated the framework on a tractable subset of 2,407 samples. Within this training set, we synthesized a challenging *Length Bias Shift*: the calibration environment disproportionately favored longer responses (an 88/12 split between long and short preference). In contrast, the deployment (test) environment of 1,000 samples strictly penalized verbosity in favor of concise, factually dense outputs (consisting entirely of 100% short preferences). As shown in Table 1, standard aggregation via ERM suffered a catastrophic collapse (41.6%), blindly exploiting the length heuristic learned during calibration. In contrast, RSE prevents the aggregation mechanism from over-indexing on this spurious correlation. This advantage translates to a robust 61.4% accuracy, significantly preserving the integrity of the reward signal under severe distribution shift.

5.2. Adversarial Distribution Shifts

To validate RSE on adversarial environments, we evaluate it on SST-5 and TREC. We partition each original training set into training and test pools with matched initial distributions, then synthesize shifts on test pools using a minority target shift protocol. We introduce a shift parameter $\alpha \in [0, 1]$ where the test distribution becomes $P_{\text{test}} = (1 - \alpha) \cdot P_{\text{train}} + \alpha \cdot \delta_{\text{hard}}$, with δ_{hard} concentrating mass on the hardest class (lowest aggregate accuracy across all M ensemble models on held-out validation set). This ranges from $\alpha = 0$ (in-distribution) to $\alpha = 1$ (test set entirely from hardest class).

SST-5 Dataset. We first evaluate RSE on SST-5, a fine-grained sentiment classification benchmark consisting of movie reviews labeled on a scale of 1 (very negative) to 5 (very positive). This domain offers a challenging test for robustness due to the nuanced semantic differences between adjacent sentiment classes (e.g., "negative" vs. "very negative"). We use the official training split (8,544 instances).

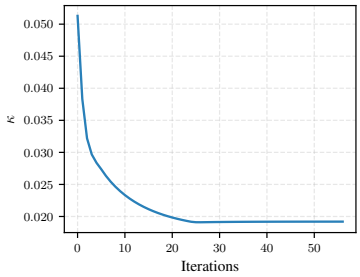
TREC Dataset. Then, we evaluate RSE on the TREC dataset, a widely used benchmark for question classification that comprises questions categorized into six semantic types: abbreviation, description, entity, human, location, and numeric. The dataset's distinct semantic categories make it an

Table 1. **Accuracy (%) on the HH-RLHF Task Under Severe Shift.** RSE mitigates overfitting to the length heuristic. Best results in **bold**.

Task	Uniform	Acc. Weighted Model	Majority Vote	ERM	IRM	KL-RS	SRM	DRO	RSE
RLHF	54.9	54.4	51.0	41.6	42.6	44.2	57.6	59.8	**61.4**

effective testbed for robust classification. We use the official train split of TREC (5452 instances).

Severe Shift Results. While RSE achieves the highest average accuracy across all shift spectrums (detailed in Appendix B), its primary mathematical objective is preventing catastrophic failure in highly unpredictable environments. Table 2 isolates the Severe Shift ($\alpha = 1.00$) regime. Standard baselines collapse under these conditions. On the nuanced sentiment analysis of SST-5, RSE achieves 36.3%, outperforming the strongest robust baseline (DRO) by a massive **+3.8%** and ERM by **+8.0%**. On the TREC dataset, RSE achieves 71.8%, surpassing DRO by **+1.9%**.



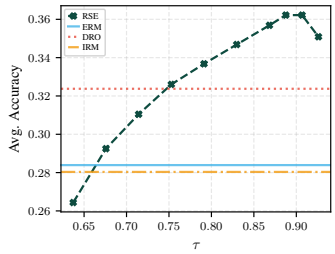
(a) RSE

Figure 2. **Convergence of the RSE optimization on the SST-5 training set.** The plot tracks κ during weight optimization.

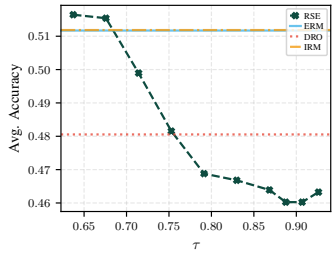
5.3. Natural Distribution Shifts

Subpopulation Shift: CivilComments. To evaluate robustness under subpopulation shift, we use the CivilComments dataset from the WILDS benchmark, which contains online comments annotated for toxicity along with demographic group labels (e.g., Christian, Muslim, LGBTQ, male, female, White, Black). Following standard protocol with official train and test splits, we report both overall accuracy and Worst-Group Accuracy (WGA), which is a critical metric for fairness-aware applications. Table 3 presents the results. Detailed per-group accuracy breakdowns are available in Table 10 of Appendix B. RSE achieves the highest overall accuracy (72.1%) and substantially outperforms all baselines on WGA (59.7%), representing 2.0 points absolute improvement over the next-best method (DRO at 57.7%).

Temporal Shift: HuffPost. Temporal distribution shifts pose a significant challenge for language models, as changes in writing style, topic prevalence, and current events lead to performance degradation over time. We evaluate temporal robustness on the HuffPost dataset, where the task is news category prediction from article headlines. We retain categories that appear consistently across all years, resulting in 9 categories. Following the same temporal split protocol from (Yao et al., 2022), we train algorithms on articles from 2012–2015 and evaluate on 2016–2018. Table 3 reports overall accuracy and Worst-Year Accuracy (WYA). Year-wise performance details are provided in Table 12 of Appendix B. RSE achieves the best WYA (77.2%) and overall accuracy (78.8%), outperforming the second-best methods (IRM and SRM) by 0.4%.



(a) Severe shift ($\alpha = 1$)



(b) No shift ($\alpha = 0$)

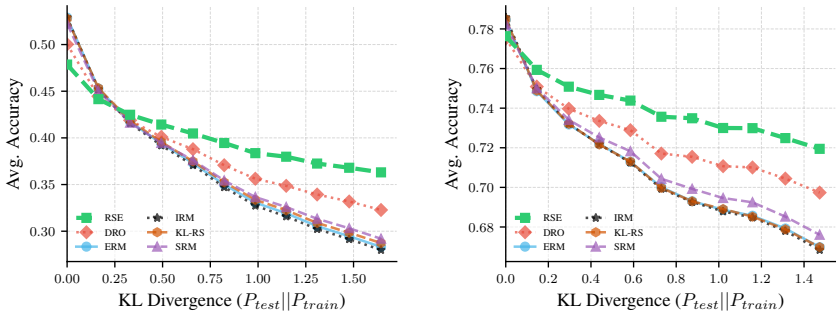
Figure 4. Sensitivity of RSE to aspiration level τ on SST-5.

5.4. Ablation Study

Aspiration level sensitivity analysis for RSE. To assess the role of τ as a sequential monitoring safeguard, we vary this hyperparameter within the solvable range $\tau \in [0.65, 0.94]$ on SST-5. The results reveal a clear robustness-accuracy trade-off governed by τ. Under severe shift (Figure 4a), increasing τ substantially improves worst-case accuracy, with RSE surpassing DRO and IRM beyond $\tau \approx 0.75$ and peaking near $\tau \approx 0.9$. Conversely, under no shift (Figure 4b), higher τ values degrade in-distribution accuracy, allowing practitioners to mathematically bound their risk tolerance.

Table 2. **Performance Under Severe Adversarial Shift ($\alpha = 1.00$).** We report accuracy when the test distribution is shifted entirely to the hardest class. RSE prevents catastrophic failure, significantly outperforming baselines in high-risk regimes. Best results in **bold**.

Dataset	Non-optimization-based			Optimization-based					
	Uniform	Acc. Weighted Model	Majority Vote	ERM	KL-RS	SRM	IRM	DRO	RSE
SST-5	25.1%	24.5%	25.3%	28.3%	28.8%	29.3%	28.0%	32.5%	**36.3%**
TREC	61.7%	62.1%	60.8%	67.1%	66.9%	67.7%	66.7%	69.9%	**71.8%**



(a) SST-5 Accuracy (b) TREC Accuracy

Figure 3. **Performance evaluation under increasing adversarial distribution shift.** We plot the average accuracy against the severity of the shift. RSE consistently maintains higher accuracy compared to baselines as shift severity increases.

Table 3. Performance under natural distribution shifts. Worst Group (Year) and overall accuracies across all groups (years) are given for CivilComments and HuffPost. Best results in **bold**.

	CivilComments		HuffPost	
Method	Acc	WGA	Acc	WYA
ERM	70.1	56.8	78.1	76.7
DRO	70.6	57.7	76.3	75.5
IRM	69.8	56.5	78.4	76.9
SRM	68.8	55.1	78.4	76.9
KL-RS	69.6	56.3	78.1	76.6
RSE (Ours)	**72.1**	**59.7**	**78.8**	**77.2**

Table 4. Per-iteration complexity and training times on HuffPost. M: models, K: classes, N: samples.

MODEL	PER-ITER. COMPLEXITY	TIME (S)
RSE	$O(M^2K + M^3)$	1.41
ERM	$O(NM)$	1.70
KL-RS	$O(MN)$	2.59
IRM	$O(MNK)$	3.03
DRO	$O(MNK)$	4.67
SRM	$O(M^2NK)$	8.30

Computational Complexity. Table 4 compares per-iteration complexity and training times on HuffPost. RSE is the fastest (1.41s), achieving 1.2–6× speedup over baselines. The key advantage is that RSE's per-iteration cost $O(M^2K + M^3)$ is independent of dataset size N: the moment matrices $\{\mathbf{A}_j, \mathbf{b}_j\}_{j=1}^K$ are precomputed once from the calibration set in $O(NM^2K)$ time, after which each RSE iteration operates only on the $M \times K$ sufficient statistics. This decoupling of precomputation from optimization is especially valuable in large-scale deployments where N may be in the tens of thousands but M and K remain modest. Memory overhead is also favorable: storing the moment matrices requires $O(M^2K)$ space, independent of N, compared to $O(NMK)$ for methods such as DRO and IRM that must retain per-sample predictions throughout training. As M grows with larger ensembles, the $O(M^3)$ term from the simplex-constrained QP solve dominates, but remains negligible for the ensemble sizes ($M \leq 16$) typical in current agentic deployments.

6. Conclusion, Limitations, and Future Work

We introduced RSE, a probabilistic aggregation framework that secures multi-agent pipelines against unpredictable distribution shifts by optimizing for *fragility*. Evaluated across generative tasks and diverse shift benchmarks, RSE consistently prevented catastrophic failures, outperforming strong robust baselines while remaining the computationally lightest optimization-based method. Currently, RSE requires token-level log-probabilities. Extending the framework to entirely black-box systems that expose only hard predictions or natural language outputs remains an open challenge. Additionally, RSE assumes a fixed calibration distribution. Adapting the framework for online sequential monitoring, dynamically updating moment matrices as the reference distribution $\mathbf{q}_{\text{ref}}$ drifts, is a critical direction for non-stationary environments. Finally, extending the quadratic risk surrogate beyond the Brier score to handle structured, open-ended generation will further broaden RSE's applicability.

Acknowledgements

This work was supported by the Scientific and Technological Research Council of Türkiye (TUBITAK) under Grant 124E065; TUBITAK 2024 Incentive Award; by the Turkish Academy of Sciences Distinguished Young Scientist Award Program (TUBA-GEBIP-2023).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024a.

Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., Harrison, M., Hewett, R. J., Javaheripi, M., Kauffmann, P., et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024b.

Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., Arbus, E., Arora, R. K., Bai, Y., Baker, B., Bao, H., et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Ai, R., Pan, Y., Simchi-Levi, D., Tambe, M., and Xu, H. Beyond majority voting: Llm aggregation by leveraging higher-order information. *arXiv preprint arXiv:2510.01499*, 2025.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Arora, S., Narayan, A., Chen, M. F., Orr, L., Guha, N., Bhatia, K., Chami, I., and Re, C. Ask me anything: A simple strategy for prompting language models. In *International Conference on Learning Representations*, 2023.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Białek, J., Kivimäki, J., Kuberski, W., and Perrakis, N. Estimating model performance under covariate shift without labels. *Advances in Neural Information Processing Systems*, 38, 2025.

Breiman, L. Bagging predictors. *Machine Learning*, 24: 123–140, 1996.

Cai, W., Jiang, J., Wang, F., Tang, J., Kim, S., and Huang, J. A survey on mixture of experts in large language models. *IEEE Transactions on Knowledge and Data Engineering*, 37:3896–3915, 2025.

Dietterich, T. G. Ensemble methods in machine learning. In *Proceedings of the First International Workshop on Multiple Classifier Systems*, MCS '00, pp. 1–15, 2000.

Elumar, E. C., Tekin, C., and Yagan, O. Cost-aware LLM-based online dataset annotation. *Advances in Neural Information Processing Systems*, 38, 2025.

Gao, T., Wettig, A., Yen, H., and Chen, D. How to train long-context language models (effectively). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7376–7399, 2025.

Guo, J., Wei, D., Luo, Y., Shang, H., Li, Z., Yang, J., Wu, Z., Rao, Z., Tao, S., and Yang, H. M-ped: Multi-prompt ensemble decoding for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 16693–16711, 2025.

Hu, W., Lau, G. K. R., Diwen, L., Jizhuo, C., Ng, S.-K., and Low, B. K. H. Dipper: Diversity in prompts for producing large language model ensembles in reasoning tasks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35558–35572, 2025.

Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J. E., and Weinberger, K. Q. Snapshot ensembles: Train 1, get m for free. In *International Conference on Learning Representations*, 2017.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D. P., and Wilson, A. G. Averaging weights leads to wider optima and better generalization. In *Conference on Uncertainty in Artificial Intelligence*, 2018.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., et al. Mistral 7b, 2023a.

Jiang, D., Ren, X., and Lin, B. Y. LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14165–14178, 2023b.

Kim, C. W., Verma, S., Tec, M., and Tambe, M. DPO-PRO: Direct preference optimization with preference robustness. In *NeurIPS 2025 Workshop: Reliable ML from Unreliable Data*, 2025.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017.

Li, J., Zhang, Q., Yu, Y., Fu, Q., and Ye, D. More agents is all you need. *arXiv preprint arXiv:2402.05120*, 2024a.

Li, X. and Roth, D. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002.

Li, Z., Xu, Y., and Zhan, R. Statistical properties of robust satisficing. In *International Conference on Machine Learning*, pp. 29112–29127. PMLR, 2024b.

Li, Z., Wen, X., Lou, J., Ji, Y., Lu, Y., Han, X., Zhang, D., and Sun, L. The devil is in the details: Tackling unimodal spurious correlations for generalizable multimodal reward models. In *International Conference on Machine Learning*, pp. 36281–36297. PMLR, 2025.

Lipton, Z., Wang, Y.-X., and Smola, A. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*, pp. 3122–3130. PMLR, 2018.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Long, D. Z., Sim, M., and Zhou, M. Robust satisficing. *Operations Research*, 71(1):61–82, 2023.

Miao, W., Pang, G., Zheng, J., and Bai, X. Long-tailed out-of-distribution detection via normalized outlier distribution adaptation. *Advances in Neural Information Processing Systems*, 37, 2024.

Mistral AI. Mistral small 24b instruct 2501, January 2025. URL https://mistral.ai/news/mistral-small-3/.

Mohajerin Esfahani, P. and Kuhn, D. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018.

Naik, R., Chandrasekaran, V., Yuksekgonul, M., Palangi, H., and Nushi, B. Diversity of thought improves reasoning abilities of llms. *arXiv preprint arXiv:2310.07088*, 2023.

Namkoong, H. and Duchi, J. C. Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in Neural Information Processing Systems*, 29, 2016.

Oh, C., Fang, Z., Im, S., Du, X., and Li, Y. Understanding multimodal LLMs under distribution shifts: An information-theoretic approach. In *International Conference on Machine Learning*, pp. 46943–46970. PMLR, 2025.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems*, 32, 2019.

Pitis, S., Zhang, M. R., Wang, A., and Ba, J. Boosted prompt ensembles for large language models. *arXiv preprint arXiv:2304.05970*, 2023.

Qiao, F., Chen, Y., and Peng, X. Structure-informed risk minimization for robust ensemble learning. In *International Conference on Machine Learning*, pp. 50130–50141. PMLR, 2025.

Razavi, A., Soltangheis, M., Arabzadeh, N., Salamat, S., Zihayat, M., and Bagheri, E. Benchmarking prompt sensitivity in large language models. In *European Conference on Information Retrieval*, pp. 303–313, 2025.

Ruan, H., Zhou, S., Chen, Z., and Ho, C. P. Robust satisficing MDPs. In *International Conference on Machine Learning*, pp. 29232–29258. PMLR, 2023.

Saday, A., Yıldırım, Y. C., and Tekin, C. Robust bayesian satisficing. *Advances in Neural Information Processing Systems*, 36, 2023.

Saday, A., Yıldırım, Y. C., and Tekin, C. Robust satisficing gaussian process bandits under adversarial attacks. *Advances in Neural Information Processing Systems*, 38, 2025.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

Shah, J., Bikshandi, G., Zhang, Y., Thakkar, V., Ramani, P., and Dao, T. Flashattention-3: Fast and accurate attention with asynchrony and low-precision. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Sim, M., Tang, Q., Zhou, M., and Zhu, T. The analytics of robust satisficing: Predict, optimize, satisfice, then fortify. *Operations Research*, 73(5):2708–2728, 2025.

Smith, J. E. and Winkler, R. L. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52(3):311–322, 2006.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, 2013.

Taubenfeld, A., Sheffer, T., Ofek, E., Feder, A., Goldstein, A., Gekhman, Z., and Yona, G. Confidence improves self-consistency in llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20090–20111, 2025.

Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Team, T. The falcon 3 family of open models, December 2024.

Tekin, C., Yoon, J., and Van Der Schaar, M. Adaptive ensemble learning with confidence bounds. *IEEE Transactions on Signal Processing*, 65(4):888–903, 2016.

Tekin, S. F., Ilhan, F., Huang, T., Hu, S., and Liu, L. LLM-TOPLA: Efficient LLM ensemble by maximising diversity. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 11951–11966, 2024.

Turkmen, Y., Buyukates, B., and Bastopcu, M. Balancing information accuracy and response timeliness in networked llms. *arXiv preprint arXiv:2508.02209*, 2025.

Vapnik, V. N. *Statistical Learning Theory*. Wiley, New York, 1998.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Wang, C., Liu, Y., Bi, B., Zhang, D., Li, Z.-Z., Ma, Y., He, Y., Yu, S., Li, X., Fang, J., Zhang, J., and Hooi, B. Safety in large reasoning models: A survey. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 3468–3482, 2025.

Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., et al. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 214–229, 2022.

Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, pp. 23965–23998. PMLR, 2022.

Yan, H., Zhou, M., and Guo, J. On the kl-divergence-based robust satisficing model. *arXiv preprint arXiv:2408.09157*, 2024.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yao, H., Choi, C., Cao, B., Lee, Y., Koh, P. W. W., and Finn, C. Wild-time: A benchmark of in-the-wild distribution shift over time. *Advances in Neural Information Processing Systems*, 35, 2022.

Zhang, Y. and Martinez, I. From stochasticity to signal: A bayesian latent state model for reliable measurement with llms. *arXiv preprint arXiv:2510.23874*, 2025.

Zhang, Z.-Y., Xie, Z., Yao, H., and Sugiyama, M. Test-time adaptation in non-stationary environments via adaptive representation alignment. *Advances in Neural Information Processing Systems*, 37, 2024.

Zhu, M., Liu, Y., Fu, Z., Zhang, Y., and Mao, Z. Leveraging robust optimization for llm alignment under distribution shifts. *Advances in Neural Information Processing Systems*, 38, 2025.

A. Implementation Details

For our computational setup, the LLMs were executed using an NVIDIA A100 GPU server (80 GB VRAM), whereas the experimental evaluations were performed locally on a personal computer.

A.1. RSE Parameters and Selection of the Aspiration Level (τ)

In all reported experiments, the reference label distribution $\mathbf{q}_{\mathrm{ref}}$ was set exactly to the empirical marginal class distribution of the calibration training set. Because all evaluated datasets contain strictly positive sample counts for every class within their respective training splits, no artificial smoothing was required or applied. The proximal weight ρ and $\varepsilon_{\text{stop}}$ are chosen as 5 and 10^{-5}, respectively. The infinite series obtained from the Lagrange–Bürmann expansion is approximated by retaining terms up to order $N = 8$. To determine the aspiration level τ, we employ a data-driven heuristic that balances performance on the source distribution with robustness against worst-case class shifts. We first compute the standard Empirical Risk Minimization (ERM) solution, denoted as $\mathbf{w}_{\text{ERM}}$ (Also used in ensemble weight initialization), on the training source distribution P_{train}. Using this baseline, we evaluate the corresponding class-conditional risks $\{r(\mathbf{w}_{\mathrm{ERM}}, y_j)\}_{j=1}^{K}$. We then define the source risk and worst-class risk as

$$R_{\mathrm{source}} := R(\mathbf{w}_{\mathrm{ERM}}, \mathbf{q}_{\mathrm{ref}}) = \sum_{j=1}^{K} (q_{\mathrm{ref}})_j \, r(\mathbf{w}_{\mathrm{ERM}}, y_j), \qquad R_{\mathrm{worst}} := \max_{j \in \{1,\dots,K\}} r(\mathbf{w}_{\mathrm{ERM}}, y_j).$$

The threshold τ is then defined as a linear interpolation between the source risk and the worst-case risk, controlled by a hyperparameter $\lambda \in [0, 1]$:

$$\tau = R_{\text{source}} + \lambda \left(R_{\text{worst}} - R_{\text{source}} \right) \tag{18}$$

The hyperparameter λ determines the tolerance for relaxing the source risk constraint to gain robustness. In our experiments, we intentionally selected τ values that correspond to the maximum feasible values within the solvable range. Specifically, we set $\lambda = 0.70$ for SST-5, $\lambda = 0.62$ for TREC, $\lambda = 0.80$ for hh-rlhf, $\lambda = 0.95$ for selective copying, $\lambda = 0.80$ for CivilComments, and $\lambda = 0.19$ for HuffPost. This strategy ensures the lowest possible fragility κ, thereby maximizing the ensemble's robustness against severe distribution shifts.

How to Pick τ in Practice: Higher values of τ typically reduce the fragility coefficient κ, but they also relax the nominal risk requirement. Thus, τ controls an intercept–slope trade-off in the certificate $R(\mathbf{w}, \mathbf{q}) \leq \tau + \kappa_\tau(\mathbf{w})\varepsilon$. Therefore, picking τ is not a blind hyperparameter search, but a deterministic reflection of the deployment environment. Our mathematical bounds provide the exact minimum and maximum feasible values for τ. Practitioners simply use an interpolation parameter $\lambda \in [0, 1]$ to select their desired risk profile:

- **Low-Risk, Predictable Environments:** If the practitioner expects minimal distribution shift, they set τ close to the nominal ERM risk ($\lambda \approx 0$). This explicitly guarantees on-par performance at low shift regimes.
- **High-Risk, Volatile Environments:** If the practitioner expects severe shifts, the optimal strategy is to select a high value of τ ($\lambda \to 1$). In practice, the upper limit of λ is dictated by the numerical stability of the inner fragility root. As τ approaches the worst-case risk, the fragility κ approaches 0. The practitioner can therefore increase λ to the highest value before the exponential tilt adversary ($e^{r_j/\kappa}$) encounters numerical overflow limits, thereby guaranteeing the maximum computationally stable protection.

This approach is completely deterministic and automatically ensures the ensemble gains the optimal level of protection without breaking the closed-form optimization. Crucially, it entirely eliminates the need for expensive hyperparameter tuning on a validation set.

A.2. Hyperparameter Settings for the Baseline Algorithms

We provide complete hyperparameter settings for all methods to ensure reproducibility. In label shift experiments, we repeated the experiments 20 times and reported the averaged results.

ERM. It optimizes ensemble weights by minimizing cross-entropy loss on the training set. We use a learning rate of 0.01 with 500 optimization iterations. The weights are parameterized via softmax to ensure simplex constraints.

Group DRO. We implement Group DRO using exponentiated gradient ascent for the adversarial distribution q and gradient descent for the ensemble weights w. The method runs for 200 iterations, utilizing a step size of $\eta_q = 0.1$ for the q update and a learning rate of $\eta_w = 0.1$ for the w update. We perform 5 inner optimization steps per iteration and optimize using the cross-entropy loss function.

IRM. For Invariant Risk Minimization, we treat each class as a distinct environment and apply the IRMv1 penalty. The model is trained for 200 iterations using a learning rate of 0.01 and a penalty weight set to $\lambda_{\text{IRM}} = 0.4$.

SRM. For Structure-informed Risk Minimization, we construct the distributional graph utilizing 2-Wasserstein distances between class-conditional prediction distributions, employing an extended feature representation of concatenated per-model predictions. We determine node importance via closeness centrality and train for 200 iterations with a KL regularization weight of $\lambda_{\text{reg}} = 1.0$, using a learning rate of 0.1 for both the weights w and the distribution q.

KL-RS. For KL-Divergence-based Robust Satisficing, we employ a bisection search over the range $\lambda \in [10^{-5}, 10^{3}]$ to identify the optimal robustness parameter λ^*. We set the target loss factor to $\tau_{\text{factor}} = 1.1$ (yielding $\tau = 1.1 \times L_{\text{ERM}}$) and utilize a convergence threshold of $\epsilon = 10^{-3}$. The inner optimization process involves 300 iterations with a learning rate of 0.05.

Non-optimization Baselines. For the static methods, the **Uniform Ensemble** assigns equal weights $w_m = 1/M$ for all models $m \in [M]$. The **Accuracy-Weighted** approach assigns weights proportional to the test accuracy of each model, such that $w_m \propto \text{Acc}_m$. Finally, the **Majority Vote** baseline utilizes hard voting derived from the $\arg\max$ predictions of each individual model.

A.3. Dataset Splits and Partitioning

To ensure rigorous evaluation and reproducibility across our diverse experimental settings, we apply specific partitioning strategies depending on the nature of the distribution shift.

Adversarial Label Shift Datasets (SST-5 and TREC): For the SST-5 and TREC datasets, we utilize their respective official training splits as our primary foundational data. Because our adversarial shift protocol requires synthesizing specific target shifts from a held-out pool, we partition this main official train split into three subsets:

- **Test Pool:** 10% of the primary split is reserved as the test pool (`TEST_POOL_FRACTION = 0.1`). We draw from this pool to synthesize the varying degrees of test-time label shift (α).
- **Validation Set:** 10% of the primary split is allocated for validation purposes (`VALIDATION_FRACTION = 0.1`), which is utilized to identify the hardest class for the severe shift protocol.
- **Training Set:** The remaining 80% of the primary split is used as the calibration training set to estimate the base class-conditional risks and compute the moment matrices for the ensemble.

Our SST-5 experiments utilize the official training split containing 8,544 instances. For TREC, we use the official train split consisting of 5,452 instances.

Natural Distribution Shift Datasets (CivilComments and HuffPost): For our evaluations targeting natural, real-world distribution shifts, we strictly adhere to the established benchmark splits from the WILDS framework to maintain comparability with prior literature.

- **CivilComments:** We use the official train and test splits provided by the WILDS benchmark to evaluate subpopulation shift. The dataset comprises 269,038 training samples and 133,782 test samples.
- **HuffPost:** We utilize the official temporal train and test splits provided by the WILDS benchmark, where the ensemble is calibrated on the earlier years (2012–2015, comprising 84,402 training samples) and evaluated on the subsequent years (2016–2018, comprising 12,058 test samples) to measure robustness against temporal drift.

A.4. Probability Vector Extraction and Prompt Templates

To obtain a probability distribution $\mathbf{p}_m(x) \in \Delta_K$ for a given input x and model m, we utilize a constrained log-probability scoring method. This approach bypasses the stochasticity of standard text generation and ensures valid, dense probability vectors even for smaller models that might otherwise hallucinate formats.

For each task, we map the K class labels to a fixed set of identifiers $\mathcal{A} = \{A, B, \dots\}$. For every candidate identifier $y_k \in \mathcal{A}$, we construct a target continuation string s_k that represents the strictly formatted answer expected from the model (e.g., `{"answer": "A"}`).

We compute the scalar score z_k for class k as the sum of the log-probabilities of the tokens in s_k, conditioned on the input prompt x_{prompt}:

$$z_k = \sum_{t=1}^{|s_k|} \log P_\theta \left(s_k^{(t)} \mid x_{\text{prompt}}, s_k^{(1:t-1)} \right) \tag{19}$$

where $s_k^{(t)}$ denotes the t-th token of the answer string. Once the raw scores $\mathbf{z} = [z_1, \dots, z_K]$ are computed for all K classes, we apply a softmax function to normalize them into a probability distribution:

$$p_m(x)_k = \frac{\exp(z_k)}{\sum_{j=1}^{K} \exp(z_j)} \tag{20}$$

This vector $\mathbf{p}_m(x)$ is then used directly as the input to the ensemble aggregation algorithms.

We employ a standard zero-shot formatting across all models. In Figure 5, the specific prompt templates used for each dataset can be seen.

A Note on Probability Calibration and the Brier Score. Constrained log-probability scoring inherently sharpens output distributions, yielding overconfident vectors that distort absolute risk. Crucially, this empirical misalignment does not compromise our theoretical guarantees. The framework's mathematical foundation, specifically the positive semi-definiteness of the moment matrices $\mathbf{A}_j$ and the convergence of proximal updates, relies solely on the simplex constraint ($\mathbf{p}_m \in \Delta_K$) and the algebraic structure of the Brier score, making it entirely agnostic to perfect calibration. In practice, this distorted risk landscape is naturally neutralized by our data-driven aspiration level τ. Because τ is formulated as a relative interpolation between the empirical source risk and the worst-case class risk, it dynamically scales to accommodate the sharpened probability space. Consequently, RSE successfully minimizes the *relative* fragility among constituent models, ensuring optimal robust aggregation even when absolute probability calibration is flawed.

B. Additional Experiment Results

In this section, we present comprehensive numerical results supplementing the main experimental evaluation. We provide granular performance breakdowns for both the adversarial distribution shift experiments (SST-5, TREC) and the natural distribution shift experiments (CivilComments, HuffPost).

B.1. Label Shift Datasets

In this subsection, we present the extended results under minority target label shift for the SST-5 and TREC datasets. We evaluate the performance stability of RSE compared to baselines across the full spectrum of shift severities, denoted by the shift magnitude parameter $\alpha \in [0, 1]$.

- $\alpha = 0.00$: Represents the **In-Distribution** setting (test distribution matches training).
- $\alpha = 1.00$: Represents the **Worst-Case** shift, where the test set is composed entirely of the class that is hardest for the ensemble to predict.

Statistical Significance. Tables 5 and 6 provide the paired t-test significance results for the SST-5 and TREC datasets, respectively, comparing RSE against all optimization-based baselines. The results mathematically validate RSE's core hypothesis: conservatively trading off marginal in-distribution performance to achieve statistically significant, superior robustness in moderate-to-severe shift regimes.

Dataset: SST-5 (Sentiment Analysis)

[System Instruction]
You are an expert at sentiment analysis. Classify the user's text into one of five categories: very negative, negative, neutral, positive, or very positive. Respond with ONLY a JSON object in this exact format: {`"answer": "A"`} where the value is A, B, C, D, or E.

[User Input]
Review: `[Input Text]`
Choices:
A. very negative B. negative C. neutral D. positive E. very positive

Target: {`"answer": "E"`}

Dataset: TREC (Question Classification)

[System Instruction]
You are an expert at question classification. Classify the user's question into one of six categories. Respond with ONLY a JSON object in this exact format: {`"answer": "A"`} where the value is A, B, C, D, E, or F.

[User Input]
Question: `[Input Question]`
Choices:
A. abbreviation B. description C. entities D. human beings E. locations F. numeric values

Target: {`"answer": "F"`}

Dataset: CivilComments (Toxicity)

[System Instruction]
You are an expert at detecting toxic comments online. Classify whether the following comment is toxic or non-toxic. Respond with ONLY: {`"answer": "X"`} where X is A or B.

[User Input]
Comment: `[Input Text]`
Choices:
A. non-toxic B. toxic

Target: {`"answer": "B"`}

Dataset: HH-RLHF (Preference Ranking)

[System Instruction]
You are an expert AI alignment judge. Given a conversation context and two candidate responses, evaluate which response is better. Respond with ONLY a JSON object in this exact format: {`"answer": "A"`} where the value is A or B.

[User Input]
Context: `[Conversation Context]`
Response A: `[Candidate Response A]`
Response B: `[Candidate Response B]`

Target: {`"answer": "A"`}

Dataset: HuffPost (News Categorization)

[System Instruction]
You are an expert at news classification. Classify the user's headline into one of 9 categories. Respond with ONLY a JSON object.

[User Input]
Headline: `[Input Headline]`
Choices:
A. Black Voices B. Business ... I. Queer Voices

Target: {`"answer": "A"`}

Dataset: Selective Copying (Input Shift)

[System Instruction]
Extract the single uppercase letter hidden in the sequence of numbers below.

[User Input — Hard split: target in MIDDLE]
Sequence: `[Noise Tokens] [Target Letter] [Noise Tokens]`
The hidden letter is:

Target: `"X"`

Figure 5. Prompt templates used for all six evaluated datasets.

Extended Accuracy Breakdowns. For complete transparency, Tables 7 and 8 supply the granular accuracy breakdowns for both datasets. These tables include not only the primary optimization algorithms but also simple non-optimization heuristics and the raw performance of all individual constituent models in the ensemble.

Table 5. **Significance testing on SST-5 under increasing minority target shift.** We report mean accuracy (%) $\pm$ standard deviation across 20 independent experiments. The highest accuracy for each shift severity is marked in **bold**. Statistical significance against our proposed method (RSE) is determined via a paired t-test: * denotes $p < 0.05$ and ** denotes $p < 0.01$.

Shift (α)	RSE (Ours)	DRO	IRM	SRM	KL-RS	ERM
0.00	48.0 ± 1.2	$50.0 \pm 1.9^{**}$	$\mathbf{52.9 \pm 1.7}^{**}$	$52.2 \pm 1.4^{**}$	$52.8 \pm 1.6^{**}$	$\mathbf{52.9 \pm 1.6}^{**}$
0.31	44.2 ± 1.7	$44.5 \pm 2.1^{*}$	$45.2 \pm 1.8^{**}$	$45.1 \pm 1.7^{**}$	$\mathbf{45.4 \pm 1.8}^{**}$	$45.3 \pm 1.8^{**}$
0.45	$\mathbf{42.5 \pm 2.1}$	$42.0 \pm 2.4^{**}$	$41.6 \pm 2.1^{**}$	$41.7 \pm 2.0^{**}$	$41.9 \pm 2.1^{**}$	$41.7 \pm 2.1^{**}$
0.56	$\mathbf{41.5 \pm 2.2}$	$40.2 \pm 2.2^{**}$	$39.2 \pm 2.0^{**}$	$39.5 \pm 1.9^{**}$	$39.6 \pm 2.0^{**}$	$39.4 \pm 2.0^{**}$
0.65	$\mathbf{40.5 \pm 2.8}$	$38.9 \pm 2.7^{**}$	$37.1 \pm 2.5^{**}$	$37.7 \pm 2.4^{**}$	$37.5 \pm 2.4^{**}$	$37.3 \pm 2.5^{**}$
0.73	$\mathbf{39.5 \pm 2.7}$	$37.2 \pm 2.6^{**}$	$34.7 \pm 2.5^{**}$	$35.5 \pm 2.5^{**}$	$35.3 \pm 2.6^{**}$	$35.0 \pm 2.6^{**}$
0.80	$\mathbf{38.4 \pm 3.2}$	$35.8 \pm 3.1^{**}$	$32.7 \pm 2.8^{**}$	$33.7 \pm 2.9^{**}$	$33.4 \pm 2.9^{**}$	$33.1 \pm 2.9^{**}$
0.86	$\mathbf{38.0 \pm 3.3}$	$35.0 \pm 3.1^{**}$	$31.6 \pm 3.0^{**}$	$32.7 \pm 3.0^{**}$	$32.3 \pm 3.0^{**}$	$31.9 \pm 3.0^{**}$
0.92	$\mathbf{37.2 \pm 3.4}$	$34.1 \pm 3.3^{**}$	$30.2 \pm 3.1^{**}$	$31.5 \pm 3.1^{**}$	$30.9 \pm 3.1^{**}$	$30.6 \pm 3.1^{**}$
0.96	$\mathbf{36.8 \pm 3.6}$	$33.4 \pm 3.4^{**}$	$29.1 \pm 3.2^{**}$	$30.4 \pm 3.3^{**}$	$29.9 \pm 3.3^{**}$	$29.5 \pm 3.3^{**}$
1.00	$\mathbf{36.3 \pm 3.5}$	$32.5 \pm 3.3^{**}$	$28.0 \pm 3.2^{**}$	$29.3 \pm 3.2^{**}$	$28.8 \pm 3.3^{**}$	$28.3 \pm 3.2^{**}$

Table 6. **Significance testing on TREC under increasing minority target shift.** We report mean accuracy (%) $\pm$ standard deviation across 20 independent experiments. The highest accuracy for each shift severity is marked in **bold**. Statistical significance against our proposed method (RSE) is determined via a paired t-test: * denotes $p < 0.05$ and ** denotes $p < 0.01$.

Shift (α)	RSE (Ours)	DRO	IRM	SRM	KL-RS	ERM
0.00	77.6 ± 1.7	77.5 ± 1.6	$\mathbf{78.6 \pm 1.6}^{**}$	$78.3 \pm 1.8^{**}$	$78.5 \pm 1.7^{**}$	$78.5 \pm 1.8^{**}$
0.32	$\mathbf{75.9 \pm 2.6}$	$75.2 \pm 2.4^{*}$	$74.9 \pm 2.5^{**}$	$75.1 \pm 2.4^{**}$	$74.8 \pm 2.4^{**}$	$74.9 \pm 2.5^{**}$
0.46	$\mathbf{75.0 \pm 3.0}$	$74.0 \pm 2.5^{**}$	$73.2 \pm 2.8^{**}$	$73.5 \pm 2.6^{**}$	$73.2 \pm 2.7^{**}$	$73.2 \pm 2.8^{**}$
0.57	$\mathbf{74.6 \pm 3.5}$	$73.4 \pm 3.2^{**}$	$72.1 \pm 3.5^{**}$	$72.6 \pm 3.2^{**}$	$72.1 \pm 3.5^{**}$	$72.2 \pm 3.5^{**}$
0.66	$\mathbf{74.3 \pm 3.7}$	$72.9 \pm 3.2^{**}$	$71.2 \pm 3.7^{**}$	$71.9 \pm 3.3^{**}$	$71.2 \pm 3.7^{**}$	$71.4 \pm 3.7^{**}$
0.74	$\mathbf{73.5 \pm 4.1}$	$71.8 \pm 3.6^{**}$	$69.8 \pm 3.9^{**}$	$70.5 \pm 3.6^{**}$	$69.9 \pm 3.9^{**}$	$70.1 \pm 3.9^{**}$
0.81	$\mathbf{73.4 \pm 4.2}$	$71.7 \pm 3.8^{**}$	$69.2 \pm 4.2^{**}$	$70.0 \pm 3.7^{**}$	$69.2 \pm 4.2^{**}$	$69.4 \pm 4.2^{**}$
0.87	$\mathbf{72.9 \pm 4.4}$	$71.1 \pm 3.9^{**}$	$68.7 \pm 4.4^{**}$	$69.6 \pm 3.8^{**}$	$68.8 \pm 4.3^{**}$	$68.9 \pm 4.4^{**}$
0.92	$\mathbf{72.9 \pm 4.5}$	$71.0 \pm 4.1^{**}$	$68.3 \pm 4.7^{**}$	$69.3 \pm 4.2^{**}$	$68.5 \pm 4.6^{**}$	$68.6 \pm 4.7^{**}$
0.96	$\mathbf{72.4 \pm 4.9}$	$70.6 \pm 4.4^{**}$	$67.7 \pm 4.8^{**}$	$68.6 \pm 4.2^{**}$	$67.8 \pm 4.6^{**}$	$68.0 \pm 4.7^{**}$
1.00	$\mathbf{71.8 \pm 5.0}$	$69.9 \pm 4.8^{**}$	$66.7 \pm 5.1^{**}$	$67.7 \pm 4.5^{**}$	$66.9 \pm 4.9^{**}$	$67.1 \pm 5.0^{**}$

B.2. Subpopulation Shift: CivilComments

In this subsection, we present the extended results for the CivilComments dataset under subpopulation shift. Because the dataset provides a single official test split, standard cross-trial variance cannot be computed directly. To rigorously evaluate the statistical significance of our results, we employ non-parametric bootstrap resampling ($B = 1000$ iterations) on the test set predictions.

Statistical Significance. Table 9 reports the bootstrapped Overall Accuracy and Worst-Group Accuracy (WGA). RSE achieves a statistically significant improvement ($p < 0.01$) over all optimization baselines in both metrics. Notably, RSE improves the critical WGA metric by nearly 2.0% absolute points compared to the strongest robust baseline (DRO). This confirms that RSE's fragility minimization effectively protects the most vulnerable demographic groups in the dataset without sacrificing average performance.

Detailed Subpopulation Breakdown. For granular transparency, Table 10 details the accuracy across the eight specific demographic groups. RSE consistently dominates or matches the best performance across almost all individual groups, demonstrating broad robustness across diverse subpopulations rather than over-optimizing for a single demographic.

B.3. Temporal Shift: HuffPost

In this subsection, we evaluate robustness against temporal distribution shifts using the HuffPost dataset. Models are trained on articles published between 2012 and 2015, and evaluated on subsequent years (2016–2018). As with the subpopulation shift analysis, we employ non-parametric bootstrap resampling ($B = 1000$ iterations) on the test predictions to rigorously

Table 7. **SST-5**: Detailed accuracy breakdown for all methods and individual models across shift magnitudes.

Method	**Shift Magnitude (α)**										
	0.00	**0.31**	**0.45**	**0.56**	**0.65**	**0.73**	**0.80**	**0.86**	**0.92**	**0.96**	**1.00**
Optimization Baselines											
RSE	48.0	44.2	**42.5**	**41.5**	**40.5**	**39.5**	**38.4**	**38.0**	**37.2**	**36.8**	**36.3**
DRO	50.0	44.5	42.0	40.2	38.9	37.2	35.8	35.0	34.1	33.4	32.5
IRM	**52.9**	45.2	41.6	39.2	37.1	34.7	32.7	31.6	30.2	29.1	28.0
SRM	52.2	45.1	41.7	39.5	37.7	35.5	33.7	32.7	31.5	30.4	29.3
KL-RS	52.8	**45.4**	41.9	39.6	37.5	35.3	33.4	32.3	30.9	29.9	28.8
ERM	**52.9**	45.3	41.7	39.4	37.3	35.0	33.1	31.9	30.6	29.5	28.3
Non-Optimization Baselines											
Uniform	52.1	43.8	39.8	37.2	34.8	32.4	30.2	28.9	27.4	26.2	25.1
Majority Vote	51.1	43.2	39.4	36.8	34.6	32.2	30.2	29.0	27.5	26.3	25.3
Acc. Weighted Model	52.7	43.9	39.8	37.1	34.6	32.1	29.9	28.5	26.9	25.7	24.5
Individual Models											
GPT-OSS-20b	48.8	44.6	42.6	41.3	40.2	38.9	37.8	37.2	36.4	35.9	35.2
Qwen3-4b	44.7	37.0	33.2	30.6	28.4	26.1	24.3	22.9	21.4	20.6	19.3
Falcon3-7b-It	48.5	41.3	37.9	35.7	33.6	31.7	30.2	29.0	27.4	26.5	25.6
Phi-4-15b	51.3	40.4	35.4	31.7	28.5	25.5	23.1	21.2	19.3	17.8	16.3
Mistral-7b-It	39.4	37.9	37.5	37.2	36.8	36.4	35.9	36.1	35.7	35.3	35.3
Gemma-3-12b	51.7	44.9	41.5	39.2	37.4	35.4	33.6	32.9	31.2	30.6	29.4
Mistral-Small-24b	52.9	44.8	41.3	38.5	36.3	34.0	31.9	30.7	29.2	28.0	26.9
Phi-3.5-mini-It	51.8	42.3	38.1	34.9	32.2	29.5	27.5	25.9	24.0	22.7	21.6

compute statistical significance over the fixed temporal split.

Statistical Significance. Table 11 presents the bootstrapped Overall Accuracy and Worst-Year Accuracy (WYA). RSE yields statistically significant improvements across both metrics. Notably, RSE maintains a definitive edge ($p < 0.05$) in worst-year performance even against strong invariant-learning baselines like IRM and SRM, while vastly outperforming DRO ($p < 0.01$). This confirms that explicitly optimizing for ensemble stability translates directly into resilience against chronological data drift.

Detailed Temporal Breakdown. For granular transparency, Table 12 details the accuracy for each specific year in the test phase. RSE shows consistent temporal resilience, achieving the best absolute performance in every single evaluation year (2016, 2017, and 2018) compared to all optimization baselines and individual constituent models.

B.4. Additional Aspiration Level Sensitivity Analysis

In this section, we extend the sensitivity analysis of the aspiration level τ to the TREC dataset to verify the generalizability of the observations made on SST-5.

B.4.1. TREC Dataset

Figure 6 illustrates the performance on TREC. Similar to the SST-5 results, we observe a trade-off governed by τ. Under severe shift (Figure 6a), increasing τ generally improves robustness. Conversely, in the absence of distribution shift (Figure 6b), aggressive constraints (higher τ) can lead to a slight degradation in in-distribution accuracy.

Table 8. **TREC**: Detailed accuracy breakdown for all methods and individual models across shift magnitudes.

Method	Shift Magnitude (α)										
	0.00	**0.32**	**0.46**	**0.57**	**0.66**	**0.74**	**0.81**	**0.87**	**0.92**	**0.96**	**1.00**
Optimization Baselines											
RSE	77.6	**75.9**	**75.0**	**74.6**	**74.3**	**73.5**	**73.4**	**72.9**	**72.9**	**72.4**	**71.8**
DRO	77.5	75.2	74.0	73.4	72.9	71.8	71.7	71.1	71.0	70.6	69.9
IRM	**78.6**	74.9	73.2	72.1	71.2	69.8	69.2	68.7	68.3	67.7	66.7
SRM	78.3	75.1	73.5	72.6	71.9	70.5	70.0	69.6	69.3	68.6	67.7
KL-RS	78.5	74.8	73.2	72.1	71.2	69.9	69.2	68.8	68.5	67.8	66.9
ERM	78.5	74.9	73.2	72.2	71.4	70.1	69.4	68.9	68.6	68.0	67.1
Non-Optimization Baselines											
Uniform	77.6	72.8	70.4	69.0	67.6	66.0	65.2	64.4	63.7	62.7	61.7
Majority Vote	76.6	71.8	69.3	68.1	66.6	65.0	64.2	63.4	62.7	61.7	60.8
Acc. Weighted Model	77.7	72.9	70.6	69.2	67.9	66.3	65.5	64.8	64.1	63.1	62.1
Individual Models											
GPT-OSS-20b	70.6	70.5	70.3	70.7	70.6	70.2	70.4	70.2	70.6	70.4	70.0
Qwen3-4b	69.8	63.5	60.5	59.0	56.9	55.3	54.3	53.3	52.1	51.3	50.3
Falcon3-7b-It	69.8	61.5	57.6	54.7	52.7	50.1	48.4	46.8	45.7	44.3	43.3
Phi-4-15b	77.2	74.1	72.5	71.9	70.9	69.6	69.3	69.0	68.5	67.7	67.2
Mistral-7b-It	62.0	54.6	51.1	48.9	46.6	45.0	43.5	42.1	40.7	39.8	38.8
Gemma-3-12b	67.4	63.2	61.1	60.0	58.9	57.0	56.5	55.9	55.2	54.3	53.8
Mistral-Small-24b	76.5	74.1	72.8	72.2	71.7	70.4	69.9	69.8	69.6	69.0	68.3
Phi-3.5-mini-It	71.2	59.6	54.1	51.1	47.2	44.4	42.5	39.8	38.3	36.2	34.9

B.5. Input Shift: Selective Copying

Real-world sequence-processing and retrieval-augmented systems often suffer from the "Lost in the Middle" phenomenon, where pre-trained models attend poorly to information positioned in the center of extended input contexts (Liu et al., 2024). To formalize this vulnerability against input shifts, we constructed a synthetic Selective Copying extraction task where models must identify a single target class hidden within a sequence of 500 noise tokens. We calibrated the ensemble on 500 input sequences where the target information was located at the *end*. We then evaluated the framework on a shifted test set of 500 sequences where the critical context was buried in the *exact middle*.

Model Selection for Selective Copying. Across all adversarial and natural distribution shift experiments, we utilize a fixed suite of 8 models (including Gemma-3-12B-Instruct). However, evaluating input shift via Selective Copying requires constituent models that actually fail to positional sequence bias. Gemma-3-12B-Instruct demonstrates an unusually high native resilience to extraction failures within 500-token noise sequences. Including it masks the aggregation dynamics we aim

Table 9. **Significance testing on CivilComments under subpopulation shift.** We report mean accuracy (%) $\pm$ standard deviation obtained via $B = 1000$ bootstrap resamples of the official test set. The highest accuracy for each metric is marked in **bold**. Statistical significance against our proposed method (RSE) is determined via a one-sided bootstrap hypothesis test: * denotes $p < 0.05$ and ** denotes $p < 0.01$. RSE significantly outperforms all baselines in both overall and worst-group performance.

Method	Overall Accuracy	Worst-Group Accuracy (WGA)
RSE (Ours)	**72.1 $\pm$ 0.1**	**59.7 $\pm$ 0.7**
DRO	70.6 $\pm$ 0.1**	57.7 $\pm$ 0.8**
IRM	69.8 $\pm$ 0.1**	56.5 $\pm$ 0.8**
SRM	68.8 $\pm$ 0.1**	55.1 $\pm$ 0.8**
KL-RS	69.6 $\pm$ 0.1**	56.3 $\pm$ 0.8**
ERM	70.1 $\pm$ 0.1**	56.8 $\pm$ 0.8**

Table 10. **CivilComments:** Detailed accuracy breakdown across subpopulation shifts for specific demographic groups.

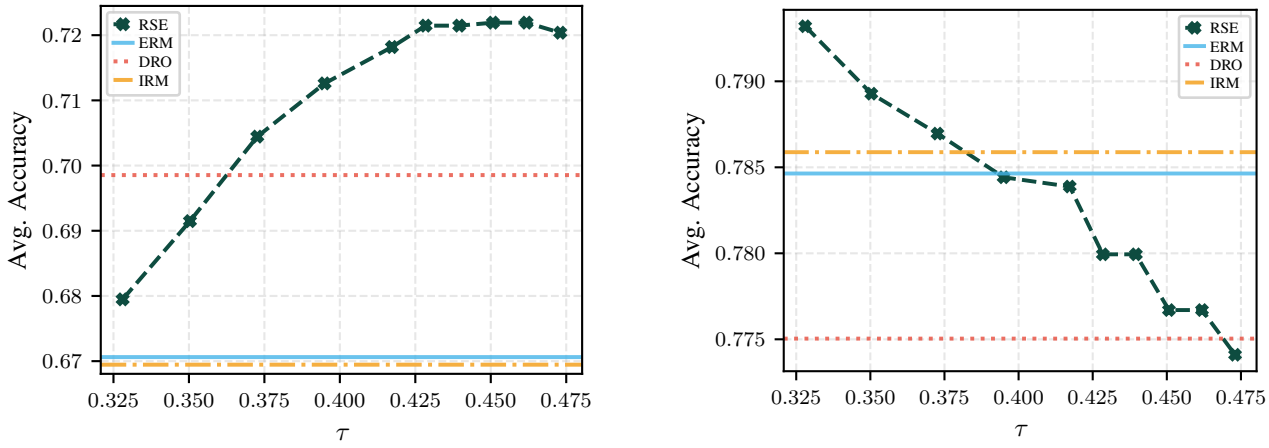
Method	Black	Christian	LGBTQ	Muslim	White	Female	Male	Other Rel.
Optimization Baselines								
RSE	**65.4**	**77.2**	**68.0**	60.8	**59.7**	**65.6**	**73.7**	**67.1**
DRO	64.3	76.2	66.3	59.1	57.8	64.2	72.2	64.8
IRM	63.5	75.4	65.5	57.8	56.5	63.1	71.5	63.8
SRM	62.8	74.3	64.2	56.9	55.1	62.1	70.5	61.8
KL-RS	63.3	75.1	65.0	57.8	56.3	62.9	71.2	63.5
ERM	63.8	75.7	65.7	58.3	56.8	63.5	71.7	64.1
Individual Models								
Falcon3-7b-It	63.9	75.1	64.9	57.7	56.0	64.5	71.5	61.1
Gemma3-12b-It	59.0	71.2	61.7	52.6	50.9	59.0	66.0	55.3
GPT-OSS-20b	47.6	55.2	47.6	38.7	39.5	45.7	57.0	37.9
Mistral-24b	64.4	76.2	65.7	61.2	57.8	64.2	71.0	67.0
Mistral-7b	50.3	59.5	51.2	42.9	42.0	47.1	55.9	46.8
Phi-3.5-Mini	65.2	75.5	66.8	**62.0**	58.6	65.2	73.3	67.0
Phi-4	**65.4**	75.1	66.0	57.7	**59.7**	64.8	72.6	64.0
Qwen3-4b	61.7	69.4	63.1	58.1	54.4	60.6	64.8	62.7

Table 11. **Significance testing on HuffPost under temporal shift.** We report mean accuracy (%) $\pm$ standard deviation obtained via $B = 1000$ bootstrap resamples of the test set (2016–2018). The highest accuracy for each metric is marked in **bold**. Statistical significance against our proposed method (RSE) is determined via a one-sided bootstrap hypothesis test: * denotes $p < 0.05$ and ** denotes $p < 0.01$. RSE significantly outperforms all baselines across both Average Yearly Accuracy (AYA) and Worst-Year Accuracy (WYA) metrics.

Method	Average Yearly Accuracy (AYA)	Worst-Year Accuracy (WYA)
RSE (Ours)	**78.8 $\pm$ 0.3**	**77.2 $\pm$ 0.4**
DRO	$76.3 \pm 0.3^{**}$	$75.5 \pm 0.4^{**}$
IRM	$78.4 \pm 0.3^{**}$	$76.9 \pm 0.4^{*}$
SRM	$78.4 \pm 0.3^{**}$	$76.9 \pm 0.4^{*}$
KL-RS	$78.1 \pm 0.3^{**}$	$76.6 \pm 0.4^{**}$
ERM	$78.1 \pm 0.3^{**}$	$76.7 \pm 0.4^{**}$

Table 12. **HuffPost:** Detailed accuracy breakdown across temporal shifts (2016–2018) for models trained on 2012–2015 data.

Method	2016	2017	2018
Optimization Baselines			
RSE	**77.2**	**78.1**	**81.0**
DRO	76.1	75.6	77.2
IRM	77.0	77.5	80.6
SRM	77.0	77.7	80.5
KL-RS	76.7	77.3	80.4
ERM	76.8	77.3	80.3
Individual Models			
Falcon3-7b-It	69.2	66.0	70.5
Gemma3-12b-It	75.0	74.4	76.8
GPT-OSS-20b	71.6	72.1	75.4
Mistral-24b	76.9	77.7	80.0
Mistral-7b-It	70.2	67.0	69.1
Phi-3.5-Mini-It	71.5	70.9	74.7
Phi-4-15b	71.6	69.7	71.9
Qwen3-4b	72.4	74.1	76.0



(a) Severe shift ($\alpha = 1$)

(b) No shift ($\alpha = 0$)

Figure 6. Sensitivity of RSE to aspiration level τ on the **TREC** dataset. (a) Performance under severe shift. (b) Performance under no shift.

to isolate, as the task would merely measure the performance of a single natively robust model. To construct a rigorous testbed where constituent models natively suffer from middle-context extraction failures, we substituted Gemma-3-12B-Instruct with Llama-3.1-8B-Instruct.

Results. RSE outperformed ERM by **+7.0%** and DRO by **+5.0%**, effectively mitigating position-induced extraction failures. Non-optimization baselines collapse because they cannot selectively downweight the majority of individual models that exploit the positional bias present in the calibration set.

Statistical Significance. Because the evaluation relies on a single fixed, shifted test set, we utilize non-parametric bootstrap resampling ($B = 1000$ iterations) to estimate variance and statistical significance. Table 13 details the bootstrapped accuracy across all baselines.

Table 13. **Significance testing on the Selective Copying task under input shift.** We report mean accuracy (%) $\pm$ standard deviation obtained via $B = 1000$ bootstrap resamples of the shifted test set. The highest accuracy is marked in **bold**. Statistical significance against our proposed method (RSE) is determined via a one-sided bootstrap hypothesis test: * denotes $p < 0.05$ and ** denotes $p < 0.01$. RSE significantly outperforms both individual models and all ensemble baselines.

Method	Overall Accuracy (%)	Significance
RSE (Ours)	**86.7 ± 1.5**	–
Optimization Baselines		
DRO	81.7 ± 1.7	**
KL-RS	80.1 ± 1.7	**
ERM	80.1 ± 1.7	**
SRM	79.3 ± 1.8	**
IRM	78.1 ± 1.8	**
Non-Optimization Baselines & Models		
Best Individual	79.3 ± 1.8	**
Acc. Weighted	55.1 ± 2.2	**
Uniform	48.6 ± 2.3	**
Majority Vote	40.1 ± 2.2	**

C. Derivation of the Quadratic Risk Form

In this section we show that for each class $y_j \in \mathcal{Y}$, the matrix $\mathbf{A}_j \in \mathbb{R}^{M \times M}$ defined in Section 4.1 is positive semidefinite.

C.1. Explicit definition of the matrix $\mathbf{A}_j$

Fix a class index $j \in \{1, \dots, K\}$. Let X be a random input drawn from the class-conditional distribution $P(X \mid Y = y_j)$. For each model $m \in \{1, \dots, M\}$, the model output is a probability vector

$$\mathbf{p}_m(X) = \begin{bmatrix} p_{m,1}(X) \\ \vdots \\ p_{m,K}(X) \end{bmatrix} \in \mathbb{R}^K, \qquad p_{m,k}(X) \geq 0, \quad \sum_{k=1}^{K} p_{m,k}(X) = 1.$$

The matrix $\mathbf{A}_j$ is defined entrywise as

$$(\mathbf{A}_j)_{ab} = \mathbb{E}_{X|Y=y_j}\big[\mathbf{p}_a(X)^\top \mathbf{p}_b(X)\big], \qquad a, b \in \{1, \dots, M\}.$$

We assume $\mathbb{E}_{X|Y=y_j}\big[\|\mathbf{p}_m(X)\|_2^2\big] < \infty$ for all m, which holds automatically here since $\mathbf{p}_m(X)$ has nonnegative entries and $\sum_{k=1}^{K} p_{m,k}(X) = 1$, hence $\|\mathbf{p}_m(X)\|_2^2 = \sum_{k=1}^{K} p_{m,k}(X)^2 \leq \sum_{k=1}^{K} p_{m,k}(X) = 1$.

Expanding the dot product coordinate-wise yields

$$\mathbf{p}_a(X)^\top \mathbf{p}_b(X) = \sum_{k=1}^{K} p_{a,k}(X)\, p_{b,k}(X),$$

and therefore

$$(\mathbf{A}_j)_{ab} = \sum_{k=1}^{K} \mathbb{E}_{X|Y=y_j}\big[p_{a,k}(X)\, p_{b,k}(X)\big].$$

Thus $\mathbf{A}_j$ is the $M \times M$ matrix

$$\mathbf{A}_j = \begin{bmatrix} \mathbb{E}_{X|Y=y_j}[\mathbf{p}_1(X)^\top \mathbf{p}_1(X)] & \cdots & \mathbb{E}_{X|Y=y_j}[\mathbf{p}_1(X)^\top \mathbf{p}_M(X)] \\ \vdots & \ddots & \vdots \\ \mathbb{E}_{X|Y=y_j}[\mathbf{p}_M(X)^\top \mathbf{p}_1(X)] & \cdots & \mathbb{E}_{X|Y=y_j}[\mathbf{p}_M(X)^\top \mathbf{p}_M(X)] \end{bmatrix}.$$

Moreover, $\mathbf{A}_j$ is symmetric because for any a, b,

$$(\mathbf{A}_j)_{ab} = \mathbb{E}_{X|Y=y_j}[\mathbf{p}_a(X)^\top \mathbf{p}_b(X)] = \mathbb{E}_{X|Y=y_j}[\mathbf{p}_b(X)^\top \mathbf{p}_a(X)] = (\mathbf{A}_j)_{ba},$$

where we used $\mathbf{p}_a(X)^\top \mathbf{p}_b(X) = \mathbf{p}_b(X)^\top \mathbf{p}_a(X)$ for real vectors.

C.2. Positive semidefiniteness via the quadratic form

Recall that a symmetric matrix $\mathbf{H} \in \mathbb{R}^{M \times M}$ is PSD if $\mathbf{v}^\top \mathbf{H} \mathbf{v} \geq 0$ for all $\mathbf{v} \in \mathbb{R}^M$.

Let $\mathbf{v} = (v_1, \dots, v_M)^\top \in \mathbb{R}^M$ be an arbitrary vector in $\mathbb{R}^M$.

$$\begin{aligned} \mathbf{v}^\top \mathbf{A}_j \mathbf{v} &= \sum_{a=1}^{M} \sum_{b=1}^{M} v_a (\mathbf{A}_j)_{ab} v_b \\ &= \sum_{a=1}^{M} \sum_{b=1}^{M} v_a v_b\, \mathbb{E}_{X|Y=y_j}\big[\mathbf{p}_a(X)^\top \mathbf{p}_b(X)\big]. \end{aligned}$$

Since the sums over a and b are finite, we may interchange the finite summations and the expectation. Therefore,

$$\mathbf{v}^\top \mathbf{A}_j \mathbf{v} = \mathbb{E}_{X|Y=y_j}\Big[\sum_{a=1}^{M} \sum_{b=1}^{M} v_a v_b\, \mathbf{p}_a(X)^\top \mathbf{p}_b(X)\Big].$$

Define, for each realization X, the vector

$$\mathbf{u}(X;\mathbf{v}) := \sum_{m=1}^{M} v_m \mathbf{p}_m(X) \in \mathbb{R}^K.$$

Expanding $\mathbf{u}(X;\mathbf{v})^\top \mathbf{u}(X;\mathbf{v})$ gives

$$\mathbf{u}(X;\mathbf{v})^\top \mathbf{u}(X;\mathbf{v}) = \sum_{a=1}^{M}\sum_{b=1}^{M} v_a v_b \, \mathbf{p}_a(X)^\top \mathbf{p}_b(X).$$

Substituting this identity yields

$$\mathbf{v}^\top \mathbf{A}_j \mathbf{v} = \mathbb{E}_{X|Y=y_j}\big[\mathbf{u}(X;\mathbf{v})^\top \mathbf{u}(X;\mathbf{v})\big] = \mathbb{E}_{X|Y=y_j}\big[\|\mathbf{u}(X;\mathbf{v})\|_2^2\big].$$

Writing the squared norm coordinate-wise,

$$\|\mathbf{u}(X;\mathbf{v})\|_2^2 = \sum_{k=1}^{K}\Big(\sum_{m=1}^{M} v_m p_{m,k}(X)\Big)^2 \geq 0 \quad \text{for every realization of } X \text{ and every } \mathbf{v} \in \mathbb{R}^M.$$

Since the integrand is nonnegative for every X, its expectation is also nonnegative. Therefore

$$\mathbf{v}^\top \mathbf{A}_j \mathbf{v} \geq 0 \quad \text{for all } \mathbf{v} \in \mathbb{R}^M,$$

which proves that $\mathbf{A}_j$ is positive semidefinite.

C.3. Finite-sample estimator and preservation of positive semidefiniteness

Fix a class index $j \in \{1,\dots,K\}$. Let

$$\mathcal{D}_j := \{x^{(i)} : y^{(i)} = y_j\}$$

denote the subset of the training data consisting of all samples whose label equals y_j, and let $|\mathcal{D}_j|$ denote its cardinality.

For each input $x \in \mathcal{D}_j$ and each model $m \in \{1,\dots,M\}$, recall that the model produces a probability vector

$$\mathbf{p}_m(x) = \begin{bmatrix} p_{m,1}(x) \\ \vdots \\ p_{m,K}(x) \end{bmatrix} \in \mathbb{R}^K, \qquad p_{m,k}(x) \geq 0, \quad \sum_{k=1}^{K} p_{m,k}(x) = 1.$$

Using these vectors, we define the matrix-valued function

$$\mathbf{P}(x) := \big[\mathbf{p}_1(x) \quad \mathbf{p}_2(x) \quad \cdots \quad \mathbf{p}_M(x)\big] \in \mathbb{R}^{K\times M}.$$

That is, the m-th column of $\mathbf{P}(x)$ is the probability vector $\mathbf{p}_m(x)$ produced by model m at input x.

The empirical (finite-sample) estimator of $\mathbf{A}_j$ is then defined as

$$\widehat{\mathbf{A}}_j = \frac{1}{|\mathcal{D}_j|} \sum_{x^{(i)} \in \mathcal{D}_j} \mathbf{P}(x^{(i)})^\top \mathbf{P}(x^{(i)}).$$

First, fix an arbitrary input $x \in \mathcal{D}_j$. Consider the matrix $\mathbf{P}(x)^\top \mathbf{P}(x) \in \mathbb{R}^{M\times M}$. To verify that this matrix is positive semidefinite, we must show that

$$\mathbf{v}^\top \mathbf{P}(x)^\top \mathbf{P}(x) \mathbf{v} \geq 0 \quad \text{for all vectors } \mathbf{v} \in \mathbb{R}^M.$$

Let $\mathbf{v} = (v_1, \ldots, v_M)^\top \in \mathbb{R}^M$ be arbitrary. We compute the quadratic form explicitly:

$$\mathbf{v}^\top \mathbf{P}(x)^\top \mathbf{P}(x)\mathbf{v} = (\mathbf{P}(x)\mathbf{v})^\top (\mathbf{P}(x)\mathbf{v}).$$

Next, we write $\mathbf{P}(x)\mathbf{v}$ explicitly.

$$\mathbf{P}(x)\mathbf{v} = \sum_{m=1}^{M} v_m \mathbf{p}_m(x) = \begin{bmatrix} \sum_{m=1}^{M} v_m p_{m,1}(x) \\ \vdots \\ \sum_{m=1}^{M} v_m p_{m,K}(x) \end{bmatrix} \in \mathbb{R}^K.$$

Substituting this expression back into the quadratic form yields

$$(\mathbf{P}(x)\mathbf{v})^\top (\mathbf{P}(x)\mathbf{v}) = \sum_{k=1}^{K} \Big(\sum_{m=1}^{M} v_m p_{m,k}(x) \Big)^2.$$

Each term in the sum is a square of a real number and is therefore nonnegative. Hence

$$(\mathbf{P}(x)\mathbf{v})^\top (\mathbf{P}(x)\mathbf{v}) \geq 0 \quad \text{for all } \mathbf{v} \in \mathbb{R}^M,$$

which implies

$$\mathbf{v}^\top \mathbf{P}(x)^\top \mathbf{P}(x)\mathbf{v} \geq 0 \quad \text{for all } \mathbf{v} \in \mathbb{R}^M.$$

By definition, this shows that the matrix $\mathbf{P}(x)^\top \mathbf{P}(x)$ is positive semidefinite.

We now consider the empirical estimator $\widehat{\mathbf{A}}_j$. For any $\mathbf{v} \in \mathbb{R}^M$, we have

$$\begin{aligned} \mathbf{v}^\top \widehat{\mathbf{A}}_j \mathbf{v} &= \mathbf{v}^\top \left(\frac{1}{|\mathcal{D}_j|} \sum_{x^{(i)} \in \mathcal{D}_j} \mathbf{P}(x^{(i)})^\top \mathbf{P}(x^{(i)}) \right) \mathbf{v} \\ &= \frac{1}{|\mathcal{D}_j|} \sum_{x^{(i)} \in \mathcal{D}_j} \mathbf{v}^\top \mathbf{P}(x^{(i)})^\top \mathbf{P}(x^{(i)}) \mathbf{v}. \end{aligned}$$

As shown above, each term $\mathbf{v}^\top \mathbf{P}(x^{(i)})^\top \mathbf{P}(x^{(i)})\mathbf{v}$ is nonnegative. Since $|\mathcal{D}_j|^{-1} > 0$, the weighted sum of these nonnegative terms is also nonnegative. Therefore,

$$\mathbf{v}^\top \widehat{\mathbf{A}}_j \mathbf{v} \geq 0 \quad \text{for all } \mathbf{v} \in \mathbb{R}^M.$$

With this, we conclude that the matrix $\widehat{\mathbf{A}}_j$ is positive semidefinite.

D. Derivation of $\Phi(\kappa; \mathbf{w})$

In this section, we derive the inner objective $\Phi(\kappa; \mathbf{w})$ starting from the robust satisficing definition in Definition 3.1.

We fix a reference label distribution $\mathbf{q}_{\text{ref}} \in \Delta_K$ and an aspiration level $\tau > 0$. Throughout this appendix, we assume that

$$(q_{\text{ref}})_j > 0 \quad \text{for all } j \in \{1, \ldots, K\},$$

which ensures that the Kullback–Leibler divergence $D_{\text{KL}}(\mathbf{q} \,\|\, \mathbf{q}_{\text{ref}})$ is finite for all $\mathbf{q} \in \Delta_K$.

Recall Definition 3.1. For a fixed ensemble weight vector $\mathbf{w} \in \mathcal{W}_\tau$, the fragility $\kappa_\tau(\mathbf{w})$ is defined as the smallest nonnegative scalar κ such that

$$R(\mathbf{w}, \mathbf{q}) \leq \tau + \kappa\, \Delta(\mathbf{q}, \mathbf{q}_{\text{ref}}) \qquad \text{for all } \mathbf{q} \in \Delta_K. \tag{21}$$

To analyze feasibility for a *candidate* value $\kappa \geq 0$, we rewrite (21) equivalently as

$$R(\mathbf{w}, \mathbf{q}) - \kappa\, \Delta(\mathbf{q}, \mathbf{q}_{\text{ref}}) \leq \tau \qquad \text{for all } \mathbf{q} \in \Delta_K. \tag{22}$$

The inequality (22) holds if and only if it holds for the worst-case distribution. Equivalently, for a given $\kappa \geq 0$,

$$\sup_{\mathbf{q}\in\Delta_K} \Big\{ R(\mathbf{w},\mathbf{q}) - \kappa\,\Delta(\mathbf{q},\mathbf{q}_{\mathrm{ref}}) \Big\} \leq \tau. \tag{23}$$

This motivates defining a scalar-valued objective $\Phi(\kappa;\mathbf{w})$ that summarizes the worst-case left-hand side. For any $\kappa \geq 0$, define

$$\Phi(\kappa;\mathbf{w}) := \sup_{\mathbf{q}\in\Delta_K} \Big\{ R(\mathbf{w},\mathbf{q}) - \kappa\,\Delta(\mathbf{q},\mathbf{q}_{\mathrm{ref}}) \Big\}. \tag{24}$$

With this definition, a candidate κ satisfies the robust satisficing condition if and only if

$$\Phi(\kappa;\mathbf{w}) \leq \tau. \tag{25}$$

Since $\kappa \mapsto \Phi(\kappa;\mathbf{w})$ is nonincreasing, the feasible set $\{\kappa \geq 0 : \ \Phi(\kappa;\mathbf{w}) \leq \tau\}$ is an interval of the form $[\kappa_\tau(\mathbf{w}),\infty)$ whenever $\kappa_\tau(\mathbf{w}) \in (0,\infty)$.

Moreover, the fragility $\kappa_\tau(\mathbf{w})$ is precisely the smallest $\kappa \geq 0$ for which (25) holds.

We now make explicit the dependence of the risk $R(\mathbf{w},\mathbf{q})$ on the label distribution $\mathbf{q}$. For each class y_j, recall the class-conditional risk (cf. (2))

$$r(\mathbf{w},y_j) := \mathbb{E}_{X\sim P(X|Y=y_j)}\Big[L\big(y_j,\widehat{\mathbf{p}}_{\mathbf{w}}(X)\big)\Big].$$

For a label distribution $\mathbf{q} \in \Delta_K$, the corresponding mixture risk is

$$R(\mathbf{w},\mathbf{q}) = \sum_{j=1}^{K} q_j\, r(\mathbf{w},y_j) = \mathbf{q}^\top \mathbf{r}(\mathbf{w}), \tag{26}$$

where $\mathbf{r}(\mathbf{w}) \in \mathbb{R}^K$ denotes the vector of class-conditional risks defined in (3).

We now instantiate the Δ with the Kullback–Leibler divergence:

$$\Delta(\mathbf{q},\mathbf{q}_{\mathrm{ref}}) = D_{\mathrm{KL}}(\mathbf{q}\,\|\,\mathbf{q}_{\mathrm{ref}}) = \sum_{j=1}^{K} q_j \log \frac{q_j}{(q_{\mathrm{ref}})_j}. \tag{27}$$

Substituting (26) and (27) into (24) yields the KL-based inner problem

$$\Phi(\kappa;\mathbf{w}) = \sup_{\mathbf{q}\in\Delta_K} \Big\{ \mathbf{q}^\top \mathbf{r}(\mathbf{w}) - \kappa\, D_{\mathrm{KL}}(\mathbf{q}\,\|\,\mathbf{q}_{\mathrm{ref}}) \Big\}. \tag{28}$$

Fix $\mathbf{w} \in \mathcal{W}_\tau$. For any $\kappa_1 \geq \kappa_2 \geq 0$ and any $\mathbf{q} \in \Delta_K$, since $D_{\mathrm{KL}}(\mathbf{q}\,\|\,\mathbf{q}_{\mathrm{ref}}) \geq 0$, we have

$$\mathbf{q}^\top \mathbf{r}(\mathbf{w}) - \kappa_1 D_{\mathrm{KL}}(\mathbf{q}\,\|\,\mathbf{q}_{\mathrm{ref}}) \leq \mathbf{q}^\top \mathbf{r}(\mathbf{w}) - \kappa_2 D_{\mathrm{KL}}(\mathbf{q}\,\|\,\mathbf{q}_{\mathrm{ref}})$$

Taking the supremum over $\mathbf{q} \in \Delta_K$ on both sides and using the definition in (28) yields

$$\Phi(\kappa_1;\mathbf{w}) \leq \Phi(\kappa_2;\mathbf{w}),$$

which shows that the function $\kappa \mapsto \Phi(\kappa;\mathbf{w})$ is nonincreasing on $[0,\infty)$.

We next derive explicit upper and lower bounds on $\Phi(\kappa;\mathbf{w})$. At $\kappa = 0$, the KL term vanishes, and

$$\Phi(0;\mathbf{w}) = \sup_{\mathbf{q}\in\Delta_K} \mathbf{q}^\top \mathbf{r}(\mathbf{w}) = \max_{j\in\{1,\ldots,K\}} r(\mathbf{w},y_j),$$

where the last equality follows because a linear function over the simplex Δ_K attains its maximum at an extreme point.

For a lower bound, note that $\mathbf{q}_{\mathrm{ref}} \in \Delta_K$ is feasible in (28). Substituting $\mathbf{q} = \mathbf{q}_{\mathrm{ref}}$ and using $D_{\mathrm{KL}}(\mathbf{q}_{\mathrm{ref}}\,\|\,\mathbf{q}_{\mathrm{ref}}) = 0$ gives, for all $\kappa \geq 0$,

$$\Phi(\kappa;\mathbf{w}) \geq \mathbf{q}_{\mathrm{ref}}^\top \mathbf{r}(\mathbf{w}) = R(\mathbf{w},\mathbf{q}_{\mathrm{ref}}).$$

Combining the monotonicity of $\Phi(\kappa; \mathbf{w})$ with the value at $\kappa = 0$ also yields the uniform upper bound

$$\Phi(\kappa; \mathbf{w}) \le \Phi(0; \mathbf{w}) = \max_{j \in \{1,\dots,K\}} r(\mathbf{w}, y_j), \qquad \forall \kappa \ge 0.$$

The preceding bounds imply that the equation $\Phi(\kappa; \mathbf{w}) = \tau$ can only be satisfied if

$$R(\mathbf{w}, \mathbf{q}_{\text{ref}}) < \tau < \max_j r(\mathbf{w}, y_j).$$

If $\tau \ge \max_j r(\mathbf{w}, y_j)$, then $\Phi(0; \mathbf{w}) \le \tau$ and the smallest feasible value is $\kappa = 0$. If $\tau \le R(\mathbf{w}, \mathbf{q}_{\text{ref}})$, then the inequality $\Phi(\kappa; \mathbf{w}) \le \tau$ fails for all $\kappa \ge 0$. Otherwise, the robust satisficing constraint becomes active at some $\kappa > 0$, which we denote by $\kappa_\tau(\mathbf{w})$, satisfying

$$\Phi\big(\kappa_\tau(\mathbf{w}); \mathbf{w}\big) = \tau. \tag{29}$$

Moreover, the equation $\Phi(\kappa; \mathbf{w}) = \tau$ admits a unique finite root $\kappa \in (0, \infty)$ precisely in the interior regime $R(\mathbf{w}, \mathbf{q}_{\text{ref}}) < \tau < \max_j r(\mathbf{w}, y_j)$ (Appendix F). The explicit solution of the maximization problem in (28), as well as the closed-form expression for $\Phi(\kappa; \mathbf{w})$, are derived in Appendix E.

E. KL Inner Problem and Exponential Tilt

For fixed $\mathbf{w}$ we abbreviate $r_j = r(\mathbf{w}, y_j)$ and consider the penalized inner problem, which is a standard KL-regularized variational optimization; the corresponding exponential-tilting optimizer appears, for example, as Proposition A.5 in Yan et al. (2024) for the uniform reference case. We include the derivation here to match our notation and allow a general $\mathbf{q}_{\text{ref}}$.

$$\Phi(\kappa; \mathbf{w}) = \sup_{\mathbf{q} \in \Delta_K} \left\{ \sum_{j=1}^K q_j r_j - \kappa \sum_{j=1}^K q_j \log \frac{q_j}{(q_{\text{ref}})_j} \right\}.$$

If all $(q_{\text{ref}})_j > 0$, then the objective is strictly concave on the open simplex and the unique maximizer lies in the interior. Hence the stationarity KKT conditions characterize the solution.

Introducing a Lagrange multiplier λ for $\sum_j q_j = 1$, the Lagrangian is

$$\mathcal{L}(\mathbf{q}, \lambda) = \sum_{j=1}^K q_j r_j - \kappa \sum_{j=1}^K q_j \log \frac{q_j}{(q_{\text{ref}})_j} + \lambda \Big(\sum_{j=1}^K q_j - 1 \Big).$$

Using $\frac{\partial}{\partial q_j}(q_j \log q_j) = \log q_j + 1$, the stationarity condition is

$$r_j - \kappa \Big(\log \tfrac{q_j}{(q_{\text{ref}})_j} + 1 \Big) + \lambda = 0.$$

Rearranging and exponentiating yields $q_j \propto (q_{\text{ref}})_j \exp(r_j/\kappa)$. Enforcing $\sum_j q_j = 1$ determines the unique normalizing constant, giving the exponentially tilted distribution

$$(q^\star)_j = \frac{(q_{\text{ref}})_j \exp(r_j/\kappa)}{\sum_{k=1}^K (q_{\text{ref}})_k \exp(r_k/\kappa)},$$

which matches (13). Substituting $\mathbf{q}^\star$ into the objective gives

$$\Phi(\kappa; \mathbf{w}) = \kappa \log \Big(\sum_{j=1}^K (q_{\text{ref}})_j \exp(r_j/\kappa) \Big),$$

which is (14).

F. Solving for $\kappa_\tau(\mathbf{w})$

Recall from Lemma 4.2 that for fixed $\mathbf{w}$ the KL fragility is characterized by the root equation (15). There are three cases: (i) if $\tau \geq \max_j r(\mathbf{w}, y_j)$ then $\kappa_\tau(\mathbf{w}) = 0$; (ii) if $R(\mathbf{w}, \mathbf{q}_{\text{ref}}) \geq \tau$ then $\kappa_\tau(\mathbf{w}) = +\infty$; (iii) if

$$R(\mathbf{w}, \mathbf{q}_{\text{ref}}) < \tau < \max_j r(\mathbf{w}, y_j),$$

then there is a unique $\kappa_\tau(\mathbf{w}) \in (0, \infty)$ satisfying (15). We focus on case (iii) below.

We now (i) establish existence and uniqueness, and (ii) derive an analytic series representation of $\kappa_\tau(\mathbf{w})$ via Lagrange–Bürmann inversion.

F.1. Existence and uniqueness

For brevity, fix $\mathbf{w}$ and set

$$\alpha_j := r(\mathbf{w}, y_j), \qquad X \sim \sum_{j=1}^{K} (q_{\text{ref}})_j \, \delta_{\alpha_j}.$$

Introduce

$$S := \frac{1}{\kappa} > 0, \qquad G(S) := \sum_{j=1}^{K} (q_{\text{ref}})_j e^{\alpha_j S}. \tag{30}$$

Then (15) becomes

$$\tau = f(S) := \frac{1}{S} \log G(S). \tag{31}$$

Since $G(0) = \sum_j (q_{\text{ref}})_j = 1$ and G is analytic near $S = 0$, $\log G(S)$ admits a Taylor expansion of the form

$$\log G(S) = \sum_{m=1}^{\infty} a_m S^m, \qquad a_m = \frac{1}{m!} \operatorname{cum}_m(X),$$

where $\operatorname{cum}_m(X)$ are the cumulants of X. Dividing by S gives

$$f(S) = a_1 + a_2 S + a_3 S^2 + a_4 S^3 + \cdots, \tag{32}$$

with

$$a_1 = \mathbb{E}[X] = R(\mathbf{w}, \mathbf{q}_{\text{ref}}), \qquad a_2 = \tfrac{1}{2}\operatorname{Var}(X).$$

We assume $\operatorname{Var}(X) > 0$ (i.e., the class-wise risks are non-degenerate under $\mathbf{q}_{\text{ref}}$); otherwise $R(\mathbf{w}, \mathbf{q}_{\text{ref}}) = \max_j r(\mathbf{w}, y_j)$ and the interval $\big(R(\mathbf{w}, \mathbf{q}_{\text{ref}}), \max_j r(\mathbf{w}, y_j)\big)$ collapses, making $\kappa_\tau(\mathbf{w})$ trivial.

Standard properties of log-moment generating functions imply:

- As $S \to 0^+$ (equivalently, $\kappa \to \infty$),
$$f(S) \longrightarrow f(0) = R(\mathbf{w}, \mathbf{q}_{\text{ref}}).$$
- As $S \to \infty$ (equivalently, $\kappa \to 0^+$), the largest class-wise risk dominates:
$$f(S) = \frac{1}{S} \log G(S) \longrightarrow \max_j \alpha_j = \max_j r(\mathbf{w}, y_j).$$

Moreover, f is strictly increasing on $(0, \infty)$. Let $g(S) := \log G(S)$. Since g is convex and $g(0) = 0$, the map $S \mapsto g(S)/S$ is nondecreasing on $(0, \infty)$, hence

$$f'(S) = \frac{Sg'(S) - g(S)}{S^2} \geq 0 \quad \text{for all } S > 0,$$

and under $\mathrm{Var}(X) > 0$ (non-degenerate risks), g is strictly convex, so $f'(S) > 0$ for all $S > 0$. Therefore f maps $(0, \infty)$ continuously and strictly monotonically onto the interval $\big(R(\mathbf{w}, \mathbf{q}_{\mathrm{ref}}),\ \max_j r(\mathbf{w}, y_j)\big)$. For any

$$\tau \in \big(R(\mathbf{w}, \mathbf{q}_{\mathrm{ref}}),\ \max_j r(\mathbf{w}, y_j)\big)$$

there is a unique $S > 0$ solving (31), and thus a unique $\kappa_\tau(\mathbf{w}) = 1/S$. Numerically, this scalar root can be recovered by standard one-dimensional solvers (bisection, Newton, and so on).

F.2. Analytic inversion via Lagrange–Bürmann

We derive an analytic series for the inverse mapping $\tau \mapsto S(\tau)$ in a neighborhood of the reference point

$$\tau_0 := f(0) = R(\mathbf{w}, \mathbf{q}_{\mathrm{ref}}).$$

Expanding (32) gives

$$f(S) = \tau_0 + a_2 S + a_3 S^2 + a_4 S^3 + \cdots, \qquad a_2 > 0.$$

Since f is analytic and $f'(0) = a_2 \neq 0$, it admits a local analytic inverse near τ_0, which we write as

$$S(\tau) = \sum_{n=1}^{\infty} c_n\,(\tau - \tau_0)^n. \tag{33}$$

Introduce

$$H(S) := \frac{S}{f(S) - \tau_0}. \tag{34}$$

By construction,

$$f(S) - \tau_0 = a_2 S + a_3 S^2 + \cdots,$$

so $H(S)$ is analytic at $S = 0$ with

$$H(S) = \frac{1}{a_2} \cdot \frac{1}{1 + \frac{a_3}{a_2} S + O(S^2)}$$

Applying the geometric series expansion $\frac{1}{1+x} = 1 - x + x^2 - x^3 + \cdots$ with $x = \frac{a_3}{a_2} S + O(S^2)$:

$$H(S) = \frac{1}{a_2}\left(1 - \left(\frac{a_3}{a_2} S + O(S^2)\right) + O(S^2)\right)$$

$$H(S) = \frac{1}{a_2}\left(1 - \frac{a_3}{a_2} S + O(S^2)\right), \qquad H(0) = \frac{1}{a_2}.$$

The Lagrange–Bürmann inversion theorem yields, for $n \geq 1$,

$$c_n = \frac{1}{n!}\left[\frac{d^{n-1}}{dS^{n-1}}\Big(H(S)^n\Big)\right]_{S=0}. \tag{35}$$

This provides a closed-form expression for each coefficient c_n in terms of the derivatives of H at the origin, hence in terms of the cumulants of X. For illustration, the first coefficients are

$$c_1 = \frac{1}{a_2}, \qquad c_2 = -\frac{a_3}{a_2^3}, \qquad c_3 = \frac{2a_3^2 - a_2 a_4}{a_2^5}. \tag{36}$$

F.3. Laurent series for $\kappa_\tau(\mathbf{w})$

We are ultimately interested in

$$\kappa_\tau(\mathbf{w}) = \frac{1}{S(\tau)}.$$

Setting

$$z := \tau - \tau_0, \qquad S(z) = \sum_{n=1}^{\infty} c_n z^n,$$

we have $c_1 > 0$ and $S(z)$ is analytic with $S(0) = 0$ and nonzero derivative at the origin. Consequently, $\kappa_\tau(\mathbf{w})$ admits a Laurent expansion with a simple pole at $z = 0$,

$$\kappa_\tau(\mathbf{w}) = \frac{1}{S(z)} = \sum_{m=-1}^{\infty} d_m z^m, \qquad z = \tau - \tau_0, \tag{37}$$

where the coefficients $\{d_m\}_{m\geq -1}$ are determined uniquely by the identity $S(z)\,\kappa_\tau(\mathbf{w}) = 1$. Multiplying (33) and (37) and matching coefficients gives the recursion

$$c_1 d_{-1} = 1 \;\Rightarrow\; d_{-1} = \frac{1}{c_1},$$

and for any integer $\ell \geq 0$,

$$d_\ell = -\frac{1}{c_1} \sum_{n=2}^{\ell+2} c_n d_{\ell+1-n}. \tag{38}$$

Equations (35), (33), (37), and (38) provide a complete analytic description of $\kappa_\tau(\mathbf{w})$ as a Laurent series in $z = \tau - \tau_0$.

In particular, the leading term is

$$\kappa_\tau(\mathbf{w}) \sim \frac{d_{-1}}{z} = \frac{1}{c_1(\tau - \tau_0)} = \frac{a_2}{\tau - \tau_0} = \frac{\frac{1}{2}\operatorname{Var}_{\mathbf{q}_{\mathrm{ref}}}\big(r(\mathbf{w}, y_j)\big)}{\tau - R(\mathbf{w}, \mathbf{q}_{\mathrm{ref}})} \quad \text{as } \tau \to \tau_0.$$

G. Closed-Form Solution of the Simplex-Constrained QP

In each iteration of Algorithm 1, the ensemble weights are updated by solving the convex quadratic program

$$\min_{\mathbf{w}\in\mathbb{R}^M} \; \frac{1}{2}\mathbf{w}^\top \mathbf{H}\mathbf{w} + \mathbf{g}^\top \mathbf{w} \quad \text{s.t.} \quad \mathbf{w} \geq 0, \;\; \mathbf{1}^\top \mathbf{w} = 1, \tag{39}$$

In our application at iteration t, we set $\mathbf{H} = \mathbf{H}_t$ and $\mathbf{g} = \mathbf{g}_t$ as defined below.

Let $\mathbf{w}^\star$ denote the unique minimizer of (39). We define the active and inactive index sets at the optimum as

$$S := \{i \in \{1, \ldots, M\} : w_i^\star > 0\}, \qquad S^c := \{1, \ldots, M\} \setminus S.$$

For any index sets $I, J \subseteq \{1, \ldots, M\}$ we use the standard block notation:

$$\mathbf{w}_I := (w_i)_{i\in I}, \quad \mathbf{g}_I := (g_i)_{i\in I}, \quad \mathbf{1}_I := (1)_{i\in I}, \quad \mathbf{H}_{IJ} := (H_{ij})_{i\in I, j\in J}.$$

In our RSE update (17), the QP has the form

$$\min_{\mathbf{w}\in\Delta_M} \; \mathbf{w}^\top \mathbf{A}_{\mathrm{adv},t}\mathbf{w} + \mathbf{b}_{\mathrm{adv},t}^\top \mathbf{w} + \frac{\rho}{2}\|\mathbf{w} - \mathbf{w}_t\|_2^2.$$

Expanding the proximal term and collecting quadratic/linear parts gives (39) (at iteration t) with $\mathbf{H} = \mathbf{H}_t$ and $\mathbf{g} = \mathbf{g}_t$, where

$$\mathbf{H}_t = 2\mathbf{A}_{\mathrm{adv},t} + \rho\mathbf{I}, \qquad \mathbf{g}_t = \mathbf{b}_{\mathrm{adv},t} - \rho\mathbf{w}_t, \tag{40}$$

(up to an additive constant independent of $\mathbf{w}$).

The feasible region is the probability simplex Δ_M. In our RSE update (40), $\rho > 0$ implies

$$\mathbf{H}_t = 2\mathbf{A}_{\text{adv},t} + \rho\mathbf{I} \succ 0,$$

so problem (39) is strictly convex and therefore has a unique global minimizer, denoted $\mathbf{w}^\star$. Since $\mathbf{H}_t \succ 0$, every principal submatrix is positive definite; hence $(\mathbf{H}_t)_{SS} \succ 0$ and is invertible.

Geometrically, this step projects the unconstrained quadratic minimizer onto the simplex. Indeed, since $\mathbf{H} \succ 0$, the unconstrained minimizer is $-\mathbf{H}^{-1}\mathbf{g}$, and completing the square gives

$$\frac{1}{2}\mathbf{w}^\top\mathbf{H}\mathbf{w} + \mathbf{g}^\top\mathbf{w} = \frac{1}{2}\left\|\mathbf{w} + \mathbf{H}^{-1}\mathbf{g}\right\|_{\mathbf{H}}^2 - \frac{1}{2}\mathbf{g}^\top\mathbf{H}^{-1}\mathbf{g},$$

where $\|\mathbf{u}\|_{\mathbf{H}}^2 := \mathbf{u}^\top\mathbf{H}\mathbf{u}$. Hence

$$\mathbf{w}^\star = \arg\min_{\mathbf{w}\in\Delta_M} \left\|\mathbf{w} - (-\mathbf{H}^{-1}\mathbf{g})\right\|_{\mathbf{H}}^2,$$

i.e., $\mathbf{w}^\star$ is the projection of $-\mathbf{H}^{-1}\mathbf{g}$ onto Δ_M in the $\mathbf{H}$-norm.

G.1. KKT Conditions and Active Set Framework

The Lagrangian of (39) is

$$\mathcal{L}(\mathbf{w}, \lambda, \boldsymbol{\mu}) = \frac{1}{2}\mathbf{w}^\top\mathbf{H}\mathbf{w} + \mathbf{g}^\top\mathbf{w} + \lambda(\mathbf{1}^\top\mathbf{w} - 1) - \boldsymbol{\mu}^\top\mathbf{w},$$

where

- $\lambda \in \mathbb{R}$ is the multiplier for the equality constraint $\mathbf{1}^\top\mathbf{w} = 1$;
- $\boldsymbol{\mu} \in \mathbb{R}_{\geq 0}^M$ are the multipliers for the nonnegativity constraints $w_i \geq 0$.

The full KKT conditions are

$$\mathbf{H}\mathbf{w} + \mathbf{g} + \lambda\mathbf{1} - \boldsymbol{\mu} = 0, \tag{41}$$

$$\mathbf{w} \geq 0, \quad \boldsymbol{\mu} \geq 0, \qquad \mu_i w_i = 0 \;\; \forall i, \tag{42}$$

$$\mathbf{1}^\top\mathbf{w} = 1. \tag{43}$$

Complementary slackness (42) implies

$$i \in S \Rightarrow \mu_i = 0, \qquad i \in S^c \Rightarrow w_i = 0, \;\; \mu_i \geq 0.$$

Splitting the gradient condition (41) yields:

$$\mathbf{H}_{SS}\mathbf{w}_S + \mathbf{g}_S + \lambda\mathbf{1}_S = 0, \tag{44}$$

$$\mathbf{H}_{S^cS}\mathbf{w}_S + \mathbf{g}_{S^c} + \lambda\mathbf{1}_{S^c} = \boldsymbol{\mu}_{S^c} \geq 0. \tag{45}$$

The equality constraint reduces to

$$\mathbf{1}_S^\top\mathbf{w}_S = 1. \tag{46}$$

Equations (44) and (46) produce the linear system

$$\begin{bmatrix} \mathbf{H}_{SS} & \mathbf{1}_S \\ \mathbf{1}_S^\top & 0 \end{bmatrix} \begin{bmatrix} \mathbf{w}_S \\ \lambda \end{bmatrix} = \begin{bmatrix} -\mathbf{g}_S \\ 1 \end{bmatrix}. \tag{A1}$$

The inactive coordinates are fixed at $\mathbf{w}_{S^c} = 0$, and feasibility requires $\boldsymbol{\mu}_{S^c} \geq 0$ via (45).

Thus, solving the QP reduces to identifying the correct active set S and then solving (A1).

G.2. Closed-Form Analytical Solution for a Fixed Active Set

System (A1) admits the explicit closed-form solution:

$$\lambda^\star = \frac{\mathbf{1}_S^\top \mathbf{H}_{SS}^{-1}(-\mathbf{g}_S) - 1}{\mathbf{1}_S^\top \mathbf{H}_{SS}^{-1}\mathbf{1}_S}, \tag{A2}$$

$$\mathbf{w}_S^\star = -\mathbf{H}_{SS}^{-1}\left(\mathbf{g}_S + \lambda^\star \mathbf{1}_S\right), \tag{A3}$$

with

$$\mathbf{w}_{S^c}^\star = \mathbf{0}.$$

These equations follow directly from block elimination (Schur complement) applied to (A1). When a candidate active set S is correct, $\mathbf{w}_S^\star$ computed from (A3) satisfies

$$\mathbf{w}_S^\star > 0, \qquad \boldsymbol{\mu}_{S^c} = \mathbf{H}_{S^cS}\mathbf{w}_S^\star + \mathbf{g}_{S^c} + \lambda^\star \mathbf{1}_{S^c} \geq 0.$$

If any component of $\mathbf{w}_S^\star$ is nonpositive, that index cannot belong to the active set and must be removed. This leads naturally to an active-set algorithm.

In our implementation we solve (39) using a standard convex QP solver; (A2)-(A3) provide the KKT characterization of the unique minimizer.

H. Convergence of the Proximal RSE Alternating Procedure

This appendix analyzes the proximal offline RSE iteration in Algorithm 1. We first differentiate the KL fragility root $\kappa_\tau(\mathbf{w})$ defined by $\Phi(\kappa; \mathbf{w}) = \tau$, and relate $\nabla_{\mathbf{w}}\kappa_\tau(\mathbf{w})$ to the gradient of the adversarially tilted quadratic risk. We then show that the proximal QP step yields a sufficient decrease in $\kappa_\tau(\mathbf{w}_t)$ and that every limit point satisfies the simplex stationarity conditions. A stopping certificate follows from the QP KKT conditions.

H.1. Standing definitions and well-posed regime

Recall the class-conditional risks $r(\mathbf{w}, y_j)$ and the risk vector $\mathbf{r}(\mathbf{w}) = (r(\mathbf{w}, y_1), \ldots, r(\mathbf{w}, y_K))^\top$ defined in (2)–(3). Under the Brier-score quadratic model (Section 4.1), each $r(\mathbf{w}, y_j)$ is the convex quadratic (11):

$$r(\mathbf{w}, y_j) = \mathbf{w}^\top \mathbf{A}_j \mathbf{w} + \mathbf{b}_j^\top \mathbf{w} + 1, \qquad \mathbf{A}_j \succeq 0.$$

For $\kappa > 0$, define the KL-penalized inner value (as in (12))

$$\Phi(\kappa; \mathbf{w}) := \sup_{\mathbf{q} \in \Delta_K} \Big\{ \mathbf{q}^\top \mathbf{r}(\mathbf{w}) - \kappa\, D_{\mathrm{KL}}(\mathbf{q} \,\|\, \mathbf{q}_{\mathrm{ref}}) \Big\}. \tag{47}$$

By Lemma 4.1, Φ has the closed form (14), and the unique maximizer is the exponential tilt $\mathbf{q}^\star(\kappa, \mathbf{w})$ from (13).

We recall the well-posed regime $\mathcal{W}_\tau$ and the extended definition of $\kappa_\tau(\mathbf{w})$ from Section 3. Throughout this appendix we work on $\mathcal{W}_\tau$ (or a compact subset $\mathcal{K} \subset \mathcal{W}_\tau$).

Assumption H.1 (Iterates remain in a compact well-posed region)**.** Assume $(q_{\mathrm{ref}})_j > 0$ for all j. Moreover, there exists a compact set $\mathcal{K} \subset \mathcal{W}_\tau$ such that the iterates satisfy $\mathbf{w}_t \in \mathcal{K}$ for all t.

Remark H.2. If $(q_{\mathrm{ref}})_j = 0$ for some j, the KL-based root defining $\kappa_\tau(\mathbf{w})$ is ill-posed, corresponding to the absence of any observations from class j. In practice, reference probabilities estimated from finite data are strictly positive, possibly after smoothing. The compactness assumption excludes degenerate boundary cases required for the well-posedness of the analysis.

Lemma H.3 (Uniform boundedness of the class-conditional risks)**.** *Under the Brier score, for every $\mathbf{w} \in \Delta_M$ and every class j,*

$$0 \leq r(\mathbf{w}, y_j) \leq 2.$$

Consequently, for every $\mathbf{q} \in \Delta_K$, $0 \leq R(\mathbf{w}, \mathbf{q}) \leq 2$.

Proof. For a probability vector $\widehat{\mathbf{p}} \in \Delta_K$ and true class y_j, the Brier loss is $L(y_j, \widehat{\mathbf{p}}) = \sum_{k=1}^{K}(\widehat{p}_k - \mathbf{1}\{k = j\})^2$. It is always nonnegative. Its maximum over Δ_K is achieved at a vertex $\widehat{\mathbf{p}} = \mathbf{e}_\ell$ with $\ell \neq j$, giving $L = (0-1)^2 + (1-0)^2 = 2$. Taking expectations yields the bounds for r and R. □

Lemma H.4 (Uniform gap and bounded fragility on compact subsets of $\mathcal{W}_\tau$)**.** *Under Assumption H.1, there exist constants $0 < \underline{\kappa} \leq \overline{\kappa} < \infty$ and $\delta > 0$ such that for all $\mathbf{w} \in \mathcal{K}$,*

$$\kappa_\tau(\mathbf{w}) \in [\underline{\kappa}, \overline{\kappa}], \qquad \mathbb{E}_{\mathbf{q}^\star(\kappa_\tau(\mathbf{w}),\mathbf{w})}[r(\mathbf{w}, Y)] - \tau \geq \delta.$$

Equivalently, $\partial_\kappa \Phi(\kappa_\tau(\mathbf{w}); \mathbf{w})$ is uniformly bounded away from 0 on $\mathcal{K}$.

Proof. On $\mathcal{W}_\tau$, $\kappa_\tau(\mathbf{w})$ is the unique root of $\Phi(\kappa; \mathbf{w}) = \tau$ with $\kappa \in (0, \infty)$, and $\kappa \mapsto \Phi(\kappa; \mathbf{w})$ is continuous and strictly decreasing. Standard root-continuity arguments imply $\mathbf{w} \mapsto \kappa_\tau(\mathbf{w})$ is continuous on $\mathcal{W}_\tau$. Hence it is continuous on the compact set $\mathcal{K}$, so it attains a finite minimum and maximum: $\underline{\kappa} := \min_{\mathcal{K}} \kappa_\tau(\mathbf{w}) > 0$ and $\overline{\kappa} := \max_{\mathcal{K}} \kappa_\tau(\mathbf{w}) < \infty$.

Define $g(\mathbf{w}) := \mathbb{E}_{\mathbf{q}^\star(\kappa_\tau(\mathbf{w}),\mathbf{w})}[r(\mathbf{w}, Y)] - \tau$. Fix any $\mathbf{w} \in \mathcal{W}_\tau$ and let $\kappa = \kappa_\tau(\mathbf{w})$ and $\mathbf{q}^\star = \mathbf{q}^\star(\kappa, \mathbf{w})$. Since $\mathbf{q}^\star$ attains the supremum in (47), at the root $\Phi(\kappa; \mathbf{w}) = \tau$ we have

$$\tau = \Phi(\kappa; \mathbf{w}) = \mathbb{E}_{\mathbf{q}^\star}[r(\mathbf{w}, Y)] - \kappa\, D_{\mathrm{KL}}(\mathbf{q}^\star \| \mathbf{q}_{\mathrm{ref}}),$$

hence

$$g(\mathbf{w}) = \mathbb{E}_{\mathbf{q}^\star}[r(\mathbf{w}, Y)] - \tau = \kappa\, D_{\mathrm{KL}}(\mathbf{q}^\star \| \mathbf{q}_{\mathrm{ref}}).$$

Moreover, on $\mathcal{W}_\tau$ we have $R(\mathbf{w}, \mathbf{q}_{\mathrm{ref}}) < \tau$, so the maximizer at the root cannot be $\mathbf{q}^\star = \mathbf{q}_{\mathrm{ref}}$; therefore $D_{\mathrm{KL}}(\mathbf{q}^\star \| \mathbf{q}_{\mathrm{ref}}) > 0$, and since $\kappa > 0$ on $\mathcal{W}_\tau$, we conclude $g(\mathbf{w}) > 0$ for all $\mathbf{w} \in \mathcal{W}_\tau$. By continuity of g and compactness of $\mathcal{K}$, it follows that

$$\delta := \min_{\mathbf{w} \in \mathcal{K}} g(\mathbf{w}) > 0.$$

□

Lemma H.5 (Lower bound on the scaling $c(\mathbf{w})$)**.** *Under Assumption H.1, define*

$$c(\mathbf{w}) := \frac{\kappa_\tau(\mathbf{w})}{\mathbb{E}_{\mathbf{q}^\star}[r(\mathbf{w}, Y)] - \tau}, \quad \mathbf{q}^\star = \mathbf{q}^\star(\kappa_\tau(\mathbf{w}), \mathbf{w}).$$

Then $c(\mathbf{w})$ is continuous on $\mathcal{K}$ and admits a uniform lower bound

$$c(\mathbf{w}) \geq c_{\min} := \min_{\mathbf{u} \in \mathcal{K}} c(\mathbf{u}) > 0.$$

Moreover, using Lemma H.3 one has the explicit bound

$$c(\mathbf{w}) \;\geq\; \frac{\underline{\kappa}}{2 - \tau} \qquad \textit{for all } \mathbf{w} \in \mathcal{K},$$

where $\underline{\kappa}$ is from Lemma H.4.

Proof. By Lemma H.4, the denominator is uniformly $\geq \delta > 0$ on $\mathcal{K}$ and $\kappa_\tau(\mathbf{w})$ is continuous and positive on $\mathcal{K}$. Thus $c(\mathbf{w})$ is continuous and positive on compact $\mathcal{K}$, so it attains a positive minimum.

For the explicit bound, Lemma H.3 gives $\mathbb{E}_{\mathbf{q}^\star}[r(\mathbf{w}, Y)] \leq 2$, hence the denominator $\leq 2 - \tau$. Using $\kappa_\tau(\mathbf{w}) \geq \underline{\kappa}$ yields $c(\mathbf{w}) \geq \underline{\kappa}/(2 - \tau)$. □

Remark H.6 (Interpretation of $c_{\min}$). The constant $c_{\min}$ is a local conditioning constant for the convergence proof, not a hyperparameter used by the algorithm. At the exact root,

$$\tau = \mathbb{E}_{\mathbf{q}^\star}[r(\mathbf{w}, Y)] - \kappa_\tau(\mathbf{w}) D_{\mathrm{KL}}(\mathbf{q}^\star \| \mathbf{q}_{\mathrm{ref}}),$$

and therefore

$$c(\mathbf{w}) = \frac{\kappa_\tau(\mathbf{w})}{\mathbb{E}_{\mathbf{q}^\star}[r(\mathbf{w}, Y)] - \tau} = \frac{1}{D_{\mathrm{KL}}(\mathbf{q}^\star \| \mathbf{q}_{\mathrm{ref}})}.$$

Hence $c_{\min}$ can be small when the adversarial tilt $\mathbf{q}^\star$ is very far from $\mathbf{q}_{\text{ref}}$, for example when $\mathbf{q}_{\text{ref}}$ assigns very small mass to a class. However, for a fixed full-support reference distribution with

$$q_{\min} := \min_j (q_{\text{ref}})_j > 0,$$

we have

$$D_{\text{KL}}(\mathbf{q}^\star \| \mathbf{q}_{\text{ref}}) \le \log \frac{1}{q_{\min}},$$

and therefore

$$c(\mathbf{w}) \ge \frac{1}{\log(1/q_{\min})}.$$

Thus $c_{\min} > 0$ under the stated full-support and compact well-posedness assumptions. The magnitude of the lower bound only affects the conservativeness of the sufficient condition on ρ; it does not change the algorithm or the validity of the convergence statement under the stated assumptions.

Lemma H.7 (Lipschitz continuity of $\nabla\kappa_\tau$ on $\mathcal{K}$)**.** *Under Assumption H.1, the map $\mathbf{w} \mapsto \kappa_\tau(\mathbf{w})$ is continuously differentiable on $\mathcal{K}$. Moreover, there exists $L < \infty$ such that*

$$\|\nabla\kappa_\tau(\mathbf{w}) - \nabla\kappa_\tau(\mathbf{w}')\|_2 \le L\|\mathbf{w} - \mathbf{w}'\|_2 \qquad \forall \mathbf{w}, \mathbf{w}' \in \mathcal{K}.$$

Proof. Each $r(\mathbf{w}, y_j)$ is a quadratic polynomial in $\mathbf{w}$, so $r(\cdot, y_j)$ is C^∞ with globally bounded Hessian on Δ_M. Also, $\Phi(\kappa; \mathbf{w}) = \kappa \log \sum_j (q_{\text{ref}})_j \exp(r(\mathbf{w}, y_j)/\kappa)$ is smooth in $(\kappa, \mathbf{w})$ on any set where κ is bounded away from 0. By Lemma H.4, on $\mathcal{K}$ we have $\kappa_\tau(\mathbf{w}) \in [\underline{\kappa}, \overline{\kappa}]$ and $\partial_\kappa \Phi(\kappa_\tau(\mathbf{w}); \mathbf{w})$ is bounded away from 0. Hence the implicit function theorem yields that κ_τ is C^1 on $\mathcal{K}$.

Furthermore, differentiating the implicit relation $\Phi(\kappa_\tau(\mathbf{w}); \mathbf{w}) = \tau$ shows $\nabla\kappa_\tau$ is a composition of smooth functions whose derivatives are continuous on the compact set $\{(\kappa, \mathbf{w}) : \mathbf{w} \in \mathcal{K},\ \kappa = \kappa_\tau(\mathbf{w})\}$. Therefore the Hessian $\nabla^2\kappa_\tau$ is bounded on $\mathcal{K}$, and the claimed Lipschitz bound follows with $L := \sup_{\mathbf{u}\in\mathcal{K}} \|\nabla^2\kappa_\tau(\mathbf{u})\|_{\text{op}} < \infty$. □

H.2. Differentiability of Φ and $\kappa_\tau(\mathbf{w})$

Define the normalizer

$$Z(\kappa, \mathbf{w}) := \sum_{j=1}^{K} (q_{\text{ref}})_j \exp\big(r(\mathbf{w}, y_j)/\kappa\big),$$

so that $\Phi(\kappa; \mathbf{w}) = \kappa \log Z(\kappa, \mathbf{w})$ by (14). Also define the KL-adversary (Lemma 4.1)

$$(q^\star(\kappa, \mathbf{w}))_j = \frac{(q_{\text{ref}})_j \exp\big(r(\mathbf{w}, y_j)/\kappa\big)}{Z(\kappa, \mathbf{w})}. \tag{48}$$

Lemma H.8 (Gradients of Φ)**.** *For any $\kappa > 0$ and $\mathbf{w} \in \Delta_M$, the function $\Phi(\kappa; \mathbf{w})$ is continuously differentiable in $(\kappa, \mathbf{w})$ and satisfies*

$$\nabla_{\mathbf{w}} \Phi(\kappa; \mathbf{w}) = \sum_{j=1}^{K} (q^\star(\kappa, \mathbf{w}))_j \, \nabla_{\mathbf{w}} r(\mathbf{w}, y_j), \tag{49}$$

$$\frac{\partial}{\partial\kappa} \Phi(\kappa; \mathbf{w}) = \log Z(\kappa, \mathbf{w}) - \frac{1}{\kappa} \sum_{j=1}^{K} (q^\star(\kappa, \mathbf{w}))_j \, r(\mathbf{w}, y_j) = \frac{\Phi(\kappa; \mathbf{w}) - \mathbb{E}_{\mathbf{q}^\star}[r(\mathbf{w}, Y)]}{\kappa}. \tag{50}$$

Proof. Using $\Phi(\kappa; \mathbf{w}) = \kappa \log Z(\kappa, \mathbf{w})$ and the chain rule,

$$\nabla_{\mathbf{w}} \Phi(\kappa; \mathbf{w}) = \kappa \frac{1}{Z(\kappa, \mathbf{w})} \nabla_{\mathbf{w}} Z(\kappa, \mathbf{w}).$$

Since

$$\nabla_{\mathbf{w}} Z(\kappa, \mathbf{w}) = \sum_{j=1}^{K} (q_{\text{ref}})_j \exp(r(\mathbf{w}, y_j)/\kappa) \cdot \frac{1}{\kappa} \nabla_{\mathbf{w}} r(\mathbf{w}, y_j),$$

substitution yields (49) after recognizing (48).

For the κ-derivative, write

$$\frac{\partial}{\partial \kappa}\Phi(\kappa;\mathbf{w}) = \log Z(\kappa,\mathbf{w}) + \kappa \frac{1}{Z(\kappa,\mathbf{w})}\frac{\partial}{\partial \kappa} Z(\kappa,\mathbf{w}).$$

Because $\frac{\partial}{\partial \kappa}\exp(r/\kappa) = \exp(r/\kappa)\cdot(-r/\kappa^2)$, we have

$$\frac{\partial}{\partial \kappa} Z(\kappa,\mathbf{w}) = -\frac{1}{\kappa^2}\sum_{j=1}^{K}(q_{\mathrm{ref}})_j \exp(r(\mathbf{w},y_j)/\kappa)\, r(\mathbf{w},y_j),$$

and substituting gives the first equality in (50). The last equality follows from $\Phi(\kappa;\mathbf{w}) = \kappa \log Z(\kappa,\mathbf{w})$ and the definition of $\mathbf{q}^\star$. □

Lemma H.9 (Implicit gradient of $\kappa_\tau(\mathbf{w})$)**.** *Under Assumption H.1, the function $\mathbf{w} \mapsto \kappa_\tau(\mathbf{w})$ is continuously differentiable along the iterates $\{\mathbf{w}_t\}$, and for any $\mathbf{w} \in \mathcal{W}_\tau$:*

$$\nabla_{\mathbf{w}}\kappa_\tau(\mathbf{w}) = -\frac{\nabla_{\mathbf{w}}\Phi(\kappa_\tau(\mathbf{w});\mathbf{w})}{\partial_\kappa \Phi(\kappa_\tau(\mathbf{w});\mathbf{w})} = \frac{\kappa_\tau(\mathbf{w})}{\mathbb{E}_{\mathbf{q}^\star}[r(\mathbf{w},Y)] - \tau}\sum_{j=1}^{K}(q^\star)_j\, \nabla_{\mathbf{w}} r(\mathbf{w},y_j), \tag{51}$$

where $\mathbf{q}^\star = \mathbf{q}^\star(\kappa_\tau(\mathbf{w}),\mathbf{w})$. Moreover,

$$\partial_\kappa \Phi(\kappa_\tau(\mathbf{w});\mathbf{w}) < 0 \quad \textit{and thus} \quad \mathbb{E}_{\mathbf{q}^\star}[r(\mathbf{w},Y)] > \tau \qquad \textit{for all } \mathbf{w} \in \mathcal{W}_\tau. \tag{52}$$

Proof. Fix $\mathbf{w} \in \mathcal{W}_\tau$ and define $F(\kappa,\mathbf{w}) := \Phi(\kappa;\mathbf{w}) - \tau$. By Lemma H.8, F is C^1. Since $\kappa_\tau(\mathbf{w})$ satisfies $F(\kappa_\tau(\mathbf{w}),\mathbf{w}) = 0$, the implicit function theorem implies differentiability provided $\partial_\kappa \Phi(\kappa_\tau(\mathbf{w});\mathbf{w}) \neq 0$.

At $\kappa = \kappa_\tau(\mathbf{w})$, we have $\Phi(\kappa;\mathbf{w}) = \tau$. Let $\mathbf{q}^\star = \mathbf{q}^\star(\kappa,\mathbf{w})$. Then

$$\tau = \Phi(\kappa;\mathbf{w}) = \mathbf{q}^{\star\top}\mathbf{r}(\mathbf{w}) - \kappa\, D_{\mathrm{KL}}(\mathbf{q}^\star \| \mathbf{q}_{\mathrm{ref}}).$$

We claim that

$$D_{\mathrm{KL}}(\mathbf{q}^\star \| \mathbf{q}_{\mathrm{ref}}) > 0.$$

Indeed, if $D_{\mathrm{KL}}(\mathbf{q}^\star \| \mathbf{q}_{\mathrm{ref}}) = 0$, then $\mathbf{q}^\star = \mathbf{q}_{\mathrm{ref}}$. Hence

$$\tau = \Phi(\kappa;\mathbf{w}) = R(\mathbf{w},\mathbf{q}_{\mathrm{ref}}),$$

contradicting $\mathbf{w} \in \mathcal{W}_\tau$, which requires

$$R(\mathbf{w},\mathbf{q}_{\mathrm{ref}}) < \tau.$$

Therefore $D_{\mathrm{KL}}(\mathbf{q}^\star \| \mathbf{q}_{\mathrm{ref}}) > 0$. Since $\kappa = \kappa_\tau(\mathbf{w}) > 0$, we obtain

$$\tau = \mathbf{q}^{\star\top}\mathbf{r}(\mathbf{w}) - \kappa\, D_{\mathrm{KL}}(\mathbf{q}^\star \| \mathbf{q}_{\mathrm{ref}}) < \mathbf{q}^{\star\top}\mathbf{r}(\mathbf{w}) = \mathbb{E}_{\mathbf{q}^\star}[r(\mathbf{w},Y)].$$

Therefore $\mathbb{E}_{\mathbf{q}^\star}[r(\mathbf{w},Y)] > \tau$, and substituting into (50) gives $\partial_\kappa \Phi(\kappa_\tau(\mathbf{w});\mathbf{w}) = (\tau - \mathbb{E}_{\mathbf{q}^\star}[r])/\kappa_\tau(\mathbf{w}) < 0$, establishing (52).

Finally, differentiate the identity $F(\kappa_\tau(\mathbf{w}),\mathbf{w}) = \Phi(\kappa_\tau(\mathbf{w});\mathbf{w}) - \tau = 0$ with respect to $\mathbf{w}$. By the chain rule,

$$\mathbf{0} = \nabla_{\mathbf{w}} F(\kappa_\tau(\mathbf{w}),\mathbf{w}) = \nabla_{\mathbf{w}}\Phi(\kappa_\tau(\mathbf{w});\mathbf{w}) + \partial_\kappa \Phi(\kappa_\tau(\mathbf{w});\mathbf{w})\, \nabla_{\mathbf{w}}\kappa_\tau(\mathbf{w}).$$

Since $\partial_\kappa \Phi(\kappa_\tau(\mathbf{w});\mathbf{w}) \neq 0$ (indeed it is negative by (52)), we can solve for $\nabla_{\mathbf{w}}\kappa_\tau(\mathbf{w})$:

$$\nabla_{\mathbf{w}}\kappa_\tau(\mathbf{w}) = -\frac{\nabla_{\mathbf{w}}\Phi(\kappa_\tau(\mathbf{w});\mathbf{w})}{\partial_\kappa \Phi(\kappa_\tau(\mathbf{w});\mathbf{w})}.$$

Using Lemma H.8 gives $\nabla_{\mathbf{w}}\Phi(\kappa_\tau(\mathbf{w});\mathbf{w}) = \sum_{j=1}^{K}(q^\star)_j \nabla_{\mathbf{w}} r(\mathbf{w},y_j)$, and at the root $\Phi(\kappa_\tau(\mathbf{w});\mathbf{w}) = \tau$ we also have from (50)

$$\partial_\kappa \Phi(\kappa_\tau(\mathbf{w});\mathbf{w}) = \frac{\tau - \mathbb{E}_{\mathbf{q}^\star}[r(\mathbf{w},Y)]}{\kappa_\tau(\mathbf{w})}.$$

Substituting these two identities into the previous display yields (51).

□

H.3. Proximal QP update, descent, and a stopping certificate

At iteration t, Algorithm 1 forms the adversary $\mathbf{q}_t^\star = \mathbf{q}^\star(\kappa_t, \mathbf{w}_t)$ where $\kappa_t = \kappa_\tau(\mathbf{w}_t)$, and defines the tilted (adversarial) quadratic risk

$$\widetilde{R}_t(\mathbf{w}) := \sum_{j=1}^{K} (q_t^\star)_j \, r(\mathbf{w}, y_j) = \mathbf{w}^\top \mathbf{A}_{\mathrm{adv},t} \mathbf{w} + \mathbf{b}_{\mathrm{adv},t}^\top \mathbf{w} + 1, \tag{53}$$

with $\mathbf{A}_{\mathrm{adv},t} = \sum_j (q_t^\star)_j \mathbf{A}_j$ and $\mathbf{b}_{\mathrm{adv},t} = \sum_j (q_t^\star)_j \mathbf{b}_j$ as in (16). We use the proximal QP update

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w} \in \Delta_M} \Big\{ \widetilde{R}_t(\mathbf{w}) + \frac{\rho}{2} \|\mathbf{w} - \mathbf{w}_t\|_2^2 \Big\}, \qquad \rho > 0, \tag{54}$$

and define the step $\mathbf{d}_t := \mathbf{w}_{t+1} - \mathbf{w}_t$.

Lemma H.10 (Well-defined proximal update and VI/KKT condition)**.** *For any $\rho > 0$, the subproblem (54) admits a* unique *minimizer $\mathbf{w}_{t+1} \in \Delta_M$. Moreover, $\mathbf{w}_{t+1}$ satisfies the variational inequality*

$$\big\langle \nabla \widetilde{R}_t(\mathbf{w}_{t+1}) + \rho(\mathbf{w}_{t+1} - \mathbf{w}_t),\ \mathbf{w} - \mathbf{w}_{t+1} \big\rangle \geq 0 \qquad \forall \mathbf{w} \in \Delta_M, \tag{55}$$

equivalently,

$$\mathbf{0} \in \nabla \widetilde{R}_t(\mathbf{w}_{t+1}) + \rho(\mathbf{w}_{t+1} - \mathbf{w}_t) + N_{\Delta_M}(\mathbf{w}_{t+1}), \tag{56}$$

where $N_{\Delta_M}(\cdot)$ is the normal cone of the simplex.

Proof. $\widetilde{R}_t$ is convex quadratic on Δ_M and the proximal term is ρ-strongly convex in $\mathbf{w}$, so the objective is ρ-strongly convex. Minimization over the closed convex set Δ_M therefore has a unique solution. The VI/KKT conditions (55)–(56) are the standard first-order optimality conditions for minimizing a differentiable convex function over a closed convex set. □

Lemma H.11 (Prox-QP implies a negative directional derivative)**.** *Fix $\mathbf{w}_t \in \mathcal{W}_\tau$ and let $\mathbf{w}_{t+1}$ be the minimizer of (54). Then*

$$\big\langle \nabla \widetilde{R}_t(\mathbf{w}_t),\ \mathbf{d}_t \big\rangle \leq -\rho \|\mathbf{d}_t\|_2^2. \tag{57}$$

Consequently,

$$\big\langle \nabla \kappa_\tau(\mathbf{w}_t),\ \mathbf{d}_t \big\rangle \leq -c_t \, \rho \|\mathbf{d}_t\|_2^2, \qquad c_t := \frac{\kappa_t}{\mathbb{E}_{\mathbf{q}_t^\star}[r(\mathbf{w}_t, Y)] - \tau} > 0, \tag{58}$$

where $\kappa_t = \kappa_\tau(\mathbf{w}_t)$ and $\mathbf{q}_t^\star = \mathbf{q}^\star(\kappa_t, \mathbf{w}_t)$.

Proof. From (56) we have $-\rho \mathbf{d}_t \in \nabla \widetilde{R}_t(\mathbf{w}_{t+1}) + N_{\Delta_M}(\mathbf{w}_{t+1})$. Equivalently, for all $\mathbf{w} \in \Delta_M$,

$$\big\langle \nabla \widetilde{R}_t(\mathbf{w}_{t+1}) + \rho \mathbf{d}_t,\ \mathbf{w} - \mathbf{w}_{t+1} \big\rangle \geq 0.$$

Taking $\mathbf{w} = \mathbf{w}_t$ yields $\langle \nabla \widetilde{R}_t(\mathbf{w}_{t+1}), \mathbf{d}_t \rangle \leq -\rho \|\mathbf{d}_t\|_2^2$. Since $\widetilde{R}_t$ is convex quadratic, its Hessian is PSD, hence $\langle \nabla \widetilde{R}_t(\mathbf{w}_t), \mathbf{d}_t \rangle \leq \langle \nabla \widetilde{R}_t(\mathbf{w}_{t+1}), \mathbf{d}_t \rangle$, which gives (57). The second inequality (58) follows from Lemma H.9, i.e., $\nabla \kappa_\tau(\mathbf{w}_t) = c_t \nabla \widetilde{R}_t(\mathbf{w}_t)$ with $c_t > 0$. □

Corollary H.12 (Stopping certificate via KKT residual)**.** *Let $\mathbf{w}_{t+1}$ be the proximal update (54). Then*

$$\operatorname{dist}\Big(\mathbf{0},\ \nabla \widetilde{R}_t(\mathbf{w}_{t+1}) + N_{\Delta_M}(\mathbf{w}_{t+1})\Big) \leq \rho \, \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2. \tag{59}$$

In particular, if the algorithm stops when $\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2 \leq \varepsilon_{\mathrm{stop}}$, then the KKT residual of the QP at $\mathbf{w}_{t+1}$ is at most $\rho \, \varepsilon_{\mathrm{stop}}$.

Proof. From (56), $-\rho(\mathbf{w}_{t+1} - \mathbf{w}_t) \in \nabla \widetilde{R}_t(\mathbf{w}_{t+1}) + N_{\Delta_M}(\mathbf{w}_{t+1})$. Thus $\operatorname{dist}(\mathbf{0}, S) \leq \|s\|$ for any $s \in S$ implies (59). □

Lemma H.13 (Uniform sufficient decrease for κ_τ)**.** *Under Assumption H.1, let $\mathcal{K}$ be the compact set containing the iterates. Since $\mathcal{W}_\tau$ is open relative to the simplex and $\mathcal{K} \subset \mathcal{W}_\tau$ is compact, there exists $r_\mathcal{K} > 0$ such that the closed neighborhood*

$$\mathcal{K}_{r_\mathcal{K}} := \{\mathbf{w} \in \Delta_M : \operatorname{dist}(\mathbf{w}, \mathcal{K}) \le r_\mathcal{K}\}$$

is contained in $\mathcal{W}_\tau$.

Let $L < \infty$ be a Lipschitz constant of $\nabla\kappa_\tau$ on $\mathcal{K}_{r_\mathcal{K}}$, let $c_{\min} > 0$ be the uniform lower bound from Lemma H.5, and let

$$M_{\text{grad}} := \sup_{t, \mathbf{w} \in \Delta_M} \|\nabla \widetilde{R}_t(\mathbf{w})\|_2 < \infty.$$

If

$$\rho > \max\left\{\frac{L}{2c_{\min}}, \frac{M_{\text{grad}}}{r_\mathcal{K}}\right\},$$

then there exists a constant $c > 0$ independent of t such that

$$\kappa_\tau(\mathbf{w}_{t+1}) \le \kappa_\tau(\mathbf{w}_t) - c\,\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2. \tag{60}$$

Proof. Let

$$\mathbf{d}_t := \mathbf{w}_{t+1} - \mathbf{w}_t.$$

We first show that the entire line segment between $\mathbf{w}_t$ and $\mathbf{w}_{t+1}$ remains inside the smooth region $\mathcal{K}_{r_\mathcal{K}} \subset \mathcal{W}_\tau$.

From the proximal KKT condition (56), taking the inner product with $\mathbf{d}_t$ gives

$$\rho\|\mathbf{d}_t\|_2^2 \le \left\langle -\nabla\widetilde{R}_t(\mathbf{w}_{t+1}), \mathbf{d}_t \right\rangle \le \|\nabla\widetilde{R}_t(\mathbf{w}_{t+1})\|_2 \|\mathbf{d}_t\|_2 \le M_{\text{grad}}\|\mathbf{d}_t\|_2.$$

If $\mathbf{d}_t \neq 0$, this implies

$$\|\mathbf{d}_t\|_2 \le \frac{M_{\text{grad}}}{\rho} < r_\mathcal{K}.$$

If $\mathbf{d}_t = 0$, the claim is immediate. Hence every point on the segment

$$\mathbf{w}_t + s\mathbf{d}_t, \qquad s \in [0, 1],$$

has distance at most $r_\mathcal{K}$ from $\mathcal{K}$, and therefore belongs to $\mathcal{K}_{r_\mathcal{K}}$.

We may now apply the descent lemma to κ_τ on $\mathcal{K}_{r_\mathcal{K}}$:

$$\kappa_\tau(\mathbf{w}_{t+1}) \le \kappa_\tau(\mathbf{w}_t) + \langle \nabla\kappa_\tau(\mathbf{w}_t), \mathbf{d}_t\rangle + \frac{L}{2}\|\mathbf{d}_t\|_2^2.$$

Using (58) and $c_t \ge c_{\min}$ gives

$$\kappa_\tau(\mathbf{w}_{t+1}) \le \kappa_\tau(\mathbf{w}_t) - \left(c_{\min}\rho - \frac{L}{2}\right)\|\mathbf{d}_t\|_2^2.$$

Since $\rho > L/(2c_{\min})$, the constant

$$c := c_{\min}\rho - \frac{L}{2}$$

is positive, proving (60). □

Proof. By the descent lemma (Lipschitz gradient),

$$\kappa_\tau(\mathbf{w}_{t+1}) \le \kappa_\tau(\mathbf{w}_t) + \langle \nabla\kappa_\tau(\mathbf{w}_t), \mathbf{d}_t\rangle + \frac{L}{2}\|\mathbf{d}_t\|_2^2.$$

Using (58) and $c_t \ge c_{\min}$ gives

$$\kappa_\tau(\mathbf{w}_{t+1}) \le \kappa_\tau(\mathbf{w}_t) - \left(c_{\min}\rho - \frac{L}{2}\right)\|\mathbf{d}_t\|_2^2.$$

Set $c := c_{\min}\rho - \frac{L}{2} > 0$ to obtain (60). □

H.4. Convergence to a stationary point

Let $N_{\Delta_M}(\mathbf{w})$ denote the normal cone of Δ_M at $\mathbf{w}$.

Definition H.14 (Stationary point of κ_τ on the simplex)**.** A point $\mathbf{w}^\star \in \mathcal{W}_\tau$ is *stationary* for the robust satisficing objective if

$$\mathbf{0} \in \nabla_{\mathbf{w}} \kappa_\tau(\mathbf{w}^\star) + N_{\Delta_M}(\mathbf{w}^\star).$$

Definition H.15 (η-stationarity on the simplex)**.** For $\eta \geq 0$, a point $\mathbf{w} \in \mathcal{W}_\tau$ is called η-stationary for $\min_{\mathbf{w}\in\Delta_M} \kappa_\tau(\mathbf{w})$ if

$$\operatorname{dist}\left(\mathbf{0},\, \nabla_{\mathbf{w}} \kappa_\tau(\mathbf{w}) + N_{\Delta_M}(\mathbf{w})\right) \leq \eta.$$

The case $\eta = 0$ recovers Definition H.14.

Theorem H.16 (Convergence of proximal RSE to a stationary point)**.** *Suppose Assumption H.1 holds and the update is the proximal QP step (54) with a fixed ρ chosen so that Lemma H.13 holds. Then:*

1. *The sequence $\{\kappa_\tau(\mathbf{w}_t)\}_{t\geq 0}$ is nonincreasing and convergent.*
2. *The steps vanish: $\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2 \to 0$.*
3. *Every limit point of $\{\mathbf{w}_t\}$ is a stationary point in the sense of Definition H.14. In particular, if the stationary point is unique, then $\mathbf{w}_t \to \mathbf{w}^\star$.*

Proof. (1) Lemma H.13 gives monotone decrease of $\kappa_\tau(\mathbf{w}_t)$, and $\kappa_\tau(\mathbf{w}_t) \geq 0$, hence convergence.

(2) Summing (60) from $t = 0$ to $T - 1$ yields

$$c \sum_{t=0}^{T-1} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \leq \kappa_\tau(\mathbf{w}_0) - \kappa_\tau(\mathbf{w}_T) \leq \kappa_\tau(\mathbf{w}_0),$$

so $\sum_{t\geq 0} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 < \infty$ and therefore $\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2 \to 0$.

(3) Since Δ_M is compact, pick a convergent subsequence $\mathbf{w}_{t_n} \to \bar{\mathbf{w}}$. By (2), also $\mathbf{w}_{t_n+1} \to \bar{\mathbf{w}}$. Let $\kappa_{t_n} = \kappa_\tau(\mathbf{w}_{t_n})$ and $\mathbf{q}^\star_{t_n} = \mathbf{q}^\star(\kappa_{t_n}, \mathbf{w}_{t_n})$. By continuity of $\mathbf{w} \mapsto \kappa_\tau(\mathbf{w})$ on $\mathcal{W}_\tau$ and of (48), we have $\kappa_{t_n} \to \bar{\kappa} := \kappa_\tau(\bar{\mathbf{w}})$ and $\mathbf{q}^\star_{t_n} \to \bar{\mathbf{q}}^\star := \mathbf{q}^\star(\bar{\kappa}, \bar{\mathbf{w}})$.

The proximal optimality condition (56) at time t_n is

$$\mathbf{0} \in \nabla \widetilde{R}_{t_n}(\mathbf{w}_{t_n+1}) + \rho(\mathbf{w}_{t_n+1} - \mathbf{w}_{t_n}) + N_{\Delta_M}(\mathbf{w}_{t_n+1}),$$

with $\nabla \widetilde{R}_{t_n}(\mathbf{w}) = \sum_{j=1}^K (q^\star_{t_n})_j \nabla r(\mathbf{w}, y_j)$. Since $\mathbf{w}_{t_n+1} - \mathbf{w}_{t_n} \to 0$, passing to the limit and using continuity of $\nabla r(\cdot, y_j)$ yields

$$\mathbf{0} \in \sum_{j=1}^K (\bar{q}^\star)_j \nabla r(\bar{\mathbf{w}}, y_j) + N_{\Delta_M}(\bar{\mathbf{w}}).$$

Finally, Lemma H.9 shows that $\nabla \kappa_\tau(\bar{\mathbf{w}})$ is a strictly positive scalar multiple of $\sum_j (\bar{q}^\star)_j \nabla r(\bar{\mathbf{w}}, y_j)$, hence

$$\mathbf{0} \in \nabla \kappa_\tau(\bar{\mathbf{w}}) + N_{\Delta_M}(\bar{\mathbf{w}}),$$

so $\bar{\mathbf{w}}$ is stationary (Definition H.14). □