SUPPRESS OR RELEARN? FEATURE GEOMETRY FOR ROBUSTNESS TO SPURIOUS CORRELATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks can achieve high average accuracy in classification tasks by learning spurious correlations between class labels and context. This leads to poor performance when these correlations break down. Reducing the model's reliance on contextual features can help, but may also suppress useful task information. To address this trade-off, we propose Geometry-Guided Error-Aware Robustness (GEAR), which uses validation errors to decide when to reduce the model's sensitivity along selected feature directions and when to relearn. GEAR identifies directions of feature change from easy to hard examples within each class, using them to guide suppression and define regions for relearning. Both fitting and selection require only task labels. In an additive Gaussian model, we show that neither suppression nor relearning uniformly dominates and give conditions under which GEAR asymptotically achieves optimal worst-group accuracy among linear heads on fixed features. GEAR improves mean worst-group accuracy over its initial classifiers on all six image and text benchmarks.

1 INTRODUCTION

Machine learning systems increasingly support decisions in finance (Hand & Henley, 1997), healthcare (Jiang et al., 2017), transportation (Bojarski et al., 2016), and criminal justice (Dressel & Farid, 2018). However, high average accuracy does not guarantee reliable performance under new conditions. Such failures can arise from reliance on *spurious correlations*: associations between class labels and contextual cues that are predictive in the training data but become unreliable in other contexts (Geirhos et al., 2020). For example, in the training sets of Waterbirds and SpuCoBirds, landbirds usually appear on land and waterbirds on water (Sagawa et al., 2020; Joshi et al., 2023). A classifier that relies on the background may mistake a waterbird photographed on land for a landbird. Similar shortcuts arise in language tasks. In the MultiNLI dataset, the task is to classify the relation between a premise and a hypothesis as entailment, contradiction, or neutral (Williams et al., 2018). Negation words in the hypothesis can be correlated with contradiction in the training data, but they do not determine the relation between the sentences (Sagawa et al., 2020). In both settings, contextual cues can work well on common examples while producing systematic errors on less common ones (see Figure 1 for an illustration).

To assess such failures, researchers evaluate performance across *groups* of examples, each defined by a particular combination of a task label and a contextual cue. Average accuracy can hide poor performance on rare groups in the data. Worst-group accuracy (WGA) instead reports the lowest accuracy across these groups (Sagawa et al., 2020). When group annotations are available, methods such as GroupDRO train classifiers to minimize the largest group loss (Sagawa et al., 2020). Group annotations can also guide data rebalancing: deep feature reweighting (DFR) refits only the last layer on group-balanced examples (Kirichenko et al., 2023; Izmailov et al., 2022). However, group annotations may be unavailable or costly to obtain, and other methods avoid this requirement during fitting by using signals from model predictions. JTT gives more weight to examples misclassified early in training, while AFR emphasizes low-confidence examples (Liu et al., 2021; Qiu et al., 2023). XRM uses disagreements between models to construct training environments (Pezeshki et al., 2024).

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

SpuCoBirds (bird photographs)



Landbird on *land*
95.2% of training landbirds

Landbird on *water*
4.8% of training landbirds

Waterbird on *land*
4.8% of training waterbirds

Waterbird on *water*
95.2% of training waterbirds

MultiNLI (illustrative examples)

Premise (assumed true): "The box is empty, and the bag holds a book."

No objects are in the box.	An object is in the box.	The bag holds nothing.	The box is red.
Entailment	**Contradiction**	**Contradiction**	**Neutral**
With negation	*Without negation*	*With negation*	*Without negation*

Figure 1: Task labels and contextual cues in image and text classification. In SpuCoBirds (Joshi et al., 2023), bird type is the task label and background is the contextual cue. The constructed examples illustrate the MultiNLI task (Williams et al., 2018): each input pairs the shared premise with one of the four hypotheses above. Assuming the premise is true, the task label is entailment (the hypothesis must be true), contradiction (must be false), or neutral (cannot be determined). Negation words such as "no" and "nothing" are correlated with contradiction in the training data (Sagawa et al., 2020), so a model may predict contradiction from these words alone.

Difficulty scores identify which examples need attention, but do not determine whether the associated feature variation should be suppressed or retained for relearning. Within the same class, feature differences between easy and difficult examples may reflect contextual cues and useful task information. Suppressing sensitivity to these differences can reduce reliance on misleading cues, but can also discard useful information. Relearning retains these features and emphasizes difficult examples, but can still produce a classifier that relies on context. These competing risks motivate choosing between suppression and relearning rather than committing to either correction in advance.

We propose *Geometry-Guided Error-Aware Robustness* (GEAR), which estimates directions of feature change from easy to difficult examples, uses these directions to fit suppression and relearning candidates, and selects the final classifier by comparing the validation scores of these candidates and the initial classifier. Throughout, GEAR uses only inputs and task labels. Under simplified models, we show that either correction can outperform the other and give conditions under which GEAR's validation selection asymptotically attains the best linear-head WGA. Across six image and text benchmarks, GEAR improves mean WGA over its initial classifiers. Among the compared methods using only task labels for training and validation, GEAR achieves the highest reported mean WGA on five of the six benchmarks. Ablations on Waterbirds and CelebA show that no single correction matches GEAR on both, supporting validation-based selection.

Our contributions are:

- We propose GEAR, a method for improving worst-group robustness without group annotations. GEAR uses within-class easy-to-hard feature changes to define directions for suppression and regions for weighted relearning, then selects among the fitted candidates and the initial classifier.
- We analyze when suppression or relearning is preferable in a simplified model with fixed features. We also give sufficient conditions under which GEAR's validation selection asymptotically achieves optimal WGA among linear heads.
- We conduct experiments on six image and text benchmarks. GEAR performs competitively with existing methods that do not use group annotations. Controlled ablations quantify the gains from GEAR's learned geometry and show the benefits of retaining both correction options for validation selection.

The remainder of the paper is organized as follows. Section 2 introduces the methodology of GEAR. Section 3 evaluates its performance and examines the contributions of its components. Section 4 provides an analysis of when suppression or relearning is preferable. Section 5 concludes the paper and discusses the limitations and future directions. The appendices provide implementation details, supplementary results, and proofs.

2 GEAR: Geometry Guides the Correction

GEAR starts with a trained classifier and asks how to correct its errors. It first measures example difficulty and estimates within-class directions of feature change from easy to difficult examples, then fits two types of candidates: one reduces directional sensitivity, while the other relearns from geometrically defined regions. Validation compares the fitted candidates with the initial classifier using only inputs and task labels. Figure 2 illustrates the procedure; Algorithm 1 in Appendix A provides pseudocode.

Model and training objective. Let $\mathcal{D} = \{(x_i, y_i)\}$ contain inputs x_i and task labels $y_i \in \{0, \dots, C-1\}$, with $\mathcal{T}$ and $\mathcal{V}$ indexing the training and validation sets, respectively. We write $f(x) = Wh(x) + b$ for a classifier in which h extracts features and the linear head (W, b) maps them to class scores. We denote the class-c score by $f_c(x)$, its softmax probability by $p_f(c \mid x)$, and the predicted label by $\widehat{y}_f(x) = \arg\max_c f_c(x)$. For evaluation, let $e \in \mathcal{E}$ denote a contextual attribute, such as the background in Waterbirds (Sagawa et al., 2020). A group $g = (y, e)$ contains examples with the same task label y and context e. For a classifier f, let $\mathrm{Acc}_g(f)$ denote its accuracy within group g, and let $\mathcal{G}$ denote the set of groups. The worst-group accuracy (WGA) is defined as

$$\mathrm{WGA}(f) = \min_{g \in \mathcal{G}} \mathrm{Acc}_g(f). \tag{1}$$

These group annotations define the evaluation metric, but GEAR does not require them for fitting or validation selection. To form its candidates, GEAR trains or refits classifiers by minimizing weighted cross-entropy with a regularization penalty:

$$\mathcal{L}_\omega(h, W, b) = -\sum_{i \in \mathcal{T}} \omega_i \log p_f(y_i \mid x_i) + \lambda \Omega(W, b), \qquad (h, W, b) \in \Theta, \tag{2}$$

where ω_i is the weight of training example i, and Θ specifies the allowed parameters. The penalty Ω controls head size or departure from the initial head; Appendix A.4 specifies the penalties and their coefficients. A key idea of GEAR is to use feature geometry to choose training weights ω_i and parameter constraints Θ for fitting candidate classifiers. It then selects the classifier with the lowest validation score among these candidates and the initial classifier. The procedure has three steps.

2.1 Step 1: Initial and reference classifiers.

GEAR starts with two models. The trained initial classifier $f_0(x) = W_0 h_0(x) + b_0$ is the model we seek to improve and serves as the starting point for fitting suppression and relearning candidates. A reference classifier f_{ref} assigns each example an observed-label probability, $p_i^{\mathrm{ref}} = p_{f_{\mathrm{ref}}}(y_i \mid x_i)$; lower values indicate greater difficulty. The choice of reference therefore determines the easy-to-hard ordering used to construct suppression directions and guide example weighting in relearning and validation. Earlier reweighting methods make different choices for this role. JTT uses an early-stopped model to identify training mistakes for upweighting (Liu et al., 2021), whereas AFR uses a converged model's observed-label probabilities on held-out data to weight last-layer refitting (Qiu et al., 2023). GEAR uses either a briefly trained network or a linear classifier fitted to early-stage features of the initial network as its reference to identify examples that remain difficult early in training.

2.2 Step 2: Fit suppression and relearning candidates.

Easy-to-hard feature changes may reflect misleading context, useful task information, or both. GEAR fits two corrections under Eq. (2): suppression uses its estimated directions to restrict Θ, while relearning constructs ω_i using geometric regions.

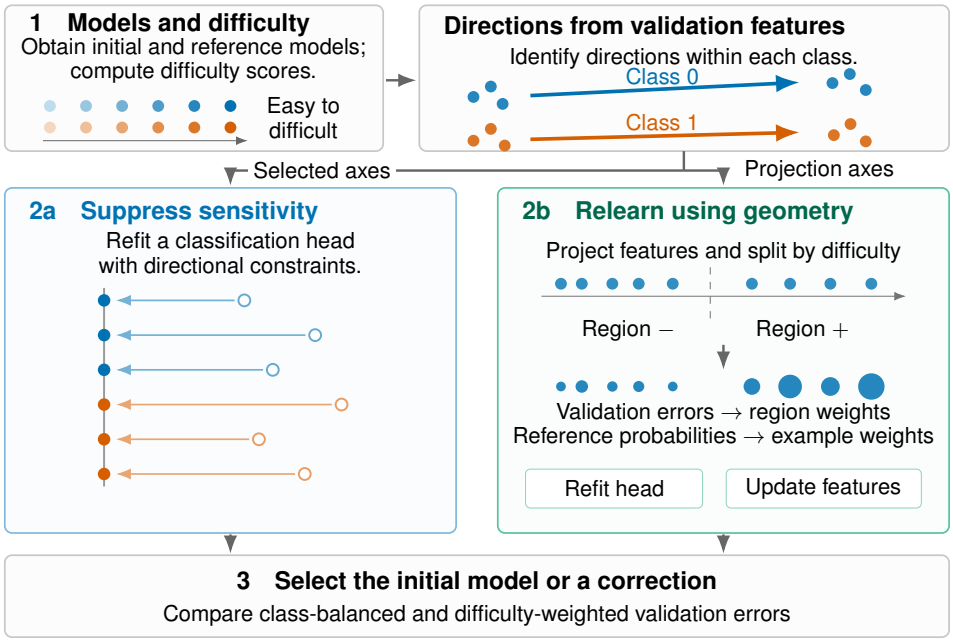

Figure 2: **Geometry-Guided Error-Aware Robustness.** (1) Train classifiers and measure example difficulty. Directions are constructed separately for the two corrections from easier to harder validation examples within each task class; blue and orange denote task classes. (2a) Suppression fits a head insensitive to the selected directions. (2b) Relearning groups examples by position along the directions and reference-model mistakes. Training weights favor regions with more validation errors and harder examples within each region. Larger dots show greater weights used to refit the classifier or update its features. (3) Validation selects among these corrections and the initial classifier.

(a) Suppress directional sensitivity. For suppression, GEAR refits the classification head under directional constraints specified by Θ. Unlike attribute-targeted concept erasure (Ravfogel et al., 2020; Belrose et al., 2023), GEAR uses task labels only to compute directions within each class and fits a classification head whose class scores remain invariant along them.

To construct suppression directions, let $\mathcal{I}^{(c)} = \{i \in \mathcal{V} : y_i = c\}$ index the class-c validation examples. Within each class, let r_i be the rank of p_i^{ref} in decreasing order, starting at one, and define the centered rank $t_i = r_i - (|\mathcal{I}^{(c)}| + 1)/2$. Positive and negative centered ranks contrast harder and easier examples within the same class. The normalized rank-weighted feature sum gives the class-c suppression direction:

$$d_c = \frac{\sum_{i \in \mathcal{I}^{(c)}} t_i h_0(x_i)}{\left\| \sum_{i \in \mathcal{I}^{(c)}} t_i h_0(x_i) \right\|_2}. \tag{3}$$

Here $h_0(x_i)$ denotes features standardized coordinatewise using training-set means and standard deviations; the initial head is transformed accordingly to preserve f_0. The unit vector d_c describes feature change from easier to harder examples within class c.

Because class-specific directions may overlap, we stack them as rows of $[d_0, \ldots, d_{C-1}]^\top$ and let U_r contain the resulting matrix's r leading right singular vectors. With $\mathcal{T}^{(c)} = \{i \in \mathcal{T} : y_i = c\}$, we minimize Eq. (2) using a ridge penalty on W, class-balanced weights, and a constrained feasible set:

$$\omega_i = \frac{1}{C|\mathcal{T}^{(y_i)}|}, \qquad \Theta = \Theta_r = \{(h, W, b) : h = h_0,\ WU_r = 0\}. \tag{4}$$

Write the fitted classifier as $f_r^{\perp}(x) = W_r h_0(x) + b_r$. Mixing its class scores with those of the initial classifier yields suppression candidates $f_{r,\alpha}(x) = (1-\alpha) f_0(x) + \alpha f_r^{\perp}(x)$, $\alpha \in [0, 1]$. Validation selects both the number of axes r and the suppression strength α.

(b) Relearn with geometric training weights. The relearning candidate uses feature geometry to place examples in fixed regions and fit classifiers with region-dependent training weights under

Eq. (2). Unlike suppression, relearning does not constrain the classifier to ignore any directions, so it can retain useful task information along them.

To define the relearning regions, let $m_i^0 = f_{0,y_i}(x_i) - \max_{k \neq y_i} f_{0,k}(x_i)$ be the initial classifier's observed-label margin. Within each class, we rank validation examples by decreasing m_i^0, center the ranks, and compute a separate relearning direction d_c using Eq. (3). We then project each class-c training example onto this direction, obtaining $d_c^\top h_0(x_i)$. The initial classifier determines the direction to examine, while the reference classifier supplies difficulty scores for splitting the projections. Let m_i^{ref} be its observed-label margin and $\xi_i = (1 + \exp m_i^{\mathrm{ref}})^{-1}$, with larger values indicating greater difficulty. For a threshold s leaving both sides nonempty, let $\bar{\xi}_{\le s}$ and $\bar{\xi}_{>s}$ be the mean difficulty scores on the two sides. We choose s_c to minimize variation in difficulty within each side:

$$s_c \in \arg\min_s \sum_{i \in \mathcal{T}^{(c)}} \begin{cases} (\xi_i - \bar{\xi}_{\le s})^2, & d_c^\top h_0(x_i) \le s, \\ (\xi_i - \bar{\xi}_{>s})^2, & d_c^\top h_0(x_i) > s. \end{cases}$$

We call the side with the higher mean difficulty score the hard side. Examples on this side misclassified by f_{ref} form the difficult region; all other class-c examples, including correctly classified examples on the hard side, form the other region. We also assign validation examples to these regions to compute the regional validation errors.

With the regions fixed, GEAR repeatedly fits weighted classifiers and updates each region's weight share using validation errors. Let R index one of the $2C$ regions, and let $\mathcal{T}_R$ and $\mathcal{V}_R$ be its training and validation examples. Before the first fit, each region receives total training-weight share $q_R^{(0)} = 1/(2C)$. For the u-th fit, the training weights are

$$\omega_i = q_R^{(u-1)} \frac{\exp(-\beta p_i^{\mathrm{ref}})}{\sum_{j \in \mathcal{T}_R} \exp(-\beta p_j^{\mathrm{ref}})}, \qquad i \in \mathcal{T}_R, \tag{5}$$

with $\beta \ge 0$. Here $q_R^{(u-1)}$ sets the total weight of region R, while β controls how strongly harder examples are favored within that region. Fitting Eq. (2) with these weights yields $f^{(u)}$. To update the regional shares for the next fit, we compute the mean soft error in each validation region, $\bar{e}_R(f) = \frac{1}{|\mathcal{V}_R|} \sum_{i \in \mathcal{V}_R} [1 - p_f(y_i \mid x_i)]$, and set $q_R^{(u)} \propto 1 + \sum_{v=1}^{u} \bar{e}_R(f^{(v)})$, with $\sum_R q_R^{(u)} = 1$.

Regions with higher accumulated validation errors receive more training weight in later fits, while region memberships and within-region relative weights remain fixed. For head relearning, Θ fixes $h = h_0$ and refits (W, b) while penalizing departures from the initial head. For representation relearning, it also permits updates to h. These choices yield f_{head} and f_{rep}, respectively. Appendix A.4 gives the fitting details.

2.3 Step 3: Candidate classifier selection by error-aware validation.

Without group annotations, GEAR selects among the initial classifier and the fitted corrections using two complementary validation errors: class-balanced error and difficulty-weighted error. Let $e_i(f) = 1 - p_f(y_i \mid x_i)$ denote the soft error of candidate f on validation example i. The two validation errors are

$$L_{\mathrm{bal}}(f) = \sum_{i \in \mathcal{V}} w_i^{\mathrm{bal}} e_i(f), \qquad L_{\mathrm{err}}(f) = \sum_{i \in \mathcal{V}} w_i^{\mathrm{err}} e_i(f),$$

Let n_c be the number of class-c validation examples. The class-balanced weights $w_i^{\mathrm{bal}} \propto n_{y_i}^{-1}$ give each task class equal total weight. The error-aware weights $w_i^{\mathrm{err}} \propto w_i^{\mathrm{bal}} \exp(-\kappa p_i^{\mathrm{ref}})$ further emphasize examples with lower reference probabilities; $\kappa \ge 0$ controls the strength of this emphasis. Both sets of weights are normalized to sum to one and remain fixed across candidates. To favor candidates that keep both errors small, we combine them with a normalized smooth maximum with smoothing parameter $\tau > 0$:

$$\widehat{R}_\tau(f) = \mathrm{smax}_\tau\{L_{\mathrm{bal}}(f), L_{\mathrm{err}}(f)\}. \tag{6}$$

GEAR returns the candidate with the lowest $\widehat{R}_\tau(f)$. This comparison uses task labels and model probabilities, not group annotations. BAM with ClassDiff and MaskTune likewise avoid group annotations in model selection (Li et al., 2024; Asgari et al., 2022). Appendix A.5.2 gives the precise formulas for the validation weights and smooth maximum.

3 EMPIRICAL EVALUATION

We evaluate GEAR on six image and text benchmarks, asking whether it improves worst-group accuracy without group annotations for training or validation. We then examine which parts of the method account for the gains and how sensitive they are to suppression and penalty strengths. Average accuracy provides a complementary measure of overall performance.

3.1 EXPERIMENTAL SETUP

Datasets. We compare GEAR to baselines on six datasets; Appendix B.1 gives sizes and groups.

- **Waterbirds and SpuCoBirds** (Sagawa et al., 2020; Joshi et al., 2023). The task is to distinguish landbirds from waterbirds. Bird type is correlated with land or water backgrounds, so a classifier relying on scenery can fail on unusual bird–background combinations.
- **CelebA** (Liu et al., 2015). The task is to predict whether a person has blond hair from a face image. Hair color is correlated with the dataset's gender annotation, making blond-haired men a minority group.
- **SpuCoDogs** (Joshi et al., 2023). The task is to distinguish small from large dogs. Small dogs appear predominantly indoors and large dogs outdoors, creating a misleading association between background and dog category.
- **CivilComments** (Koh et al., 2021). The task is to classify comments as toxic or non-toxic. Mentions of demographic identities can become misleading cues for toxicity.
- **MultiNLI** (Williams et al., 2018; Sagawa et al., 2020). The task is to determine whether a hypothesis follows from, contradicts, or is neutral to a premise. Negation words in the hypothesis are correlated with contradiction.

Models and metrics. For comparability with prior work, we use ResNet-50 (He et al., 2016) pretrained on ImageNet-1K for the image benchmarks and BERT (Devlin et al., 2019) pretrained on BookCorpus and English Wikipedia for the text benchmarks. These networks provide the feature extractor h defined in Section 2. Our primary metric is test WGA (Eq. (1)); we also report average test accuracy.

Hyperparameters. Within each run, we use validation to select the suppression rank and strength, retain relearning candidates, and choose the final classifier. Configures such as training schedules are detailed in Appendix A.

3.2 WORST-GROUP ROBUSTNESS ACROSS BENCHMARKS

We compare GEAR with methods that use only task labels during training and validation. Table 1 reports worst-group accuracy, while Table 2 reports average accuracy, allowing us to assess robustness alongside overall predictive performance. Complete comparisons, including methods with access to group annotations, are provided in Appendix B.3.

GEAR improves mean WGA over its paired ERM classifiers on all six benchmarks and achieves the highest reported mean among the task-label-only comparators on five of them (Table 1). Average accuracy rises on some benchmarks and falls on others (Table 2), so the worst-group gains do not extend uniformly to overall accuracy.

3.3 WHAT DRIVES GEAR'S ROBUSTNESS GAINS?

These ablations test the roles of learned directions, correction choice, validation selection, training weights, and suppression strength. Each varies one choice while holding the others fixed; Appendix B gives the detailed configurations.

Learned directions. Randomly rotating the learned directions while preserving their lengths and relative angles lowers mean WGA on Waterbirds and CelebA after refitting (Figure 3(a)). Their orientation in feature space carries information useful for correction in these studies.

Table 1: **Test worst-group accuracy (%).** GEAR reports mean ± sample standard deviation over five independent runs. ERM reports the same five source classifiers before GEAR's correction. We use published results when they match our settings; otherwise, we reproduce the method under those conditions. Appendix B.2 lists sources and replication counts. Bold indicates the largest mean in each dataset. "n/a" denotes an inapplicable method.

Method	Waterbirds	CelebA	SpuCo Birds	SpuCo Dogs	Civil Comments	MultiNLI
ERM	76.39 ± 2.97	30.22 ± 2.31	81.56 ± 1.52	50.60 ± 1.98	58.03 ± 1.53	64.22 ± 0.84
Class-balanced ERM	62.75 ± 2.59	65.56 ± 1.57	81.08 ± 0.50	52.28 ± 2.06	40.90 ± 4.88	66.80 ± 0.50
BAM-CM (local)	84.39 ± 2.46	80.99 ± 0.87	82.20 ± 29.07	76.44 ± 7.27	14.69 ± 6.08	35.96 ± 13.18
BAM + ClassDiff	89.10 ± 0.15	80.10 ± 3.32	93.28 ± 0.30	75.40 ± 9.12	49.44 ± 5.17	70.80 ± 1.52
CoBalT-IG	89.00	89.20	81.72 ± 3.58	49.20 ± 13.46	*n/a*	*n/a*
CCDB	90.48 ± 0.28	85.27 ± 0.28	84.76 ± 1.37	61.00 ± 1.39	44.09 ± 2.73	59.77 ± 2.23
MaskTune	56.07 ± 1.17	78.00	85.28 ± 0.30	61.84 ± 0.86	*n/a*	*n/a*
FG-CCDB	90.56 ± 0.24	$\mathbf{89.22 \pm 0.19}$	88.12 ± 1.43	65.40 ± 1.10	14.84 ± 7.59	1.90 ± 1.21
XRM	86.10 ± 1.28	88.50 ± 1.23	88.80 ± 1.17	80.52 ± 2.37	42.96 ± 4.99	72.80 ± 1.77
GEAR (5)	$\mathbf{91.06 \pm 2.77}$	88.61 ± 1.69	$\mathbf{95.12 \pm 0.41}$	$\mathbf{84.20 \pm 0.24}$	$\mathbf{68.46 \pm 1.91}$	$\mathbf{74.09 \pm 1.68}$

Table 2: **Average test accuracy (%).** Methods, predictors, and reporting conventions follow Table 1. Bold indicates the largest mean in each dataset. Dashes indicate unavailable results.

Method	Waterbirds	CelebA	SpuCo Birds	SpuCo Dogs	Civil Comments	MultiNLI
ERM	89.87 ± 1.83	$\mathbf{95.57 \pm 0.07}$	91.41 ± 0.43	76.26 ± 0.36	91.96 ± 0.04	$\mathbf{82.13 \pm 0.27}$
Class-balanced ERM	81.50 ± 1.73	93.58 ± 0.24	91.36 ± 0.07	76.69 ± 0.50	84.35 ± 0.60	81.50 ± 0.00
BAM-CM (local)	89.81 ± 1.08	87.05 ± 0.65	90.95 ± 13.34	86.32 ± 1.85	70.07 ± 4.04	62.43 ± 3.16
BAM + ClassDiff	91.40	88.40	94.51 ± 0.17	83.73 ± 2.65	86.76 ± 0.66	80.30
CoBalT-IG	92.50	92.30	90.37 ± 2.52	75.15 ± 4.90	—	—
CCDB	92.59 ± 0.10	90.08 ± 0.19	93.64 ± 0.37	80.49 ± 0.49	$\mathbf{92.01 \pm 0.06}$	78.47 ± 0.45
MaskTune	88.18 ± 0.23	91.30	93.75 ± 0.09	81.15 ± 0.16	—	—
FG-CCDB	92.50 ± 0.52	89.71 ± 0.54	94.38 ± 0.52	82.27 ± 0.74	58.00 ± 2.26	29.88 ± 2.37
XRM	90.60	91.00	93.90 ± 0.54	86.24 ± 0.74	78.70 ± 0.69	—
GEAR (5)	$\mathbf{93.64 \pm 0.78}$	91.69 ± 0.43	$\mathbf{97.35 \pm 0.07}$	$\mathbf{87.71 \pm 0.30}$	90.23 ± 0.26	79.27 ± 1.32

Suppression or relearning. Restricted variants retain only suppression, head relearning, or representation relearning, and may keep the initial classifier. Suppression alone matches GEAR in mean WGA on Waterbirds, whereas representation relearning alone matches it on CelebA (Figure 3(a)). None matches GEAR on both benchmarks, supporting validation selection between the corrections.

Validation selection. We keep the fitted candidates fixed and compare selection by $L_{\mathrm{bal}}(f)$, $L_{\mathrm{err}}(f)$, and $\widehat{R}_{\tau}(f)$. In mean WGA, the stronger single criterion is L_{err} on Waterbirds and L_{bal} on CelebA. The combined score matches the better single criterion in each study (Figure 3(a)).

Suppression strength. To examine how much suppression is useful, we vary the suppression coefficient α from 0 to 2. Here, $\alpha = 0$ retains the initial classifier, while $\alpha = 1$ completely suppresses the selected directions. Within GEAR's search range $\alpha \in [0, 1]$, worst-group error falls on both benchmarks. Beyond that range, error rises on Waterbirds, while CelebA improves briefly before it too rises (Figure 3(b)). Appendix B.5.2 also examines sensitivity to the regularization coefficient λ used when fitting the correction heads.

Training weights. To test the contribution of the geometric training weights in Eq. (5), we refit both relearning branches using class-balanced weights $\omega_i^{\mathrm{CB}} = \frac{1}{C|\mathcal{T}^{(y_i)}|}$ instead, then repeat final validation selection. The proposed weights improve mean WGA on several benchmarks but match class-balanced weights on MultiNLI (Table 3). Their contribution varies across datasets.

Supplementary masking and linear-probe studies suggest that the learned directions can aid correction without isolating context from task information (Appendix B.6). On Waterbirds (Sagawa et al., 2020),

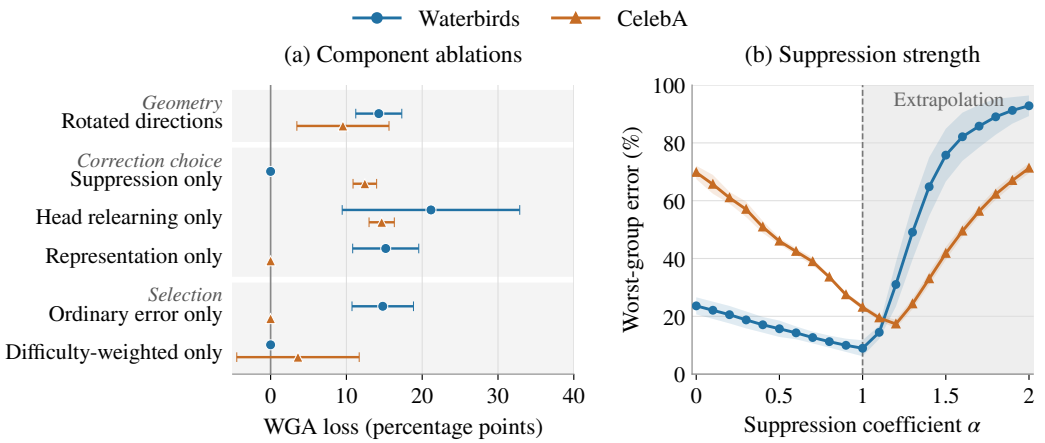

Figure 3: **Component ablations and suppression strength.** (a) The vertical axis lists the tested variants, grouped by geometry, correction choice, and validation selection. The horizontal axis shows mean paired WGA loss (percentage points) relative to each study's reference configuration; positive values indicate lower WGA than the reference. (b) The horizontal axis shows the suppression coefficient α, and the vertical axis shows mean worst-group error (%). The dashed line marks complete suppression ($\alpha = 1$); the gray region shows extrapolation beyond GEAR's range ($\alpha > 1$). Blue circles represent Waterbirds, and orange triangles represent CelebA. Error bars in (a) and colored bands in (b) show sample standard deviations over five paired runs per dataset. Experimental settings for panel (a) are given in Appendices B.4.1 and B.4.2, and those for panel (b) in Appendix B.5.1.

Table 3: **Weighting ablation of GEAR.** "Class-balanced only" replaces GEAR's geometric training weights in both relearning branches; final validation selection is unchanged. Entries report test WGA (%), mean ± sample standard deviation over five paired runs. Appendix B.4.3 gives the fitting details, average accuracies, and paired gains.

GEAR variant	SpuCoDogs	SpuCoBirds	CivilComments	MultiNLI
Class-balanced only	47.92 ± 0.33	90.32 ± 0.46	67.17 ± 2.20	74.09 ± 1.68
GEAR	84.20 ± 0.24	95.12 ± 0.41	68.46 ± 1.91	74.09 ± 1.68

they respond to both background and bird removal. This mixture of task and contextual information motivates the fixed-feature analysis in Section 4.

4 When to Suppress and When to Relearn

The empirical results in Section 3.3 show that suppression and relearning perform differently across datasets. We show that feature structure can favor either suppression or relearning. We then study when GEAR's validation criterion can select a classifier with optimal WGA.

Deep Feature Reweighting (DFR) (Kirichenko et al., 2023) suggests that learned features can retain task information even when they also encode spurious cues. Motivated by this finding, we model the features produced by the fixed extractor h_0 from Section 2.1 as $h_0(X) = Z = (Z_{\text{task}}, Z_{\text{context}})$. The features Z_{task} contain information about the class, while the features Z_{context} describe contextual attributes such as background. We consider a binary setting in which $Y \in \{-1, +1\}$ denotes the task label and $E \in \{-1, +1\}$ denotes a contextual attribute. Their four possible pairs define the groups from Section 2. We assume $\mathbb{P}(Y = 1) = 1/2$ and $\mathbb{P}(E = Y \mid Y) = p \in (1/2, 1)$, so context usually agrees with the label. To examine how task and contextual signals affect WGA in a tractable setting, we assume the additive Gaussian feature model $Z_{\text{task}} = Y\mu + \xi$ and $Z_{\text{context}} = E\nu + \zeta$. Here $\mu \in \mathbb{R}^{d_{\text{task}}}$ and $\nu \in \mathbb{R}^{d_{\text{context}}}$ are fixed nonzero signal vectors. The noise vectors satisfy $\xi \sim \mathcal{N}(0, \Sigma_{\text{task}})$ and $\zeta \sim \mathcal{N}(0, \Sigma_{\text{context}})$, with both covariance matrices nonsingular. We also assume $\xi \perp\!\!\!\perp \zeta$ and $(\xi, \zeta) \perp\!\!\!\perp (Y, E)$, where $\perp\!\!\!\perp$ denotes statistical independence. Write P for the joint law of (Z, Y, E) under this model and $\text{WGA}_P(f)$ for the worst-group accuracy of f under P.

A linear classifier on Z with head parameter $w = (u, v) \neq 0$ is defined by the score $f_w(Z) = u^\top Z_{\text{task}} + v^\top Z_{\text{context}}$ and predicts $\operatorname{sign} f_w(Z)$. We next show that neither suppression nor relearning uniformly outperforms the other across distributions, motivating GEAR to retain both corrections.

Proposition 4.1 (Neither correction uniformly dominates)**.** *There exist two joint distributions P_1, P_2 of (Z, Y, E) satisfying the Gaussian feature model for which the same correction settings give*

$$\max_{f \in \mathcal{F}_{\text{sup}}} \text{WGA}_{P_1}(f) > \max_{f \in \mathcal{F}_{\text{rel}}} \text{WGA}_{P_1}(f), \qquad \max_{f \in \mathcal{F}_{\text{rel}}} \text{WGA}_{P_2}(f) > \max_{f \in \mathcal{F}_{\text{sup}}} \text{WGA}_{P_2}(f). \tag{7}$$

Here $\mathcal{F}_{\text{sup}}$ and $\mathcal{F}_{\text{rel}}$ denote the classes of suppression and head-relearning classifiers fitted under the corresponding distribution.

Proposition 4.1 (proved in Appendix C.2) shows that committing to one correction method across distributions can miss a better classifier. This motivates GEAR to retain both corrections and use validation to choose between them. We next give conditions under which GEAR's validation step yields asymptotically optimal WGA among linear classifiers. Let $\mathcal{F} = \{f_w : w \neq 0\}$ be the class of zero-intercept linear classifiers, $\mathcal{F}_n \subseteq \mathcal{F}$ GEAR's candidate set, and $\widehat{f} \in \arg\min_{f \in \mathcal{F}_n} \widehat{R}_\tau(f)$ the selected classifier.

Theorem 4.2 (Worst-group accuracy of validation selection)**.** *Set $[x]_+ = \max\{x, 0\}$. For any comparison scale $c > 0$, define $g_c(Z) = c\,\mu^\top \Sigma_{\text{task}}^{-1} Z_{\text{task}}$. Then*

$$\max_{f \in \mathcal{F}} \text{WGA}(f) - \text{WGA}(\widehat{f}) \leq C_{\text{model}} \sqrt{\varepsilon_{\text{fit}} + \varepsilon_{\text{ref}} + \varepsilon_{\text{val}}}, \tag{8}$$

$$\varepsilon_{\text{fit}} = \min_{g \in \mathcal{F}_n} \|g - g_c\|_\Sigma + [\|\widehat{f}\|_\Sigma - \|g_c\|_\Sigma]_+, \quad \varepsilon_{\text{val}} = \sup_{g \in \mathcal{F}_n} |\widehat{R}_\tau(g) - R_\tau(g)|,$$

$$\varepsilon_{\text{ref}} = [b_\kappa \widehat{f}((0, \nu))]_+ + \kappa[-\operatorname{Cov}(\widehat{f}(Z), f_{\text{ref}}(Z) \mid Y, E)]_+ + \kappa c[f_{\text{ref}}((\mu, 0))]_+.$$

Here $\|g\|_\Sigma^2 = \operatorname{Var}(g(Z) \mid Y, E)$, $R_\tau(f)$ is the population counterpart of $\widehat{R}_\tau(f)$, obtained by replacing the within-class validation averages with their conditional expectations, and $b_\kappa \in \mathbb{R}$ is a constant depending only on the model distribution, the reference classifier, and κ. The constant $C_{\text{model}} > 0$ depends only on the signals (μ, ν), noise covariances $(\Sigma_{\text{task}}, \Sigma_{\text{context}})$, reference weighting strength κ, and comparison scale c.

The fitting error ε_{fit} reflects the candidates' approximation power and the selected classifier's excess within-group variability, and vanishes as $\widehat{f}$ approaches g_c in the Σ-norm. The validation error ε_{val} measures uniform sampling error over $\mathcal{F}_n$ and converges to zero in probability as each class's validation sample size grows under the regularity conditions. The reference error ε_{ref} captures validation bias from context imbalance and reference weighting, and is zero when the selected classifier recovers the optimal task-only score g_c and either reference weighting is disabled or the reference does not use the task signal. Here the scale c changes the magnitude of g_c, but not its predictions, and is used only in the analysis.

Under suitable fitting and reference conditions, all three errors converge to zero in probability for some fixed $c > 0$ as the fitting and validation samples grow. Theorem 4.2 then shows that GEAR approaches the best WGA among all linear classifiers. The proof of Theorem 4.2, the validation sampling bound, and sufficient conditions for the three error terms to vanish are given in Appendices C.3, C.4, and C.5, respectively.

5 Discussion

GEAR answers the question of how to improve a classifier after identifying difficult examples by linking their errors to feature variation. The resulting directions guide sensitivity suppression and define regions for relearning. Validation then compares the fitted classifiers. Our experiments show the value of this choice: which correction works better depends on the dataset.

GEAR currently uses linear directions in an existing feature space. Higher-dimensional or curved models of feature variation could describe more complex contexts, and sharing computation across candidates could reduce fitting costs. Our theory is limited to linear classifiers under a simplified Gaussian feature model, but it captures the trade-off between suppressing misleading context and preserving useful task information, explaining why neither correction is uniformly preferable. GEAR also depends on the reference model highlighting errors relevant to group robustness. Selecting such references using task labels alone remains an important practical challenge.

Reproducibility statement. Code, configurations, per-run results, and table and figure scripts are available at https://anonymous.4open.science/r/gear-review-FC09. See `reproduction/README.md` for entry points and data and model requirements. Numerical checks and table generation use the included records without downloading datasets or pretrained weights.

ETHICS STATEMENT

CelebA is an imperfect robustness benchmark. Its annotations do not imply that gender is binary or visually determined; its `Male` attribute is used only as the provided evaluation coordinate.

Use of large language models. Large language models assisted writing and derivations; the authors remain responsible for all statements, proofs, and numerical claims.

REFERENCES

Md Rifat Arefin, Yan Zhang, Aristide Baratin, Francesco Locatello, Irina Rish, Dianbo Liu, and Kenji Kawaguchi. Unsupervised concept discovery mitigates spurious correlations. In *International Conference on Machine Learning (ICML)*, pp. 1672–1688, 2024.

Saeid Asgari, Aliasghar Khani, Fereshte Khani, Ali Gholami, Linh Tran, Ali Mahdavi Amiri, and Ghassan Hamarneh. MaskTune: Mitigating spurious correlations by forcing to explore. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 23284–23296, 2022.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, and Karol Zieba. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4171–4186, 2019.

Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580, 2018.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Yujin Han and Difan Zou. Improving group robustness on spurious correlation requires preciser group inference. In *International Conference on Machine Learning (ICML)*, pp. 17480–17504, 2024.

D. J. Hand and W. E. Henley. Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(3):523–541, 1997.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew Gordon Wilson. On feature learning in the presence of spurious correlations. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 38516–38532, 2022.

Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology*, 2(4):230–243, 2017.

Siddharth Joshi, Yu Yang, Yihao Xue, Wenhan Yang, and Baharan Mirzasoleiman. Towards mitigating spurious correlations in the wild: A benchmark & a more realistic dataset. *arXiv preprint arXiv:2306.11957v1*, 2023.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations (ICLR)*, 2023.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, pp. 5637–5664, 2021.

Gaotang Li, Jiarui Liu, and Wei Hu. Bias amplification enhances minority group performance. *Transactions on Machine Learning Research*, 2024.

Evan Z. Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning (ICML)*, pp. 6781–6792, 2021.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3730–3738, 2015.

Mohammad Pezeshki, Diane Bouchacourt, Mark Ibrahim, Nicolas Ballas, Pascal Vincent, and David Lopez-Paz. Discovering environments with XRM. In *International Conference on Machine Learning (ICML)*, pp. 40551–40569, 2024.

Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. Simple and fast group robustness by automatic feature reweighting. In *International Conference on Machine Learning (ICML)*, pp. 28448–28467, 2023.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

Nguyen Nhat Minh To, Paul F. R. Wilson, Viet Nguyen, Mohamed Harmanani, Michael Cooper, Fahimeh Fooladgar, Purang Abolmaesumi, Parvin Mousavi, and Rahul Krishnan. Diverse prototypical ensembles improve robustness to subpopulation shift. In *International Conference on Machine Learning (ICML)*, pp. 59761–59783, 2025.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 1112–1122, 2018.

Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. Change is hard: A closer look at subpopulation shift. In *International Conference on Machine Learning (ICML)*, pp. 39584–39622, 2023.

Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2:527–550, 2002.

Miaoyun Zhao and Qiang Zhang. Fine-grained class-conditional distribution balancing for debiased learning. In *International Conference on Learning Representations (ICLR)*, 2026.

Miaoyun Zhao, Chenrong Li, and Qiang Zhang. Class-conditional distribution balancing for group robust classification. *arXiv preprint arXiv:2504.17314*, 2025.

A METHOD AND IMPLEMENTATION DETAILS

This section gives the complete procedure, fitting settings, and validation rules for GEAR in Section 2.

A.1 ALGORITHM OVERVIEW

Algorithm 1 GEAR

Step 1: Initial and reference classifiers
Fit f_0 and f_{ref}; keep both fixed and compute $p_i^{\mathrm{ref}} = p_{f_{\mathrm{ref}}}(y_i \mid x_i)$.
Step 2: Fit suppression and relearning candidates
(a) Suppression
Rank validation examples within class by p_i^{ref}; compute suppression directions d_c by Eq. (3).
For each r, obtain U_r, fit $f_r^{\perp}$ with $h = h_0$ and $WU_r = 0$ (Eq. (4)), and form $f_{r,\alpha} = (1-\alpha)f_0 + \alpha f_r^{\perp}$.
(b) Relearning
Rank validation examples within class by the initial margins m_i^0; recompute d_c by Eq. (3).
Split projections along d_c by reference difficulty; mark mistakes on the hard side (Appendix A.3.2).
for each relearning mode and within-region weighting strength **do**
    Start with equal regional shares; favor lower-reference-probability examples.
    for each fitting round **do**
        Fit a candidate with the current weights (Eq. (2)).
        Update regional shares from accumulated mean validation soft errors.
    end for
end for
Step 3: Select a classifier
return the candidate minimizing $\widehat{R}_\tau(f)$ among f_0, $f_{r,\alpha}$, f_{head}, and f_{rep} (Eq. (6)).

A.2 INITIAL AND REFERENCE CLASSIFIERS

Image models use ImageNet-1K-pretrained ResNet-50 with 2,048-dimensional features; text models use pretrained BERT with 768-dimensional features. Table 4 specifies source training. All sources minimize ordinary cross-entropy. For the image SGD fits, the learning rate is $\eta_{\mathrm{img}} = 5.5828866 \times 10^{-4}$, momentum is 0.9, and weight decay is 4.0511994×10^{-4}.

Table 4: Source-model training settings. SpuCo denotes both binary SpuCo datasets. ERM reports these models before correction.

Dataset	Duration	Optimizer / rate	Batch	Checkpoint criterion
Waterbirds	12 epochs	AdamW; backbone 10^{-4}, head 10^{-3}	64	Minimum class-balanced validation cross-entropy
CelebA, SpuCo	1,001 updates	SGD, η_{img}	128	Minimum mean validation cross-entropy
CivilComments	5 epochs	AdamW, 10^{-5}	16	Final epoch
MultiNLI	3 epochs	AdamW, 2×10^{-5}	32	Final epoch

Waterbirds uses weight decay 10^{-4}. CivilComments uses token length 300, weight decay 0.01, linear learning-rate decay, mixed precision, and gradient clipping at norm one. MultiNLI uses token length 128 and zero weight decay. Image source seeds are $1{,}884{,}001$ plus run index; source initialization and data order are independent across runs.

The reference classifiers supply the observed-label probabilities p_i^{ref}. Waterbirds uses its first-epoch source features with a fitted linear classifier. CelebA and both SpuCo datasets use a fixed 50-update ResNet-50 reference trained with 64 examples per task class per batch, the image SGD settings above, random resized crops, and horizontal flips. CivilComments trains a separate 50-update BERT reference in each run, with eight examples per task class per batch and its source optimizer settings, including learning-rate decay over the 50 updates. MultiNLI selects a full-model BERT reference from 50, 100, or 500 updates by class-balanced validation cross-entropy; batches contain 16 examples per task class, with AdamW at 2×10^{-5}, linear decay, norm-one clipping, and mixed precision.

Feature extraction uses evaluation mode with gradients disabled. Fitting-set means and standard deviations standardize each coordinate, with standard deviations bounded below by 10^{-6}; source heads are transformed into the same coordinates.

A.3 Direction and region construction

A.3.1 Suppression directions

For the within-class ranks in Eq. (3), ties receive average ranks. With $\mathbf{1}\{\cdot\}$ denoting an indicator,

$$t_i = 1 + \sum_{j \in \mathcal{I}^{(c)}} \mathbf{1}\{a_j > a_i\} + \frac{1}{2} \sum_{j \in \mathcal{I}^{(c)} \setminus \{i\}} \mathbf{1}\{a_j = a_i\} - \frac{|\mathcal{I}^{(c)}| + 1}{2}. \tag{9}$$

A zero feature sum stops the construction. Singular vectors are ordered by decreasing singular value, using a fixed numerical SVD convention for nonunique vectors. Suppression fits every rank $r = 1, \ldots, C$ using source features and reference probabilities.

A.3.2 Relearning regions

Region directions instead rank the observed-label margins of a linear proxy on source-network features. Image regions use standardized ResNet features. CivilComments uses unpooled classification-token features for regions and pooled features for suppression and anchored-head fitting. Its auxiliary linear proxy supplies direction ranks. MultiNLI uses an ordinary multinomial head for direction ranks and an auxiliary early linear probe for the split; its full-model reference still supplies p_i^{ref}.

Let v_c be the projection axis for class c. We use

$$v_c = \begin{cases} v_{\mathrm{shared}}, & \text{if the class directions align,} \\ d_c, & \text{otherwise,} \end{cases} \tag{10}$$

where v_{shared} maximizes $\sum_{j=0}^{C-1} (v^\top d_j)^2$ over unit vectors v. The shared axis is used when $\min_{j<k} |d_j^\top d_k| \geq \chi$: $\chi = 0$ for CivilComments and $\chi = \sqrt{1/2}$ otherwise.

Project the training features onto v_c. Let f_{reg} be the split classifier: the reference image/BERT classifier for images and CivilComments, or the early linear probe for MultiNLI. Writing $f = f_{\mathrm{reg}}$, it supplies bounded difficulty scores

$$e_i = \mathrm{error}\big(f(x_i), y_i\big) \in [0, 1]. \tag{11}$$

Larger e_i indicates a harder example. Its margin and bounded score are

$$m_i^{\mathrm{reg}} = f_{\mathrm{reg},y_i}(x_i) - \max_{c \neq y_i} f_{\mathrm{reg},c}(x_i), \qquad e_i = \frac{1}{1 + \exp(m_i^{\mathrm{reg}})}.$$

Let $\mathcal{P}_c(s)$ contain the two sets of class-c training examples with projections $v_c^\top z_i \leq s$ and $v_c^\top z_i > s$. Choose the split threshold

$$s_c \in \arg\min_s \sum_{B \in \mathcal{P}_c(s)} \sum_{i \in B} (e_i - \bar{e}_B)^2, \tag{12}$$

where $\bar{e}_B$ is the mean difficulty score on side B. Evaluate Eq. (12) at midpoints between successive distinct training projections with two nonempty sides. Ties choose the smallest midpoint. The side with larger mean training difficulty is hard; equal means choose the upper side. Region $(c, +)$ contains examples misclassified by f on that side, using $m_i^{\mathrm{reg}} < 0$ for a mistake; for a unique majority task class in a binary task, it contains the entire hard side. Region $(c, -)$ contains the remaining examples. Apply these thresholds and membership rules to validation examples. The binary majority-class adjustment is absent when class counts tie; MultiNLI uses the intersection of hard geometry and mistakes in every class. Stop if no threshold exists or a final region is empty in fitting or validation data. Both relearning modes use these fixed memberships.

A.4 CORRECTION FITTING

Head fitting uses the complete training inventory except on MultiNLI, where it uses the adaptation subset. Binary solvers store one class-one-minus-class-zero logit; ridge and anchor coefficients apply to that slope and intercept. In centered two-row softmax coordinates, $W = (-w/2, w/2)^\top$ gives $\|W\|_F^2 = \|w\|^2/2$, so doubling the coefficient preserves the binary objective, including the intercept anchor. Multiclass fits use one row per task class.

A.4.1 SUPPRESSION

Suppression uses class-balanced weights $\omega_i = 1/(C|\mathcal{T}^{(y_i)}|)$, where $\mathcal{T}^{(c)} = \{i \in \mathcal{T} : y_i = c\}$. Its ridge is 0.05 on Waterbirds, 0.003 on MultiNLI, and 0.1 elsewhere. After fitting, binary thresholds maximize minimum task-class validation accuracy; multiclass fits use additive class-intercept calibration.

A.4.2 HEAD RELEARNING

For head relearning, write $\mathrm{CE}(f(x_i), y_i) = -\log p_f(y_i \mid x_i)$ and solve

$$\begin{aligned}\min_{W,b} \quad & \sum_{i\in\mathcal{T}} \omega_i \,\mathrm{CE}(Wh_0(x_i) + b, y_i) \\ & + \lambda\{\|W - W_0\|_F^2 + \|b - b_0\|^2\}.\end{aligned} \tag{13}$$

L-BFGS fits eight heads with $\lambda = 0.1$ and within-region strength $\beta = 8$, each initialized at the source. Between fits, validation soft errors update the region shares. With $\mathcal{S}_{c,\sigma}$ indexing a validation region, the update is

$$q_{c,\sigma}^{(t)} \propto D_0 + \sum_{u=1}^{t} \frac{1}{|\mathcal{S}_{c,\sigma}|} \sum_{i\in\mathcal{S}_{c,\sigma}} \left[1 - p_{f^{(u)}}(y_i \mid x_i)\right], \tag{14}$$

where $D_0 = 1$ and $\sum_{c,\sigma} q_{c,\sigma}^{(t)} = 1$. Retain the head minimizing $\widehat{R}_\tau$, breaking ties by greater class-balanced validation accuracy and then the earlier fit.

A.4.3 REPRESENTATION RELEARNING

Images and CivilComments start from the pretrained backbone and update the encoder and head for 3,001 steps. Images use SGD with the settings above and ordinary batches of 122, except CelebA uses 128. CivilComments uses ordinary batches of 16 and its source AdamW settings. Both set the explicit head penalty in Eq. (2) to zero. Validation is measured every 50 updates, with a 50-update warm-up before accumulating regional errors. The checkpoint maximizes minimum observable-region validation accuracy, then average accuracy, then update number. Training seeds are $3{,}301{,}000$ plus run index for images and $4{,}604{,}000$ plus run index for CivilComments.

MultiNLI freezes the source encoder and performs eight weighted multinomial-head refits, with ridge $0.1\|W\|_F^2$, an unpenalized intercept, and at most 140 L-BFGS iterations per refit. The first fit starts from the ordinary head; later fits warm-start from the preceding fit. Checkpoints maximize minimum observable-region accuracy, then class-balanced accuracy, then iteration number.

Within each region, weights are normalized over its complete fitting sample. Ordinary-batch losses correct for the probability that a region appears in the batch; Appendix A.4.4 gives the estimator and proof. Compare $\beta \in \{0, 2, 4, 8\}$ using the same initialization and data order and separate regional-error accumulators. Select the retained predictor by ordinary class-balanced validation soft error, with smaller β breaking ties. CelebA uses $\beta = 0$ for all five reported runs.

A.4.4 REGIONAL MINIBATCH ESTIMATOR

Let θ collect the trainable parameters among (h, W, b), with score function f_θ. Let $R_{c,\sigma}(\theta)$ be the mean training cross-entropy in region (c, σ). At a fixed iteration, D collects the positive accumulated scores in Eq. (14) and q their normalized regional shares. Sums over (c, σ) include all task classes and both region signs. First take equal weights within each region.

For a batch B of size b_{batch} sampled uniformly without replacement from N training examples, let $N_{c,\sigma}$ be the full size of region (c,σ), $I_{c,\sigma}$ indicate whether it appears in the batch, and $\bar{\ell}_{B,c,\sigma}$ be its mean batch loss when present. Its exact inclusion probability and the corrected loss are

$$\pi_{c,\sigma} = 1 - \frac{\binom{N-N_{c,\sigma}}{b_{\text{batch}}}}{\binom{N}{b_{\text{batch}}}}, \qquad \widehat{L}_B(\theta, q) = \sum_{c,\sigma} q_{c,\sigma} \frac{I_{c,\sigma}}{\pi_{c,\sigma}} \bar{\ell}_{B,c,\sigma}. \tag{15}$$

For a binary task, order the four regions by $(0,+),(0,-),(1,+),(1,-)$ and let Δ_4 be the set of their nonnegative weights summing to one.

Proposition A.1 (Worst-region training identities)**.** *For positive accumulated regional scores $D \in \mathbb{R}^4$, the unique maximizer of $q^\top D - (\|D\|_1/2)\|q\|_2^2$ over $q \in \Delta_4$ is $q^\star = D/\|D\|_1$. Conditional on fixed q, a uniform batch satisfies $\mathbb{E}\{\widehat{L}_B(\theta, q)\} = \sum_{c,\sigma} q_{c,\sigma} R_{c,\sigma}(\theta)$.*

The same identities hold for any finite number of nonempty regions.

Proof. Writing $q^\star = D/\|D\|_1$ completes the first objective to $-(\|D\|_1/2)\|q - q^\star\|_2^2$ plus a constant. For the second identity, exchangeability gives $\mathbb{E}(\bar{\ell}_{B,c,\sigma} \mid I_{c,\sigma} = 1) = R_{c,\sigma}$ and $\Pr(I_{c,\sigma} = 1) = \pi_{c,\sigma}$; summing $q_{c,\sigma} I_{c,\sigma} \bar{\ell}_{B,c,\sigma}/\pi_{c,\sigma}$ proves the claim. □

For confidence-weighted fitting, let $\nu_i(\beta)$ denote the normalized within-region factor in the training weights:

$$\nu_i(\beta) = \frac{\exp(-\beta p_i^{\text{ref}})}{\sum_{j \in \mathcal{T}_{c,\sigma}} \exp(-\beta p_j^{\text{ref}})}, \qquad i \in \mathcal{T}_{c,\sigma}.$$

The multiplier $N_{c,\sigma}\nu_i(\beta)$ has mean one over its complete training region. For a region present in batch B, use

$$\begin{aligned} \bar{\ell}^{(\beta)}_{B,c,\sigma}(\theta) &= \frac{1}{|B \cap \mathcal{T}_{c,\sigma}|} \sum_{i \in B \cap \mathcal{T}_{c,\sigma}} N_{c,\sigma}\nu_i(\beta)\, \mathrm{CE}(f_\theta(x_i), y_i), \\ \widehat{L}^{(\beta)}_B(\theta, q) &= \sum_{c,\sigma} q_{c,\sigma} \frac{I_{c,\sigma}}{\pi_{c,\sigma}} \bar{\ell}^{(\beta)}_{B,c,\sigma}(\theta). \end{aligned} \tag{16}$$

An absent region contributes zero. For fixed parameters and weights and a uniform batch, the same conditional-mean argument gives

$$\mathbb{E}\{\widehat{L}^{(\beta)}_B(\theta, q)\} = \sum_{c,\sigma} q_{c,\sigma} \sum_{i \in \mathcal{T}_{c,\sigma}} \nu_i(\beta)\, \mathrm{CE}(f_\theta(x_i), y_i).$$

A.5 Calibration and validation selection

A.5.1 Calibration

After choosing the anchored head, hold its slopes fixed and calibrate intercepts along

$$b_c(t) = b_c + t\big(\log \widehat{\pi}_c - \log \widehat{\pi}_0\big), \qquad t \in \{-1, -0.99, \ldots, 1\},$$

where $\widehat{\pi}_c$ is the validation frequency of task class c. Maximize validation macro-F_1 subject to mean task-class accuracy decreasing by at most 0.005. Ties prefer higher mean task-class accuracy, smaller $|t|$, then smaller t. This calibration applies on every dataset.

A.5.2 Final validation criterion

Let $n_c = |\{i \in \mathcal{V} : y_i = c\}|$ be the sample size for class c in the validation set. For difficulty strength $\kappa \geq 0$, define the normalized validation soft error

$$\widehat{L}_{\text{val}}(f; \kappa) = \frac{\displaystyle\sum_{i \in \mathcal{V}} \frac{\exp(-\kappa p_i^{\text{ref}})}{n_{y_i}} \big[1 - p_f(y_i \mid x_i)\big]}{\displaystyle\sum_{i \in \mathcal{V}} \frac{\exp(-\kappa p_i^{\text{ref}})}{n_{y_i}}}. \tag{17}$$

At $\kappa = 0$, this averages errors equally across task classes; positive κ emphasizes examples with smaller reference correct-label probabilities. For smoothing parameter $\tau > 0$, the validation loss used to compare corrections in Algorithm 1 is

$$\widehat{R}_\tau(f) = \tau \log \frac{\exp(\widehat{L}_{\mathrm{val}}(f;0)/\tau) + \exp(\widehat{L}_{\mathrm{val}}(f;\kappa)/\tau)}{2}. \tag{18}$$

For each suppression rank, choose $\alpha \in \{0, 1/4, 1/2, 2/3, 3/4, 1\}$ by $\widehat{R}_\tau$, preferring smaller α in a tie. Compare these predictors with the source, calibrated head, and retained representation predictor using $\tau = 0.1$ and $\kappa = 8$. Final ties prefer larger mean task-class accuracy, smaller fitting capacity (source 0, fitted head 1, full network 2), then the fixed candidate name.

A.5.3 Properties of the validation criterion

For the derivation below, write $A(f) = \widehat{L}_{\mathrm{val}}(f;0)$ and $B(f) = \widehat{L}_{\mathrm{val}}(f;\kappa)$. For the constrained formulation, index the source, suppression, head-relearning, and representation-relearning families by $j \in \{0, S, H, R\}$, respectively. Let $\mathcal{C}_j$ contain the fitted predictors from family j retained for the final comparison, and let $\mathcal{F} = \bigcup_j \mathcal{C}_j$. Define the constraint indicator $\iota_{\mathcal{C}_j}(f)$ to be zero for $f \in \mathcal{C}_j$ and $+\infty$ otherwise.

The common validation objective has the following worst-mixture form:

$$\widehat{R}_\tau(f) = \max_{\rho \in [0,1]} \left\{ \rho A(f) + (1-\rho)B(f) - \tau\big[\rho \log(2\rho) + (1-\rho)\log(2(1-\rho))\big] \right\}. \tag{19}$$

Proposition A.2 (Unified correction and worst-mixture risk)**.** *For $\tau > 0$, define $K(\rho) = \rho \log(2\rho) + (1-\rho)\log(2(1-\rho))$, with $0 \log 0 = 0$. Then Eq. (19) holds, with unique maximizing weight*

$$\rho^\star = \frac{\exp(A(f)/\tau)}{\exp(A(f)/\tau) + \exp(B(f)/\tau)}.$$

Consequently, the minimum validation score over $\mathcal{F}$ equals

$$\min_{j,f} \max_{\rho \in [0,1]} \{\rho A(f) + (1-\rho)B(f) - \tau K(\rho) + \iota_{\mathcal{C}_j}(f)\}.$$

Moreover, the selected score has no larger $\widehat{R}_\tau$ than the source score among the evaluated candidates.

Proof. For fixed $A = A(f)$ and $B = B(f)$, differentiation of the inner objective gives $A - B - \tau \log\{\rho/(1-\rho)\}$. Its second derivative is $-\tau\{1/\rho + 1/(1-\rho)\} < 0$, so its unique maximum occurs at the displayed $\rho^\star$. Substitution gives $\tau \log\{(\exp(A/\tau) + \exp(B/\tau))/2\}$. For each j, adding the indicator restricts f to $\mathcal{C}_j$. Minimizing over j is therefore minimizing over the union of these families. Finally, the source is explicitly included in the numerical comparison. □

The objective has positive derivatives $\partial \widehat{R}_\tau / \partial A = \rho^\star$ and $\partial \widehat{R}_\tau / \partial B = 1 - \rho^\star$. Its approximation to the hard maximum satisfies

$$\max\{A, B\} - \tau \log 2 \le \widehat{R}_\tau(f) \le \max\{A, B\}, \qquad \lim_{\tau \downarrow 0} \widehat{R}_\tau(f) = \max\{A, B\}.$$

B Experimental Protocol and Supplementary Results

The following results extend the benchmark comparison to all information tiers and provide paired gains and controlled studies supporting Section 3. Paired GEAR studies use five source models and report means and sample standard deviations.

B.1 Datasets, evaluation, and replication

We use the official training, validation, and test inventories in Table 5. GEAR uses inputs and task labels for fitting and validation selection. Group annotations are used only for official-test evaluation of the selected predictors. MultiNLI divides its training inventory into 80% for source training and 20% for head adaptation, so replacement heads use examples unseen during source training.

Table 5: Exact evaluation inventories. MultiNLI uses the group-robustness partition. Its training inventory is divided into 164,940 source-fitting and 41,235 head-adaptation examples.

Dataset	Training	Validation	Test	Task classes
Waterbirds	4,795	1,199	5,794	2
CelebA	162,770	19,867	19,962	2
SpuCoBirds	21,000	1,050	2,000	2
SpuCoDogs	21,000	1,050	2,000	2
CivilComments	269,038	45,180	133,782	2
MultiNLI	206,175	82,462	123,712	3

Waterbirds groups cross bird type with background; CelebA groups cross hair color with the provided `Male` attribute. Each SpuCo test inventory has 500 examples per task–background group. MultiNLI has six label–negation groups. CivilComments WGA uses sixteen overlapping identity–label groups from eight identity annotations. Its average accuracy counts each of the 133,782 test comments once.

GEAR and ERM use the same five source models, indexed 0–4. Accuracies are computed per run and summarized by their mean and sample standard deviation; paired changes are computed before averaging. Local five-run summaries use the first five runs in the specified seed order, selected by identity. SpuCo AFR and JTT use three runs; published entries retain their reported counts and uncertainty. Averaging within a comparator's algorithm is retained.

B.2 Comparator sources and information access

T0 permits task labels only in training and validation; T1 additionally permits validation groups; T2 permits both training and validation groups. These tiers include annotation-dependent preprocessing. Comparisons use the same benchmark sources and test WGA definitions, with each method's native fitting procedure. Table 6 identifies published and locally reproduced entries. Image and language comparisons use ImageNet-1K and BERT pretraining, respectively.

Table 6: Sources for the full comparison in Table 7. "SpuCo" denotes both binary SpuCo datasets. Local entries use five independent runs, except SpuCo JTT and AFR use three.

Method	Published entries and source	Local entries	Tier
ERM, BAM-CM	—	All six	T0
Class-balanced ERM	MultiNLI: ReSample, Yang et al. (2023), Table 3	Other five	T0
BAM + ClassDiff	Waterbirds, CelebA, MultiNLI: Li et al. (2024), Tables 1–2	SpuCo, CivilComments	T0
CoBalT-IG	Waterbirds, CelebA: Arefin et al. (2024)	SpuCo	T0
CCDB	Waterbirds, CelebA: Zhao et al. (2025), Table 2	Other four	T0
MaskTune	CelebA: Asgari et al. (2022), Table 1	Waterbirds, SpuCo	T0
FG-CCDB	Waterbirds, CelebA: Zhao & Zhang (2026), Table 1	Other four	T0
XRM	Waterbirds, CelebA, MultiNLI: Pezeshki et al. (2024), Table 2 and Appendix Table 5	SpuCo, CivilComments	T0
DPE	Waterbirds, CelebA, MultiNLI: To et al. (2025), Table 2	SpuCo, CivilComments	T1
GIC	Waterbirds, CelebA, CivilComments: Han & Zou (2024), Table 1	SpuCo, MultiNLI	T1
JTT, AFR, DFR	Waterbirds, CelebA, CivilComments, MultiNLI: Liu et al. (2021); Qiu et al. (2023), Table 1; Kirichenko et al. (2023), Table 2	SpuCo	T1
GroupDRO	Waterbirds, CelebA, CivilComments, MultiNLI: Liu et al. (2021), Table 1	SpuCo	T2
XRM / DPE preprocessing variants	CivilComments: Pezeshki et al. (2024); To et al. (2025), Table 2	—	T1 / T2

Published AFR and BAM results use three runs; DFR, CCDB, and FG-CCDB use five; XRM uses ten; DPE and the MultiNLI ReSample entry use three. Other entries retain the source's available point estimates and uncertainty. DFR averages ten group-balanced validation heads within each run; DPE uses its native 15-head prototype ensemble. Local AFR fits the source on 80% of training data and reweights the head on the remaining 20%. JTT fits a fresh network after upweighting early mistakes. These T1 methods use validation groups for fitting or selection as prescribed by their respective algorithms.

CivilComments and dataset compatibility. Published XRM filters and repeats validation comments using protected identities, so this entry is classified as T1. Published DPE uses identity-based training and validation preprocessing, giving T2 access. Local T0 XRM uses all validation comments without identity-dependent filtering; local T1 DPE uses sources trained on all training comments and group-balanced validation prototypes. All use the sixteen-group test metric. Published BAM, CCDB, and FG-CCDB CivilComments results use four coarse groups and are excluded; local reproductions provide the aligned entries. The GroupDRO reference trains with four coarse groups but evaluates the sixteen overlapping groups. MaskTune's published Waterbirds inventory changes images and labels, so the table uses a local reproduction on the original benchmark. MaskTune and CoBalT have no native language implementation; these cells are marked n/a.

Local implementation conventions. XRM retains two phases of 10,001 updates on SpuCo and 60,001 updates on CivilComments; the latter uses the first fixed native hyperparameter draw. FG-CCDB follows its printed weighting equations and four bias-model stages, then weighted sampling with ordinary cross-entropy for 5,000 updates, including the printed additional inverse confusion-cell factor. Its published image entries lack released five-run configuration bindings. BAM retains bias amplification, error resampling, and ClassDiff selection; local BAM-CM uses 1,001 updates. CCDB uses full-covariance balancing, GIC uses selective Mixup, and CoBalT-IG uses a paper-specification implementation of spatial slots, teacher distillation, and concept sampling. MaskTune uses predicted-class XGradCAM and one extra training epoch, with 256×256 resizing, a 224×224 center crop, and nearest-neighbor mask mapping. Code, complete configurations, seeds, and per-run results are indexed in the anonymous repository's reproduction guide.

B.3 Benchmark comparisons and selected corrections

Tables 7 and 8 report WGA and average accuracy across the information tiers defined in Appendix B.2, including comparisons with AFR, DFR, and GroupDRO (Qiu et al., 2023; Kirichenko et al., 2023; Sagawa et al., 2020). On both text benchmarks, GroupDRO uses group annotations during training and validation and achieves higher mean WGA than GEAR (Table 7).

Table 7: **Test worst-group accuracy (%) across information tiers.** GEAR results are means and standard deviations over five independent runs. ERM reports the same five source classifiers before GEAR's correction. We combine published baseline results with local reproductions; sources and replication counts are provided in Appendix B.2. Bold indicates the largest mean among methods using only task labels during training and validation. "n/a" denotes an inapplicable method or preprocessing variant.

Method	Waterbirds	CelebA	SpuCo Birds	SpuCo Dogs	Civil Comments	MultiNLI
T0: task labels only on training and validation						
ERM	76.39 ± 2.97	30.22 ± 2.31	81.56 ± 1.52	50.60 ± 1.98	58.03 ± 1.53	64.22 ± 0.84
Class-balanced ERM	62.75 ± 2.59	65.56 ± 1.57	81.08 ± 0.50	52.28 ± 2.06	40.90 ± 4.88	66.80 ± 0.50
BAM-CM (local)	84.39 ± 2.46	80.99 ± 0.87	82.20 ± 29.07	76.44 ± 7.27	14.69 ± 6.08	35.96 ± 13.18
BAM + ClassDiff	89.10 ± 0.15	80.10 ± 3.32	93.28 ± 0.30	75.40 ± 9.12	49.44 ± 5.17	70.80 ± 1.52
CoBalT-IG	89.00	89.20	81.72 ± 3.58	49.20 ± 13.46	*n/a*	*n/a*
CCDB	90.48 ± 0.28	85.27 ± 0.28	84.76 ± 1.37	61.00 ± 1.39	44.09 ± 2.73	59.77 ± 2.23
MaskTune	56.07 ± 1.17	78.00	85.28 ± 0.30	61.84 ± 0.86	*n/a*	*n/a*
FG-CCDB	90.56 ± 0.24	$\mathbf{89.22 \pm 0.19}$	88.12 ± 1.43	65.40 ± 1.10	14.84 ± 7.59	1.90 ± 1.21
XRM	86.10 ± 1.28	88.50 ± 1.23	88.80 ± 1.17	80.52 ± 2.37	42.96 ± 4.99	72.80 ± 1.77
GEAR (5)	$\mathbf{91.06 \pm 2.77}$	88.61 ± 1.69	$\mathbf{95.12 \pm 0.41}$	$\mathbf{84.20 \pm 0.24}$	$\mathbf{68.46 \pm 1.91}$	$\mathbf{74.09 \pm 1.68}$
T1: group annotations available on validation						
DPE	94.10	90.30	93.80 ± 0.62	80.28 ± 1.69	68.10 ± 0.72	75.30
GIC	85.40	89.50	89.44 ± 1.56	55.04 ± 8.75	72.30	62.23 ± 1.38
JTT	86.70	81.10	86.93 ± 1.36	76.87 ± 7.31	69.30	72.60
AFR	90.40 ± 1.10	82.00 ± 0.50	90.27 ± 0.61	84.07 ± 3.21	68.70 ± 0.60	73.40 ± 0.60
DFR	92.90 ± 0.20	88.30 ± 1.10	92.32 ± 1.03	81.28 ± 0.70	70.10 ± 0.80	74.70 ± 0.70
XRM (val. prep.)	*n/a*	*n/a*	*n/a*	*n/a*	70.10 ± 4.41	*n/a*
T2: group annotations available on training and validation						
DPE (train prep.)	*n/a*	*n/a*	*n/a*	*n/a*	70.80	*n/a*
GroupDRO	91.40	88.90	91.56 ± 0.65	84.16 ± 1.44	69.90	77.70

Table 8: Average accuracies (%) for the full comparison, using the same predictors and replication counts as Table 7. Local entries report individual-run means $\pm$ sample standard deviations. Published entries retain their native dataset-averaging conventions; in particular, JTT, AFR, DFR, and GroupDRO report training-prevalence-weighted Waterbirds averages. GEAR reports ordinary official-test averages. Dashes indicate unavailable results.

Method	Waterbirds	CelebA	SpuCo Birds	SpuCo Dogs	Civil Comments	MultiNLI
ERM	89.87 ± 1.83	95.57 ± 0.07	91.41 ± 0.43	76.26 ± 0.36	91.96 ± 0.04	82.13 ± 0.27
Class-balanced ERM	81.50 ± 1.73	93.58 ± 0.24	91.36 ± 0.07	76.69 ± 0.50	84.35 ± 0.60	81.50 ± 0.00
BAM-CM (local)	89.81 ± 1.08	87.05 ± 0.65	90.95 ± 13.34	86.32 ± 1.85	70.07 ± 4.04	62.43 ± 3.16
BAM + ClassDiff	91.40	88.40	94.51 ± 0.17	83.73 ± 2.65	86.76 ± 0.66	80.30
CoBalT-IG	92.50	92.30	90.37 ± 2.52	75.15 ± 4.90	—	—
CCDB	92.59 ± 0.10	90.08 ± 0.19	93.64 ± 0.37	80.49 ± 0.49	92.01 ± 0.06	78.47 ± 0.45
MaskTune	88.18 ± 0.23	91.30	93.75 ± 0.09	81.15 ± 0.16	—	—
FG-CCDB	92.50 ± 0.52	89.71 ± 0.54	94.38 ± 0.52	82.27 ± 0.74	58.00 ± 2.26	29.88 ± 2.37
XRM	90.60	91.00	93.90 ± 0.54	86.24 ± 0.74	78.70 ± 0.69	—
GEAR (5)	93.64 ± 0.78	91.69 ± 0.43	97.35 ± 0.07	87.71 ± 0.30	90.23 ± 0.26	79.27 ± 1.32
DPE	—	—	96.19 ± 0.19	87.04 ± 0.70	90.02 ± 0.18	—
GIC	89.30	92.10	94.08 ± 1.02	78.90 ± 2.85	89.70	81.02 ± 0.38
JTT	93.30	88.00	94.17 ± 0.42	85.02 ± 2.66	91.10	78.60
AFR	94.20 ± 1.20	91.30 ± 0.30	95.52 ± 0.06	89.23 ± 1.05	89.80 ± 0.60	81.40 ± 0.20
DFR	94.20 ± 0.40	91.30 ± 0.30	94.28 ± 0.80	85.52 ± 0.19	87.20 ± 0.30	82.10 ± 0.20
GroupDRO	93.50	92.90	95.66 ± 0.08	87.99 ± 0.85	88.90	81.40

Table 9 gives GEAR's within-run changes from its source classifiers and selected correction counts.

Table 9: Paired accuracy changes in percentage points, mean ± standard deviation across the same five runs per dataset as Table 1, and action counts. Head and Representation denote the two relearning modes. The unchanged source is also a candidate and is never selected here.

	GEAR minus source		Selected action count		
Dataset	ΔWGA	ΔAverage	Suppress	Head	Representation
Waterbirds	$+14.67 \pm 3.83$	$+3.77 \pm 1.39$	5	0	0
CelebA	$+58.39 \pm 2.05$	-3.87 ± 0.45	0	0	5
SpuCoBirds	$+13.56 \pm 1.47$	$+5.94 \pm 0.45$	0	0	5
SpuCoDogs	$+33.60 \pm 1.99$	$+11.45 \pm 0.44$	0	5	0
CivilComments	$+10.43 \pm 2.18$	-1.72 ± 0.23	1	4	0
MultiNLI	$+9.88 \pm 1.86$	-2.86 ± 1.05	5	0	0

B.4 Component ablations

B.4.1 Learned geometry and correction families

These controls test direction orientation and the available correction families (Figure 3(a)); Table 10 gives absolute WGA. Waterbirds uses classification error for final validation selection and suppression ridge 0.1. CelebA uses classification error for the representation-region shares. Each comparison holds its remaining fitting and selection settings fixed.

For each source, apply one random orthogonal transformation. On Waterbirds, rotate the suppression basis, refit the constrained heads, recalibrate intercepts, and select rank and strength by validation. On CelebA, jointly rotate the region directions while preserving their Gram matrix, recompute regions, and refit representation relearning with $\beta = 0$. Initializations, data orders, and optimizers match the learned-direction controls. Rotation seeds are $9{,}401{,}000$ plus source index on Waterbirds and $9{,}402{,}000$ plus source index on CelebA. Correction-family controls reselect from the fitted candidates within the permitted family, always including the source.

Table 10: Mechanism ablations: test WGA (%), mean ± sample standard deviation over five paired runs. The reference uses learned geometry and all correction families. Rotation preserves direction lengths, relative angles, and rank, with suppression fixed on Waterbirds and representation relearning fixed on CelebA; the learned-geometry controls match the reference in every run. The three restricted-family rows retain learned geometry. Every comparison also permits the unchanged source.

Configuration	Waterbirds	CelebA
Reference (all corrections)	90.65 ± 2.19	89.32 ± 1.26
Randomly rotated geometry	76.38 ± 2.97	79.78 ± 5.13
Suppression only	90.65 ± 2.19	76.89 ± 0.93
Head relearning only	69.50 ± 13.65	74.67 ± 1.28
Representation relearning only	75.47 ± 5.18	89.32 ± 1.26

B.4.2 Final validation criterion

These controls provide the validation-selection results in Figure 3(a). Rank the same fitted Waterbirds and CelebA candidate collections by ordinary soft error, difficulty-weighted soft error, or their smooth maximum. Fitting, calibration, ranks, and interpolation strengths remain fixed. Ordinary error gives lower WGA on Waterbirds; difficulty-weighted error gives lower WGA on CelebA. The combined criterion matches the better alternative on each dataset (Table 11).

Table 11: Final-selection ablation on fixed candidate collections. Test accuracy (%), mean ± sample standard deviation over five paired runs.

	Waterbirds		CelebA	
Final criterion	WGA	Average	WGA	Average
Ordinary error	76.27 ± 3.21	90.13 ± 1.99	88.61 ± 1.69	91.69 ± 0.43
Difficulty-weighted error	91.06 ± 2.77	93.64 ± 0.78	85.00 ± 6.93	91.18 ± 1.19
Combined	91.06 ± 2.77	93.64 ± 0.78	88.61 ± 1.69	91.69 ± 0.43

B.4.3 Training weights

The class-balanced variant in Table 3 replaces both regional allocation and within-region weighting with $\omega_i^{\mathrm{CB}} = 1/(Cn_{y_i})$, where n_{y_i} is the fitting class count. Refit both relearning branches, then repeat calibration and final selection with the source and suppression candidates available. Inputs, geometry, initialization, optimizer, schedule, and regularization are fixed. For the region-indexed batch estimator, assign each region the sum of its class-balanced example weights and set within-region multipliers to one. Geometric regions remain in the checkpoint rule. Table 12 reports average accuracy and paired gains, measuring the joint contribution of the proposed weights under these fitting and selection rules.

Table 12: Supporting results for the weighting ablation of GEAR (Table 3). Average accuracies are percentages; paired gains are full GEAR minus the class-balanced variant in percentage points. Summaries are means ± sample standard deviations over five paired sources. The last column counts strictly positive per-source WGA gains.

	Average accuracy			
Dataset	Class-balanced	GEAR	Paired WGA gain	Sources improved
SpuCoDogs	73.19 ± 0.36	87.71 ± 0.30	36.28 ± 0.36	5 / 5
SpuCoBirds	95.22 ± 0.07	97.35 ± 0.07	4.80 ± 0.37	5 / 5
CivilComments	89.91 ± 0.59	90.23 ± 0.26	1.29 ± 2.64	3 / 5
MultiNLI	79.27 ± 1.32	79.27 ± 1.32	0.00 ± 0.00	0 / 5

B.5 Hyperparameter sensitivity

Figures 3(b) and 4 use the same five independent source models, indexed 0–4, at every point of each curve. These sweeps evaluate fixed correction families on the complete official test splits. Each sweep varies one parameter while holding the remaining configuration fixed. Summaries average individual-run errors and report sample standard deviations.

B.5.1 Suppression strength

Evaluate $\alpha = 0, 0.1, \ldots, 2$ using the score path in Section 2.2. Reuse the fitted source and calibrated projected heads without refitting or recalibration. For each source, fix the rank of its best suppression candidate under the original task-label-only validation rule. The ranks are $(2, 2, 2, 2, 1)$ on Waterbirds and $(1, 2, 1, 2, 1)$ on CelebA, with ridge coefficients 0.05 and 0.1, respectively. Here $\alpha = 0$ retains the source, and $\alpha = 1$ gives complete suppression. The extension above one reverses sensitivity along the suppressed directions; it is an extrapolation diagnostic outside the method's $[0, 1]$ range.

B.5.2 Penalty strength

Moderate regularization gives lower worst-group error than weak or excessive regularization on both Waterbirds and SpuCoDogs (Figure 4). Although the preferred penalty scale differs between the two corrections, neighboring moderate values retain relatively low error. These patterns support searching a moderate range of penalties and refining the choice through validation.

The penalty grid is 0, 0.001, 0.003, 0.01, 0.02, 0.03, 0.05, 0.075, 0.1, 0.2, 0.3, 1, 3, and 10. On Waterbirds, refit the constrained suppression head at each penalty, retaining the source, standardized

features, directions, rank, and suppression strength. Apply the same task-class validation intercept calibration after each fit. On SpuCoDogs, vary the quadratic anchor to the source weights and intercept, retaining the eight adaptive head fits and the task-label-only validation rule for selecting among them. Coefficients use the native binary one-logit parameterization. Thus $\lambda = 0$ fits a correction without regularization.

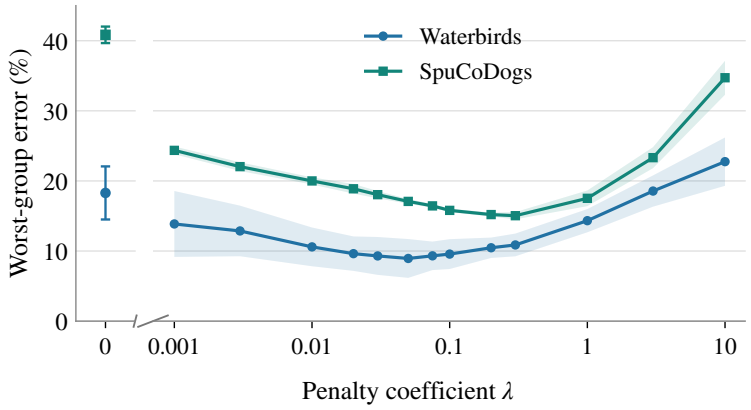

Figure 4: **Penalty-strength sensitivity.** Waterbirds varies the suppression-head penalty; SpuCoDogs varies the anchor in head relearning. Lines show mean test worst-group error (%, lower is better); bands and error bars show sample standard deviations over five paired runs. Positive coefficients use a log scale, with zero shown separately.

B.6 Background interventions and linear probes

These experiments examine which image changes the directions capture and what a common linear-classifier fit recovers from the altered features. They use the five source models and retained candidates underlying Table 1.

B.6.1 Background interventions

For all 5,794 Waterbirds test images, use the CUB foreground segmentations[1] to replace either background or bird pixels with RGB $(128, 128, 128)$. Images and masks undergo the source pipeline's aspect-preserving resize to 256 and center crop to 224×224; masks are thresholded at 0.5. Masks are used only for evaluation. Predictors and batch-normalization statistics remain fixed.

For the standardized feature change $\Delta z = z(x') - z(x)$, measure $\|U^\top \Delta z\|_2^2 / \|\Delta z\|_2^2$, where U contains the selected orthonormal suppression directions. Compare a random subspace of the same rank, with seed $9{,}413{,}000$ plus source index. Each run averages these fractions over images. Table 13 also reports label changes and mean absolute changes in class-1 probability; Table 14 gives accuracy under each intervention.

The learned directions respond to both background and bird removal. Background removal changes GEAR's labels less often than ERM's, but the substantial response to bird removal shows that these directions also contain task-related variation. Gray replacement is an artificial shift, and bird removal may retain outline cues; the results concern these interventions.

B.6.2 Linear probes

Fit a new class-balanced logistic head on each of four feature pipelines: source features, source features projected off learned directions, source features projected off random directions, and relearned features. Use the complete training split and select ridge from $\{0.001, 0.01, 0.1, 1\}$ by ordinary class-balanced validation soft error, preferring stronger ridge in ties. Projection ranks and relearning

[1] https://data.caltech.edu/records/w9d68-gec53.

Table 13: Waterbirds intervention diagnostics, expressed as percentages. Means ± sample standard deviations over five source models; each run first averages over the full test-image inventory. Probability change is the mean absolute change in class-1 probability.

	Feature-change energy		Changed predictions		Probability change	
Intervention	Learned	Random	Source	GEAR	Source	GEAR
Background removed	19.47 ± 6.41	0.08 ± 0.03	8.84 ± 1.51	5.28 ± 0.67	9.35 ± 1.33	8.15 ± 0.94
Bird removed	41.62 ± 17.03	0.09 ± 0.06	36.52 ± 3.65	44.35 ± 12.45	34.36 ± 3.49	37.24 ± 6.36

Table 14: Accuracy (%) under the Waterbirds interventions, mean ± sample standard deviation over five runs. Original-input rows use the same predictors as Table 1. The modified-input rows are artificial diagnostic conditions.

	Source classifier		GEAR	
Input	WGA	Average	WGA	Average
Original	76.39 ± 2.97	89.87 ± 1.83	91.06 ± 2.77	93.64 ± 0.78
Background removed	92.97 ± 0.80	94.70 ± 1.42	88.66 ± 2.58	96.24 ± 0.43
Bird removed	16.00 ± 7.85	55.47 ± 3.83	37.12 ± 16.53	52.69 ± 12.98

strengths are inherited from the task-only candidate selection. Source projections preserve source standardization; relearned features use their training means and standard deviations. Each pipeline retains its preprocessing: Waterbirds relearning resizes to 256×256 before cropping, while its source preserves aspect ratio. The comparison therefore concerns the complete feature pipelines.

Learned projection improves mean probe WGA on both datasets; random projection has little effect (Table 15). Relearned features do not improve WGA under this common fitting rule. On CelebA, their new probe achieves $65.00 \pm 2.78\%$ WGA, compared with $88.61 \pm 1.69\%$ for the retained GEAR classifier on the same features. The fitted head is therefore part of the robustness gain.

Table 15: Linear-probe test accuracy (%), mean ± sample standard deviation over five paired source models. All rows refit and select a class-balanced linear head using the same task-label-only procedure.

	Waterbirds		CelebA	
Feature pipeline	WGA	Average	WGA	Average
Source	69.43 ± 5.67	88.28 ± 1.06	66.22 ± 1.44	93.81 ± 0.09
Random projection	69.33 ± 5.76	88.28 ± 1.06	66.22 ± 1.27	93.80 ± 0.08
Learned projection	82.18 ± 4.67	93.62 ± 1.11	70.67 ± 1.54	93.13 ± 0.14
Relearned	65.14 ± 1.46	85.86 ± 1.32	65.00 ± 2.78	94.15 ± 0.18

C Proofs of the Main Results

This section proves the two results in Section 4. Appendices C.2 and C.3 give their proofs; Appendices C.4 and C.5 give the validation sampling bound and sufficient conditions for asymptotic optimality.

C.1 Common setup and preliminary results

We collect the common notation and preliminary results for the Gaussian feature model in Section 4. Write a linear head as $f_{u,v}(Z) = u^\top Z_{\text{task}} + v^\top Z_{\text{context}}$, with $(u, v) \neq (0, 0)$. Its score is the difference between the two class logits, and its sign determines the prediction. Expectations evaluate the fitted heads on a fresh draw from this model. All heads have zero intercept unless stated otherwise.

C.1.1 GROUP ACCURACY AND THE COMPARISON OF FITTED HEADS

Deriving the accuracy formula. For $f = f_{u,v}$, multiplying the score by its label gives the margin

$$M_f = Yf(Z) = u^\top \mu + (YE)v^\top \nu + u^\top Y\xi + v^\top Y\zeta.$$

A positive margin means a correct prediction. A centered Gaussian vector has the same distribution after a sign change. Thus, in group (y, e), M_f is normal with mean $u^\top \mu + ye\, v^\top \nu$ and standard deviation

$$s_f = \sqrt{u^\top \Sigma_{\text{task}} u + v^\top \Sigma_{\text{context}} v} > 0 \qquad \text{for } (u, v) \neq (0, 0).$$

Subtracting the mean and dividing by s_f gives a standard normal variable. Let Φ denote its distribution function. By symmetry,

$$\mathbb{P}(M_f > 0 \mid Y = y, E = e) = \Phi\left(\frac{u^\top \mu + ye\, v^\top \nu}{s_f}\right).$$

The two groups with $YE = 1$ have the same margin distribution, as do the two with $YE = -1$. Their smaller mean is $u^\top \mu - |v^\top \nu|$, so

$$\text{WGA}(f_{u,v}) = \Phi\left(\frac{u^\top \mu - |v^\top \nu|}{\sqrt{u^\top \Sigma_{\text{task}} u + v^\top \Sigma_{\text{context}} v}}\right). \tag{20}$$

The numerator is the useful task signal minus the adverse context contribution; the denominator is the standard deviation of the score's noise. Both corrections change how the classifier uses these two blocks. Equation (20) compares their fitted heads, accounting for the directions of their coefficients as well as their magnitudes. Positive definiteness of the covariances ensures that every nonconstant head has a continuous margin distribution, so ties have probability zero.

An intercept b adds yb to the mean in group (y, e). The four groups allow y and ye to vary independently, so the smallest mean becomes $u^\top \mu - |v^\top \nu| - |b|$. The variance is unchanged. Therefore zero intercept maximizes WGA for every fixed nonzero slope. This justifies the zero-intercept family studied here; it uses the symmetry of the feature distributions across classes, in addition to equal class probabilities.

Proposition C.1 (Task and context feature blocks)**.** *Under the Gaussian feature model in Section 4, fix u with $u^\top \mu > 0$. For every $v \neq 0$,*

$$\text{WGA}(f_{u,0}) > \text{WGA}(f_{u,v}).$$

The largest WGA among all nonconstant zero-intercept linear heads is

$$\max_{(u,v)\neq(0,0)} \text{WGA}(f_{u,v}) = \Phi\left(\sqrt{\mu^\top \Sigma_{\text{task}}^{-1} \mu}\right). \tag{21}$$

It is attained exactly by $u = c\Sigma_{\text{task}}^{-1}\mu$, $v = 0$, with $c > 0$.

Proof. For $u^\top \mu > 0$ and $v \neq 0$, positive definiteness gives

$$\begin{aligned}\frac{u^\top \mu - |v^\top \nu|}{\sqrt{u^\top \Sigma_{\text{task}} u + v^\top \Sigma_{\text{context}} v}} &\leq \frac{u^\top \mu}{\sqrt{u^\top \Sigma_{\text{task}} u + v^\top \Sigma_{\text{context}} v}} \\ &< \frac{u^\top \mu}{\sqrt{u^\top \Sigma_{\text{task}} u}}.\end{aligned}$$

Applying the increasing function Φ proves the first claim. To maximize the last expression over u, apply Cauchy–Schwarz to $\Sigma_{\text{task}}^{1/2} u$ and $\Sigma_{\text{task}}^{-1/2} \mu$:

$$u^\top \mu \leq \sqrt{u^\top \Sigma_{\text{task}} u}\, \sqrt{\mu^\top \Sigma_{\text{task}}^{-1} \mu}.$$

Here the matrix square roots are obtained by taking square roots of the positive eigenvalues. Equality with $u^\top \mu > 0$ holds precisely when u is a positive multiple of $\Sigma_{\text{task}}^{-1}\mu$. If $u^\top \mu \leq 0$, Eq. (20) gives WGA at most $1/2$, which is smaller than the stated optimum because $\mu \neq 0$. These observations prove the maximum and all its equality conditions. □

For isotropic task noise, $\Sigma_{\text{task}} = \sigma^2_{\text{task}} I$, the optimum is $\Phi(\|\mu\|/\sigma_{\text{task}})$, attained by positive multiples of μ. For general covariance, the optimal task coefficient also accounts for which feature directions have more noise. Consequently, the ratio $\|v\|/\|u\|$ alone does not order arbitrary fitted heads: changing the direction of u changes both its useful signal and its noise. Equation (20) gives the full comparison. The covariance formulation also permits any fixed invertible rescaling within each block, including coordinate standardization.

C.1.2 The direction and the fitted corrections

The expected direction from reference ranks. The signed feature vector is $YZ = (\mu + Y\xi, (YE)\nu + Y\zeta)$. Write the reference head as

$$f_{\text{ref}}(Z) = u_{\text{ref}}^\top Z_{\text{task}} + v_{\text{ref}}^\top Z_{\text{context}}.$$

Binary softmax probabilities satisfy $p_f(Y \mid Z) = \sigma(M_f)$, where $\sigma(t) = 1/(1+e^{-t})$. Let F_M and h_M denote the distribution function and density of the reference margin $M_{f_{\text{ref}}}$. Since σ is increasing, ranking the reference probabilities gives the same order as ranking $M_{f_{\text{ref}}}$. The signed features have the same distribution in both task classes: their agreement probabilities are both p, and changing the noise signs preserves their laws. Hence the two class directions span the same axis. Up to a nonzero scale and sign, this axis from Eq. (3) is

$$d = \mathbb{E}[YZ\{F_M(M_{f_{\text{ref}}}) - 1/2\}]. \tag{22}$$

For a continuous variable, $\mathbb{E}F_M(M_{f_{\text{ref}}}) = 1/2$: between two independent copies, either is larger with probability $1/2$.

Define

$$A_d = \mathbb{E}h_M(M_{f_{\text{ref}}}) > 0, \qquad B_d = \mathbb{E}[YE\{F_M(M_{f_{\text{ref}}}) - 1/2\}].$$

Then the two blocks of the rank direction are

$$\begin{aligned} d_{\text{task}} &= A_d \Sigma_{\text{task}} u_{\text{ref}}, \\ d_{\text{context}} &= A_d \Sigma_{\text{context}} v_{\text{ref}} + \nu B_d. \end{aligned} \tag{23}$$

To derive this identity, the constant task mean first disappears because $\mathbb{E}\{F_M(M_{f_{\text{ref}}}) - 1/2\} = 0$. For a standard normal density φ, $x\varphi(x) = -\varphi'(x)$. Integration by parts gives

$$\int x g(x)\varphi(x)\,dx = \int g'(x)\varphi(x)\,dx$$

for the bounded smooth functions used here. Write $\xi = \Sigma_{\text{task}}^{1/2} G$, where the coordinates of G are independent standard normals. Apply the identity to each coordinate of G while holding the others, the context noise, and (Y, E) fixed. The derivative of $F_M(M_{f_{\text{ref}}})$ with respect to that coordinate is the corresponding entry of $Y\Sigma_{\text{task}}^{1/2} u_{\text{ref}}$ times $h_M(M_{f_{\text{ref}}})$. Multiplying by $Y\Sigma_{\text{task}}^{1/2}$ and averaging gives

$$\mathbb{E}[Y\xi F_M(M_{f_{\text{ref}}})] = \Sigma_{\text{task}} u_{\text{ref}} \mathbb{E}h_M(M_{f_{\text{ref}}}).$$

The context noise gives the analogous term, and its conditional mean contributes νB_d. This proves Eq. (23).

For explicit expressions and later use, put

$$a_{\text{ref}} = u_{\text{ref}}^\top \mu, \qquad b_{\text{ref}} = v_{\text{ref}}^\top \nu.$$

The reference margins conditional on $YE = 1$ and $YE = -1$ are normal with means $a_{\text{ref}} + b_{\text{ref}}$ and $a_{\text{ref}} - b_{\text{ref}}$, respectively. Their common variance is $s^2_{f_{\text{ref}}}$, using the standard deviation defined above. Comparing independent margins and integrating the squared mixture density gives

$$\begin{aligned} A_d &= \frac{p^2 + (1-p)^2 + 2p(1-p)e^{-b^2_{\text{ref}}/s^2_{f_{\text{ref}}}}}{2\sqrt{\pi}\, s_{f_{\text{ref}}}}, \\ B_d &= 2p(1-p)\left\{\Phi\left(\frac{\sqrt{2}\, b_{\text{ref}}}{s_{f_{\text{ref}}}}\right) - \frac{1}{2}\right\}. \end{aligned}$$

Thus a reference that uses task features generally produces a direction with a task component: within-class task noise affects its confidence.

Fitting the corrections. Write $w = (u, v)$ and $\ell(m) = \log(1 + e^{-m})$ for the logistic training loss. The vector w is the difference between the two rows of W. Minimizing over their common row component reduces the Frobenius penalties to $\|w\|^2/2$ for suppression and $\|w - w_0\|^2/2$ for anchored relearning. The algorithm constructs directions from empirical within-class ranks. Using the class-balanced weights ω_i from Section 2.2, suppression of one learned direction $\widehat{d}$ fits

$$w^{\perp} = \arg\min_{w:\, w^{\top}\widehat{d}=0} \left\{ \sum_{i\in\mathcal{T}} \omega_i \ell(y_i w^{\top} z_i) + \frac{\lambda}{2}\|w\|^2 \right\}, \qquad \lambda > 0. \tag{24}$$

For several directions, their constraints are imposed together. The interpolation uses $w_\alpha = (1 - \alpha)w_0 + \alpha w^{\perp}$ at the prescribed values of $\alpha \in [0, 1]$.

For head relearning, fix the difficult and other regions within each task class as in Appendix A.3.2. Their accumulated validation errors determine the region shares at iteration j. Equation (5) then gives training weights $\omega_i^{(j)}$ with $\sum_{i\in\mathcal{T}} \omega_i^{(j)} = 1$. Within the zero-intercept family, the anchored head solves

$$w_{\text{head}}^{(j)} = \arg\min_{w} \left\{ \sum_{i\in\mathcal{T}} \omega_i^{(j)} \ell(y_i w^{\top} z_i) + \frac{\lambda}{2}\|w - w_0\|^2 \right\}, \qquad \lambda > 0.$$

Differentiating gives the equation for its unique minimizer,

$$\lambda(w_{\text{head}}^{(j)} - w_0) = \sum_{i\in\mathcal{T}} \omega_i^{(j)} y_i z_i \sigma(-y_i (w_{\text{head}}^{(j)})^{\top} z_i). \tag{25}$$

The fitting objectives determine the coefficients to be compared in Eq. (20); reweighting does not prescribe their alignment. Updating the nonlinear feature extractor is outside this fixed-feature analysis.

When full suppression preserves an optimal head. In block coordinates, the suppression constraint is $u^{\top}\widehat{d}_{\text{task}} + v^{\top}\widehat{d}_{\text{context}} = 0$. Even if $\widehat{d}_{\text{task}} = 0$, the remaining condition $v^{\top}\widehat{d}_{\text{context}} = 0$ need not remove the entire context block. By Proposition C.1, an optimal linear head remains feasible after full suppression if and only if

$$\widehat{d}_{\text{task}}^{\top}\Sigma_{\text{task}}^{-1}\mu = 0. \tag{26}$$

Indeed, substitute $u = c\Sigma_{\text{task}}^{-1}\mu$ and $v = 0$, with $c > 0$, into the constraint. These are all the optimal heads, so the condition is both necessary and sufficient. It describes feasibility; the logistic fitting objective determines which feasible head is fitted.

For the expected rank direction d in Eq. (23),

$$d_{\text{task}}^{\top}\Sigma_{\text{task}}^{-1}\mu = A_d u_{\text{ref}}^{\top}\mu.$$

Thus suppressing this expected direction preserves an optimal head exactly when $u_{\text{ref}}^{\top}\mu = 0$. When the reference uses task signal, suppressing this direction excludes the optimal task-only direction. A fitted relearning head can still use that direction; whether it improves WGA is decided by its actual coefficients in Eq. (20).

Partial suppression for feature blocks. Write $w_0 = (u_0, v_0)$ and $w^{\perp} = (u^{\perp}, v^{\perp})$. The context coefficient along the interpolation is

$$v_\alpha = (1 - \alpha)v_0 + \alpha v^{\perp}.$$

If $v_0 \neq 0$, an interior interpolation has $v_\alpha = 0$ precisely when $v^{\perp} = -cv_0$ for some $c > 0$; solving the equation gives $\alpha = 1/(1 + c)$. If $v^{\perp} = 0$, the suppression endpoint itself has zero context coefficient. A head with $v_\alpha = 0$ attains the global linear optimum only when its task coefficient u_α is a positive multiple of $\Sigma_{\text{task}}^{-1}\mu$. For other configurations, Eq. (20) directly compares the available interpolation points and relearning heads. A finite interpolation grid may miss a value that the continuous path attains.

C.2 Proof of Proposition 4.1: Neither correction uniformly dominates

We prove Proposition 4.1. The first regime needs suppression to remove a strong context coefficient. In the second, the reference mixes context with task-block noise: suppression leaves some nuisance dependence, whereas relearning removes it. The conditions below specify the feature law and the initial/reference heads; optimality of a fitted head is derived.

Feature distribution and correction settings. In original feature coordinates, consider

$$Z_1 = mY + \xi_1, \qquad Z_2 = \xi_2, \qquad Z_c = \nu E + \zeta, \qquad m, \nu > 0, \tag{27}$$

with independent standard normal noises and the label/context law of the Gaussian feature model in Section 4. Thus $Z_{\text{task}} = (Z_1, Z_2)$, $\mu = (m, 0)$, and both noise covariances are identity matrices. The signed variables

$$X = YZ_1 \sim \mathcal{N}(m, 1), \qquad N = YZ_2 \sim \mathcal{N}(0, 1), \qquad V = YZ_c \sim p\mathcal{N}(\nu, 1) + (1-p)\mathcal{N}(-\nu, 1)$$

are mutually independent. Proposition C.1 gives

$$\mathrm{WGA}(cZ_1 + nZ_2 + vZ_c) = \Phi\left(\frac{cm - \nu|v|}{\sqrt{c^2 + n^2 + v^2}}\right) \le \Phi(m), \tag{28}$$

with equality exactly when $c > 0$ and $n = v = 0$.

Fix one suppression penalty λ_s, one head penalty λ_h, one within-region strength β, and any finite number of relearning rounds, with

$$0 < \lambda_h < \lambda_s, \qquad \beta > 0, \quad \kappa > 0, \quad \tau > 0. \tag{29}$$

The pool retains the initial head, rank-one suppression with a fixed interpolation grid containing $\alpha = 1$, and all these relearning rounds. The population comparisons below also cover the continuous suppression path $\alpha \in [0, 1]$. The reference supplies the split difficulty and mistake restriction, and the initial head supplies the relearning direction, as in Section 2.2. For the empirical conclusion, take equally many observations from each task class in fitting and validation. This specifies class-stratified sampling, so the majority-class adjustment in Appendix A.3.2 is inactive. All samples are used by the stated procedure. These conditions concern the fixed-feature, rank-one pool above; additional ranks or hyperparameter settings are not part of this asymptotic claim.

Population standardization gives the original-coordinate penalty factors $D_x = 1 + m^2$ and $D_v = 1 + \nu^2$. For $\lambda > 0$, let c_λ be the unique positive solution of

$$\mathbb{E}[X\sigma(-c_\lambda X)] = \lambda D_x c_\lambda. \tag{30}$$

It minimizes $\mathbb{E}\ell(cX) + \lambda D_x c^2/2$: strict convexity gives uniqueness, and the derivative at zero is $-m/2 < 0$. Write $c_s = c_{\lambda_s}$, $c_h = c_{\lambda_h}$, $q(c) = \mathbb{E}\sigma(-cX)$, and $H = q(c_h)$. Then

$$c_h > c_s > 0, \qquad q(c_h) < q(c_s) < 1/2. \tag{31}$$

The first inequality follows from the strictly decreasing left side of Eq. (30). Pairing the positive and negative values of X shows $q'(c) < 0$ for $c > 0$.

At the baseline reference $f_{\text{ref}} = Z_c$, require a unique nondegenerate population difficulty split at a positive threshold in V. The difficult region, after intersecting with reference mistakes, is then $\{V < 0\}$. Require also

$$\begin{aligned} \bar{v}_\beta &:= \frac{1}{2}\frac{\mathbb{E}[Ve^{-\beta\sigma(V)}\mathbf{1}\{V < 0\}]}{\mathbb{E}[e^{-\beta\sigma(V)}\mathbf{1}\{V < 0\}]} + \frac{1}{2}\frac{\mathbb{E}[Ve^{-\beta\sigma(V)}\mathbf{1}\{V \ge 0\}]}{\mathbb{E}[e^{-\beta\sigma(V)}\mathbf{1}\{V \ge 0\}]} < 0, \\ p_\kappa &:= \frac{\mathbb{E}[e^{-\kappa\sigma(V)}\mathbf{1}\{YE = 1\}]}{\mathbb{E}e^{-\kappa\sigma(V)}} < 1 - p. \end{aligned} \tag{32}$$

Step 5 below establishes a nonempty range where all these conditions hold with finite parameters.

Two sufficient source/reference configurations. In the suppression regime, take

$$f_{\text{ref}} = Z_c, \qquad f_0 = B_0 Z_c, \qquad B_0 > 0 \text{ sufficiently large and fixed.} \tag{33}$$

In the relearning regime, take

$$f_{\text{ref},\epsilon} = Z_c + \epsilon Z_2, \qquad f_{0,\epsilon} = s_N Z_2 + s_V Z_c, \qquad s_V > 0, \quad 0 < \epsilon \text{ sufficiently small.} \tag{34}$$

For $r = s_N/s_V$, use the actual population rank direction, difficulty split, and reference-mistake restriction to form the regions. Let $(m_N(r,\epsilon), m_V(r,\epsilon)) = \mathbb{E}_{\omega^{(0)}}(N, V)$ be the nuisance-feature mean under the first-round weights in Eq. (5). These regions and normalized weights depend on r, ϵ, but not on the positive source scale s_V. The required structural condition is

$$s_N = -\frac{H}{\lambda_h} m_N(r,\epsilon), \qquad s_V = -\frac{H}{\lambda_h D_v} m_V(r,\epsilon), \qquad r = s_N/s_V. \tag{35}$$

This condition balances the reweighted nuisance means against the anchor to the initial head. It is an additional moment condition, not a consequence of Gaussianity alone. Step 4 proves that the algorithm's own weights admit a continuous family of sources satisfying it.

Step 1: derive the relearning minimizer. The first relearning objective in original coordinates is

$$J(c,n,v) = \mathbb{E}_{\omega^{(0)}}\ell(cX + nN + vV) + \frac{\lambda_h}{2}\{D_x c^2 + (n - s_N)^2 + D_v(v - s_V)^2\}.$$

All regions and within-region weights depend only on (N, V). Thus X retains its original law and remains independent of the nuisance features. At $(c, n, v) = (c_h, 0, 0)$,

$$\begin{aligned}\partial_c J &= -\mathbb{E}[X\sigma(-c_h X)] + \lambda_h D_x c_h = 0,\\ \partial_n J &= -Hm_N - \lambda_h s_N = 0,\\ \partial_v J &= -Hm_V - \lambda_h D_v s_V = 0.\end{aligned}$$

Equations (30) and (35) give the equalities. Strong convexity therefore makes $f(Z) = c_h Z_1$ the unique minimizer, and Eq. (28) gives its globally optimal WGA $\Phi(m)$. Every region has the same population soft error $q(c_h)$, since its membership uses only nuisance features. The accumulated-error update keeps all region shares equal. Induction shows that every subsequent population relearning round has this same minimizer.

Step 2: show that suppression misses the optimum in this regime. For reference $Z_c + \epsilon Z_2$, Eq. (23) gives the standardized nuisance components

$$d_N = A_d\epsilon, \qquad d_c = \frac{A_d + \nu B_d}{\sqrt{D_v}}, \qquad A_d > 0,\ B_d > 0.$$

The original-coordinate suppression constraint is consequently

$$A_d\epsilon n + (A_d + \nu B_d)v = 0. \tag{36}$$

If a constrained minimizer were globally optimal, it would be task-only; minimization along that feasible line would fix it to $c_s Z_1$. At $(c_s, 0, 0)$, the nuisance gradient of the suppression objective is

$$\big(0, -q(c_s)(2p-1)\nu\big).$$

This vector is not parallel to the constraint normal in Eq. (36), whose noise component is nonzero. The constrained first-order equations rule out optimal WGA.

At $\epsilon = 0$, Eq. (35) gives $s_N = 0$ and $s_V > 0$. Continuity therefore gives $A_d\epsilon s_N + (A_d + \nu B_d)s_V > 0$ for small positive ϵ. For every interpolation with $\alpha < 1$, the inner product with this normal is $(1 - \alpha)$ times that positive value. Its nuisance coefficients cannot both vanish. The endpoint is also suboptimal, as just shown. Each noninitial interpolation has a positive task coefficient: at task coefficient zero the task derivative is $-m\mathbb{E}\sigma(-nN - vV) < 0$. Thus the path contains no zero score. Compactness and continuity give

$$\max_{f\in\mathcal{F}_{\text{sup}}} \text{WGA}(f) < \Phi(m) = \max_{f\in\mathcal{F}_{\text{rel}}} \text{WGA}(f), \tag{37}$$

even for the entire continuous suppression path.

Step 3: derive the suppression-regime advantage. Now consider Eq. (33). The rank direction is the context axis, so full suppression fits $c_s Z_1$ and attains $\Phi(m)$.

Relearning regions are independent of the positive source scale B_0. Within each region, the weighted first moments of $|X|$ and $|V|$ are finite; all regional shares give convex combinations of these moments. The anchored first-order equation, Eq. (25), therefore gives, uniformly over all rounds and regional shares,

$$c_{\text{rel}} = O(1), \qquad n_{\text{rel}} = 0, \qquad v_{\text{rel}} = B_0 + O(1) \quad (B_0 \to \infty).$$

The noise coefficient is zero by symmetry and strict convexity. Equation (28) gives relearning WGA tending to $\Phi(-\nu) < 1/2$.

For sufficiently large finite B_0, full suppression therefore has strictly greater WGA than every relearning round. The same fixed penalties $\lambda_h < \lambda_s$ work in both regimes.

Step 4: establish the moment-balanced source family. At $r = \epsilon = 0$, the difficult region is $V < 0$, so $m_N(0,0) = 0$ and $m_V(0,0) = \bar{v}_\beta < 0$. For nearby (r, ϵ), orient and rescale the initial-head projection as $V + z(r)N$. Equation (23), with the standardization factors included, gives

$$z(r) = \frac{D_v A(r) r}{A(r) + \nu B(r)} = O(r),$$

where $A(r) > 0$ and $B(r) > 0$ are its rank coefficients. Uniqueness and nondegeneracy of the baseline split give a continuous nearby threshold, bounded below by a positive constant $s_* > 0$. The reference mistake event is $V + \epsilon N < 0$. Its difference from the difficult region requires $(z(r) - \epsilon)N > s_*$. Gaussian tails make the probability and the required weighted first moments of this difference $O(\exp\{-C/(|r| + |\epsilon|)^2\})$ for some $C > 0$. The normalizing region masses stay positive. Integrating the smooth Gaussian densities across these boundaries shows that the weighted moments are continuously differentiable near the origin; the geometric restriction has zero first derivative there with respect to r.

Consequently, for

$$F(r, \epsilon) = r m_V(r, \epsilon) - D_v m_N(r, \epsilon),$$

we have $F(0,0) = 0$ and $\partial_r F(0,0) = \bar{v}_\beta < 0$. The implicit function theorem supplies a continuous local solution $r(\epsilon) \to 0$ of $F(r(\epsilon), \epsilon) = 0$. Taking

$$s_V = -\frac{H}{\lambda_h D_v} m_V(r(\epsilon), \epsilon) > 0, \qquad s_N = r(\epsilon) s_V$$

proves Eq. (35). Thus the exact GEAR regions and weights admit this structural condition for every sufficiently small nonzero reference perturbation.

Step 5: verify that the distributional conditions are nonempty. An analytic sufficient range is $0 < \nu \leq 0.3$ and any $p \in (1/2, 1)$. For reference difficulty $D = \sigma(-V)$, let $g(d)$ be its density, $t = \log(d/(1-d))$, and $\eta = \frac{1}{2}\log((1-p)/p)$. Changing variables in the Gaussian mixture gives

$$\begin{aligned}[d(1-d)]^2 (\log g)''(d) = & -2d(1-d) - t(2d-1) \\ & + \nu^2 \operatorname{sech}^2(\nu t + \eta) + \nu(2d-1)\tanh(\nu t + \eta).\end{aligned}$$

For $|t| \leq 1$, the first negative term has magnitude at least $2e/(1+e)^2 > 0.393$, while the positive terms are at most $0.09 + 0.3\tanh(1/2) < 0.229$. For $|t| \geq 1$, an upper bound is $0.09 - 0.7\tanh(1/2) < 0$. Thus g is strictly log-concave.

To see why this gives a unique split, write G_D for its distribution function. The reverse hazard g/G_D decreases and the hazard $g/(1-G_D)$ increases, as follows by differentiating their integrals using concavity of $\log g$. In particular,

$$\frac{d}{dd}\mathbb{E}[D \mid D \leq d] = \frac{g(d)}{G_D(d)} \int_0^d \frac{G_D(x)}{G_D(d)}\, dx < 1;$$

bound the integrand by $\exp\{-(g(d)/G_D(d))(d-x)\}$ to obtain the strict inequality. The upper truncated mean has the analogous derivative below one. Hence $\mathbb{E}[D \mid D \leq d] + \mathbb{E}[D \mid D > d] - 2d$ strictly decreases from a positive to a negative value. Its unique zero is the unique nondegenerate squared-error split. For $p > 1/2$, the conditional density ratio of $V \mid V > 0$ to $-V \mid V < 0$ increases with magnitude. Thus the former strictly stochastically dominates the latter. Since $\sigma(-v)$ decreases, the two mean difficulties on either side of $V = 0$ sum to less than one. The zero just identified lies below $d = 1/2$, corresponding to a strictly positive split threshold in V.

Increasing β decreases both conditional weighted means in Eq. (32). As $\beta \to \infty$, the negative-region mean tends to $-\infty$, while the positive-region mean tends to zero. This follows by comparing the weights on any two separated intervals, using strict monotonicity of σ and Gaussian integrability. Thus $\bar{v}_\beta < 0$ for a finite β. Likewise, as $\kappa \to \infty$, reference weighting concentrates on increasingly negative V. The context-component likelihood ratio is $e^{2\nu V}$, so $p_\kappa \to 0$; a finite κ gives $p_\kappa < 1 - p$. All other parameters and sources above can then be fixed at finite values. The feature covariances remain nonsingular and every group has positive probability. Taking two slightly different admissible values of ν and common sufficiently large finite β, κ gives distinct distributions P_1, P_2 with the same correction settings. Their initial and reference heads may differ as specified above. Steps 1–3 prove the two inequalities in Proposition 4.1.

C.3 Proof of Theorem 4.2: Worst-group accuracy of validation selection

We prove the finite-sample bound in Theorem 4.2 for the combined candidate pool, without assuming that either population correction is optimal. Sufficient conditions for the bound to vanish are given in Appendix C.5. Let $\widehat{f}$ minimize $\widehat{R}_\tau$ over the nonconstant linear candidates $\mathcal{F}_n$. Both validation labels are present. All population risks below evaluate the realized fitted functions on a fresh model draw.

Use the proof abbreviations

$$\Sigma = \mathrm{diag}(\Sigma_{\mathrm{task}}, \Sigma_{\mathrm{context}}), \quad \gamma = \sqrt{\mu^\top \Sigma_{\mathrm{task}}^{-1} \mu}, \quad \gamma_c = \sqrt{\nu^\top \Sigma_{\mathrm{context}}^{-1} \nu}.$$

For the arbitrary $c > 0$ in the theorem, set

$$g_c(Z) = c\mu^\top \Sigma_{\mathrm{task}}^{-1} Z_{\mathrm{task}}, \qquad S = c\gamma > 0, \qquad s = \max\{S, \|\widehat{f}\|_\Sigma\}.$$

These are comparison functions and scales used only in the proof. Write $q_f = \mathrm{Cov}(f(Z), f_{\mathrm{ref}}(Z) \mid Y, E)$ and $\omega_{\mathrm{ref}} = e^{-\kappa\sigma(Y f_{\mathrm{ref}}(Z))}$, so

$$p_\kappa = \frac{\mathbb{E}[\omega_{\mathrm{ref}} \mathbf{1}\{YE = 1\}]}{\mathbb{E}\omega_{\mathrm{ref}}}, \qquad q_{f_{u,v}} = u^\top \Sigma_{\mathrm{task}} u_{\mathrm{ref}} + v^\top \Sigma_{\mathrm{context}} v_{\mathrm{ref}}.$$

The imbalance coefficient in Theorem 4.2 is $b_\kappa = p + p_\kappa - 1$. The conditional variances and covariances are the same in every group. For a candidate $f = f_{u,v}$, put

$$B_f = \frac{1}{4}[(p + p_\kappa - 1)v^\top \nu]_+ + \frac{\kappa e^\kappa}{32}[-q_f]_+.$$

Step 1: control the effect of reference weighting. For a fixed candidate $f = f_{u,v}$, use the temporary abbreviations $a = u^\top \mu$, $b = v^\top \nu$, and $r = \|w\|_\Sigma > 0$. For $G \sim \mathcal{N}(0,1)$, put $Q_r(m) = \mathbb{E}\sigma(-m - rG)$. The ordinary population validation loss is

$$L_0(f) = pQ_r(a + b) + (1 - p)Q_r(a - b).$$

The weighted loss is $L_\kappa(f) = \mathbb{E}[\omega_{\mathrm{ref}}\sigma(-M_f)]/\mathbb{E}\omega_{\mathrm{ref}}$, where $\omega_{\mathrm{ref}} = e^{-\kappa\sigma(M_{\mathrm{ref}})}$. Both expressions equal the class-balanced population losses because the signed feature distribution is the same for both task labels.

Conditional on either value of YE, the candidate margin M and reference margin R are jointly Gaussian. Holding their means and variances fixed and differentiating with respect to their covariance q gives

$$\frac{d}{dq}\mathbb{E}[e^{-\kappa\sigma(R)}\sigma(-M)] = \kappa\mathbb{E}[e^{-\kappa\sigma(R)}\sigma'(R)\sigma'(M)] \geq 0.$$

Indeed, their joint density g_q satisfies $\partial_q g_q = \partial_M \partial_R g_q$; integrating by parts twice proves the identity. Bounded logistic derivatives and Gaussian tails remove the boundary terms. Singular endpoint covariances follow by continuity. Since $\mathbb{E}\omega_{\mathrm{ref}} \geq e^{-\kappa}$ and $0 \leq \sigma' \leq 1/4$, the derivative of the normalized weighted loss is at most $\kappa e^\kappa/16$.

At zero covariance, candidate and reference margins are conditionally independent. Their weighted mixture gives

$$L_\kappa(f) \geq p_\kappa Q_r(a + b) + (1 - p_\kappa)Q_r(a - b) - \frac{\kappa e^\kappa}{16}[-q_f]_+.$$

The normalized smooth maximum is at least the arithmetic average of its inputs. Moreover, Q_r is decreasing and $1/4$-Lipschitz, so $Q_r(a + b) - Q_r(a - b)$ has the opposite sign to b and absolute value at most $|b|/2$. Therefore

$$\begin{aligned} R_\tau(f) \geq \frac{Q_r(a + b) + Q_r(a - b)}{2} \\ - \frac{1}{4}[(p + p_\kappa - 1)b]_+ - \frac{\kappa e^\kappa}{32}[-q_f]_+. \end{aligned} \tag{38}$$

Step 2: relate balanced soft error to global WGA. The function $-Q_r'$ is the convolution of the logistic density σ' and the centered Gaussian density. It is even and nonincreasing on $[0,\infty)$. To verify the latter property, write the even decreasing logistic density as a mixture of indicators of centered intervals. Convolving an interval $[-h,h]$ with a Gaussian gives derivative $\varphi_r(x+h)-\varphi_r(x-h)\le 0$ for $x\ge 0$. Integrating the mixture proves the assertion. For $a\ge 0$, differentiating $\{Q_r(a+b)+Q_r(a-b)\}/2$ with respect to $|b|$ consequently gives a nonnegative derivative. For $a<0$, symmetry gives a lower bound of $1/2$. Hence

$$\frac{Q_r(a+b)+Q_r(a-b)}{2}\ge\begin{cases}Q_r(a), & a\ge 0,\\ 1/2, & a<0.\end{cases} \tag{39}$$

Put $t=a/r\in[-\gamma,\gamma]$. For $t>0$, $Q_r(rt)=\mathbb{E}\sigma[-r(t+G)]$ decreases strictly with r. Writing $X=t+G$ and pairing positive and negative x shows

$$\frac{d}{dr}Q_r(rt)=-\int_0^\infty x\sigma'(rx)\{\varphi(x-t)-\varphi(x+t)\}\,dx<0.$$

It is constant at $1/2$ when $t=0$.

We also need a quantitative bound in the mean. For any $s>0$, the evenness of σ' gives

$$\begin{aligned}\frac{-Q_s'(m)}{\varphi_s(m)}&=\int_{\mathbb{R}}\sigma'(x)e^{-x^2/(2s^2)}\cosh(mx/s^2)\,dx\\ &\ge e^{-1/2}\int_{-s}^{s}\sigma'(x)\,dx=e^{-1/2}\tanh(s/2)>\tfrac14\min\{1,s\}.\end{aligned}$$

The last step uses concavity of $\tanh$ on $[0,1/2]$, its monotonicity for larger arguments, and $e^{-1/2}\tanh(1/2)>1/4$. For a comparison scale $s\ge\max\{S,r\}$, if $t\ge 0$, use $r\le s$ and integrate this derivative bound from st to $s\gamma$. On that interval the standard normal density is at least $\varphi(\gamma)$. If $t<0$, integrate from zero to $s\gamma$ and use $\gamma-t\le 2\gamma$. Equation (39) then yields, in both cases,

$$\frac{Q_r(a+b)+Q_r(a-b)}{2}-Q_s(s\gamma)\ge\frac{\min\{1,S\}\varphi(\gamma)}{8}(\gamma-t). \tag{40}$$

To turn this into a WGA bound, define unit vectors only for this step:

$$e=\frac{\Sigma^{1/2}w}{r},\qquad e^\circ=\frac{(\Sigma_{\text{task}}^{-1/2}\mu,0)}{\gamma}.$$

Their inner product is t/γ, so $\|e-e^\circ\|^2=2(\gamma-t)/\gamma$. Also $|b|/r\le\gamma_c\|e-e^\circ\|$ and $\gamma-t\le\gamma\|e-e^\circ\|$. The WGA formula and the $1/\sqrt{2\pi}$ Lipschitz constant of Φ therefore give

$$\Phi(\gamma)-\mathrm{WGA}(f)\le\frac{\gamma+\gamma_c}{\sqrt{\pi\gamma}}\sqrt{\gamma-t}. \tag{41}$$

Combining Eqs. (38), (40), and (41) proves

$$\Phi(\gamma)-\mathrm{WGA}(f)\le(\gamma+\gamma_c)\sqrt{\frac{8\{R_\tau(f)-Q_s(s\gamma)+B_f\}}{\pi\gamma\min\{1,S\}\varphi(\gamma)}}. \tag{42}$$

The expression inside the square root is nonnegative by the preceding inequalities.

Step 3: compare with either fitted correction. For a coefficient difference h, Cauchy–Schwarz and the Gaussian first absolute moment give

$$\mathbb{E}|Yh^\top Z|\le\left(\sqrt{\gamma^2+\gamma_c^2}+\sqrt{2/\pi}\right)\|h\|_\Sigma.$$

The logistic soft error is $1/4$-Lipschitz, normalized reference weights are at most e^κ, and the smooth maximum is 1-Lipschitz in its inputs' maximum norm. Hence both R_τ and the ordinary soft error are Lipschitz in $\|\cdot\|_\Sigma$ with constant

$$L=\frac{e^\kappa}{4}\left(\sqrt{\gamma^2+\gamma_c^2}+\sqrt{2/\pi}\right).$$

The comparison score g_c has ordinary risk $Q_S(S\gamma)$ and conditional reference covariance $cf_{\mathrm{ref}}(\mu, 0)$. At zero covariance, its weighted risk equals its ordinary risk, because its conditional margin law is independent of context. The covariance derivative in Step 1 therefore gives

$$R_\tau(g_c) \le Q_S(S\gamma) + \frac{\kappa e^\kappa}{16} c[f_{\mathrm{ref}}(\mu, 0)]_+.$$

The score $(s/S)g_c$ has ordinary risk $Q_s(s\gamma)$, and its distance from g_c is $s - S$. The same Lipschitz bound gives $Q_S(S\gamma) - Q_s(s\gamma) \le L(s - S)$.

Empirical selection implies $R_\tau(\widehat{f}) \le R_\tau(g) + 2\varepsilon_{\mathrm{val}}$ for every $g \in \mathcal{F}_n$. Comparing each such g with g_c and minimizing their distance yields

$$R_\tau(\widehat{f}) - Q_s(s\gamma) \le L\varepsilon_{\mathrm{fit}} + \frac{\kappa e^\kappa}{16} c[f_{\mathrm{ref}}(\mu, 0)]_+ + 2\varepsilon_{\mathrm{val}}.$$

Thus the fitting term can be supplied by suppression, relearning, or the initial classifier. Substituting in Eq. (42) for $f = \widehat{f}$ proves Eq. (8) with

$$C_{\mathrm{model}} = (\gamma + \gamma_c)\sqrt{\frac{8\max\{L, 1/4, e^\kappa/16, 2\}}{\pi\gamma \min\{1, c\gamma\}\varphi(\gamma)}}.$$

In particular, the last reference term is needed when the reference uses useful task signal: weighting then changes even an optimal task head's soft error. It vanishes in both constructions of Appendix C.2, where $f_{\mathrm{ref}}(\mu, 0) = 0$ despite the noise component in the second reference.

C.4 Uniform convergence of the validation criterion

This section bounds the validation sampling error in Theorem 4.2.

The candidates and reference can depend on validation data, so sampling error must be controlled for every head that the procedure could produce. Let $\mathcal{H} \subseteq \mathcal{F}$ contain all possible fitted heads and reference heads. Both $\mathcal{H}$ and the feature extractor h_0 are specified before examining validation data; we may condition on the training data used to obtain them. Let $\mathcal{L}$ consist of $1 - p_f(Y \mid Z)$, $\omega_r[1 - p_f(Y \mid Z)]$, and ω_r, for $f, r \in \mathcal{H}$. Their averages give the ordinary loss and the numerator and denominator of the weighted loss.

For a finite sample $(z_i, y_i)_{i=1}^m$, where $z_i = h_0(x_i)$, approximate each $g \in \mathcal{L}$ by a representative function h satisfying

$$\frac{1}{m}\sum_{i=1}^m |g(z_i, y_i) - h(z_i, y_i)| \le a.$$

The uniform empirical covering number $\mathcal{N}_1(a, \mathcal{L})$ is the smallest upper bound, over all finite samples, on the minimum number of representatives needed. Each function in $\mathcal{L}$ takes values in $[0, 1]$. Pollard's covering-number inequality, as stated in Zhang (2002, Theorem 1), therefore gives, for m independent observations from any common distribution,

$$\mathbb{P}\left\{\sup_{g\in\mathcal{L}}\left|\frac{1}{m}\sum_{i=1}^m g(z_i, y_i) - \mathbb{E}g(Z, Y)\right| > t\right\} \le 8\,\mathcal{N}_1(t/8, \mathcal{L})e^{-mt^2/128}.$$

Given $n_{+1} > 0$ and $n_{-1} > 0$, the samples with label y are independent draws from the distribution conditional on $Y = y$. Apply the inequality separately for each label and add the two failure probabilities. Writing

$$q_{n,\delta} = \sqrt{\frac{128}{\min\{n_{+1}, n_{-1}\}}\log\frac{16\,\mathcal{N}_1(1/n, \mathcal{L})}{\delta}},$$

we have $q_{n,\delta}/8 \ge 1/n$ because $n \ge 2$, $\min\{n_{+1}, n_{-1}\} \le n/2$, and $\log(16\,\mathcal{N}_1(1/n, \mathcal{L})/\delta) \ge \log 16$. Monotonicity of covering numbers then bounds the total failure probability by

$$16\mathcal{N}_1(1/n, \mathcal{L})\exp\left(-\frac{\min\{n_{+1}, n_{-1}\}q_{n,\delta}^2}{128}\right) \le \delta.$$

Define the class-balanced empirical average by

$$\widehat{P}_{\mathrm{bal}} g = \frac{1}{2} \sum_{y \in \{-1,+1\}} \frac{1}{n_y} \sum_{i \in \mathcal{V}:\, y_i = y} g(z_i, y_i).$$

Because the model gives each task class probability $1/2$, the event above implies $\sup_{g \in \mathcal{L}} |\widehat{P}_{\mathrm{bal}} g - \mathbb{E} g| \le q_{n,\delta}$. This event holds simultaneously for all candidate and reference heads in $\mathcal{H}$, including those chosen using validation data.

The empirical weighted loss is

$$\widehat{L}_{\mathrm{val}}(f; \kappa) = \frac{\widehat{P}_{\mathrm{bal}}[\omega_{f_{\mathrm{ref}}}(1 - p_f(Y \mid Z))]}{\widehat{P}_{\mathrm{bal}} \omega_{f_{\mathrm{ref}}}}.$$

It lies in $[0, 1]$, and $\mathbb{E}\omega_{f_{\mathrm{ref}}} \ge e^{-\kappa}$. The identity

$$\begin{aligned}(\mathbb{E}\omega_{f_{\mathrm{ref}}})\{\widehat{L}_{\mathrm{val}}(f; \kappa) - L_{\mathrm{val}}(f; \kappa)\} \\ = (\widehat{P}_{\mathrm{bal}} - \mathbb{E})[\omega_{f_{\mathrm{ref}}}(1 - p_f(Y \mid Z))] \\ - \widehat{L}_{\mathrm{val}}(f; \kappa)(\widehat{P}_{\mathrm{bal}} - \mathbb{E})\omega_{f_{\mathrm{ref}}}\end{aligned}$$

gives $|\widehat{L}_{\mathrm{val}}(f; \kappa) - L_{\mathrm{val}}(f; \kappa)| \le 2e^{\kappa} q_{n,\delta}$. The ordinary loss has error at most $q_{n,\delta}$. Changing each input of a smooth maximum by at most a changes its value by at most a. Consequently,

$$\sup_{f \in \mathcal{H}} |\widehat{R}_\tau(f) - R_\tau(f)| \le 2e^{\kappa} q_{n,\delta}, \tag{43}$$

simultaneously for every reference head in $\mathcal{H}$.

Consequently, with probability at least $1 - \delta$, conditional on the positive class counts,

$$\varepsilon_{\mathrm{val}} \le 16\sqrt{2}\, e^{\kappa} \sqrt{\frac{\log\{16\mathcal{N}_1(1/n, \mathcal{L})/\delta\}}{\min\{n_{+1}, n_{-1}\}}}. \tag{44}$$

If $\log \mathcal{N}_1(1/n, \mathcal{L}) = o(n)$ and both class counts grow proportionally to n, the bound tends to zero with $\delta = 1/n$. For fixed feature dimension and bounded candidate and reference coefficients, finite coefficient nets and an integrable Gaussian Lipschitz envelope also give uniform convergence directly. The uniform bound permits the existing reuse of validation data for directions, regional shares, and final selection; no additional holdout is required.

C.5 Sufficient conditions for asymptotic optimality

C.5.1 General sufficient conditions

We give sufficient conditions for consistency under the general Gaussian feature model, using the notation of Appendix C.3. The feature dimension, initial and reference classifiers, correction grid, and number of relearning rounds are fixed. Let $\mathcal{F}_\infty$ be the finite candidate pool obtained by replacing the empirical quantities in GEAR by their population counterparts. It includes the initial classifier and both correction pools. The conditions below concern their fitting objectives and validation losses; convergence of the final predictor will be derived from them.

Condition 1: a correction supplies the comparison score. Write $T = YZ$ and let $w_c = c(\Sigma_{\mathrm{task}}^{-1}\mu, 0)$ be the coefficients of g_c. Let D collect the constraint directions for a retained full-suppression candidate, so its feasible coefficients satisfy $D^\top w = 0$. For a retained relearning round, let ρ denote its actual normalized population training weight, with $\mathbb{E}\rho = 1$. Let $A_s, A_h \succ 0$ denote the quadratic penalty matrices. In the fitting coordinates they are $\lambda_s I$ and $\lambda_h I$; in original coordinates they also include the feature scales. For some fixed $c > 0$, require at least one of the following moment conditions:

$$D^\top w_c = 0, \qquad A_s w_c - \mathbb{E}[T\sigma(-w_c^\top T)] \in \mathrm{col}(D), \tag{45}$$

$$A_h(w_c - w_0) = \mathbb{E}[\rho T \sigma(-w_c^\top T)]. \tag{46}$$

These conditions use the directions and weights produced by GEAR. Equation (45) says that the gradient of the suppression objective is perpendicular to every feasible direction. Equation (46) sets the gradient of the anchored relearning objective to zero. The objectives are strongly convex, so the corresponding fitted classifier is uniquely g_c. Thus either correction can supply $g_c \in \mathcal{F}_\infty$.

Condition 2: validation favors this score. Require

$$\kappa f_{\mathrm{ref}}(\mu, 0) = 0. \tag{47}$$

For $\kappa > 0$, the reference may use context and task noise, but not the useful task signal. This condition is necessary for the present reference error to vanish at g_c: direct substitution gives $\varepsilon_{\mathrm{ref}}(g_c) = \kappa c|f_{\mathrm{ref}}(\mu, 0)|$. By Step 3 of Appendix C.3, Eq. (47) also gives $R_\tau(g_c) = Q_S(S\gamma)$.

For every competing population candidate $f = f_{u,v} \neq g_c$, use $a = u^\top \mu$, $b = v^\top \nu$, and $r = \|f\|_\Sigma$ as in Step 1 of Appendix C.3, and require

$$\frac{Q_r(a+b) + Q_r(a-b)}{2} - B_f > Q_S(S\gamma), \qquad f \in \mathcal{F}_\infty \setminus \{g_c\}. \tag{48}$$

Here B_f is the explicit reference correction defined before Step 1 of Appendix C.3; all terms in this condition are one-dimensional Gaussian expectations or moments of the fitted coefficients. Equation (38) then proves $R_\tau(f) > R_\tau(g_c)$ for every such candidate. The comparison covers the initial classifier and both corrections. Candidates with the same score g_c may tie.

A simple sufficient way to check Eq. (48) is to verify

$$a \leq c\gamma^2, \qquad (p + p_\kappa - 1)b \leq 0, \qquad q_f \geq 0.$$

These inequalities give $B_f = 0$. If $a \leq 0$, the balanced soft error is at least $1/2 > Q_S(S\gamma)$. If $a > 0$, Cauchy–Schwarz gives $r \geq a/\gamma$, and adding Gaussian noise increases $Q_r(a)$: $\partial_r Q_r(a) = rQ_r''(a) > 0$ for $a, r > 0$. Together with Step 2 of Appendix C.3 this gives

$$\frac{Q_r(a+b) + Q_r(a-b)}{2} \geq Q_r(a) \geq Q_{a/\gamma}(a) \geq Q_S(S\gamma).$$

The last inequality is strict if $a < c\gamma^2$. If $a = c\gamma^2$ but $f \neq g_c$, nonsingularity of Σ gives $r > a/\gamma$, so the middle inequality is strict. The more general condition (48) also allows violations of these three simpler inequalities when the remaining loss difference is large enough.

Consistency of the selected predictor. Let the fitting and validation sample sizes tend to infinity, with each task class having a positive limiting fraction in each sample. Assume the population direction subspaces are separated by nonzero eigenvalue gaps, the population splits are unique and nondegenerate, and every used region has positive probability and a boundary of probability zero. All fitting penalties are positive and fixed. If a count-dependent rule has several possible limiting choices, require both conditions above for each choice with the same c.

These regularity conditions give consistency of the entire fitted pool. Empirical means and feature scales converge by the law of large numbers. Within-class rank directions are order-two averages with integrable Gaussian kernels; their convergence and the eigenvalue gaps give convergent direction subspaces. Unique splits and zero-probability boundaries give convergent regions and weighted moments. Positive region masses keep the weight normalizations away from zero. Gaussian integrability then gives uniform convergence of the fitting objectives on bounded coefficient sets.

The positive penalties bound all fitted coefficients: comparing each objective with its value at zero bounds its quadratic penalty. Strong convexity gives unique population minimizers. For suppression, convergent constraint subspaces allow a feasible population coefficient to be projected to a convergent empirical feasible coefficient, and conversely. Uniform convergence of the objectives therefore gives convergence of their minimizers. For relearning, apply the same argument successively over the fixed number of rounds; uniform validation convergence also gives convergence of the regional-error updates and hence the next round's weights. All empirical candidates thus converge to their population counterparts simultaneously.

Appendix C.4 gives $\varepsilon_{\mathrm{val}} \xrightarrow{\mathbb{P}} 0$ even though the fitted candidates use validation data. Since $\mathcal{F}_\infty$ is finite, the strict validation comparisons above give a positive loss gap between g_c and every distinct competing score. The Lipschitz bound in Step 3 of Appendix C.3 and convergence of the fitted pool transfer this gap to the empirical validation losses with probability tending to one. Consequently, any minimizer $\widehat{f}$ of the empirical validation loss satisfies

$$\|\widehat{f} - g_c\|_\Sigma \xrightarrow{\mathbb{P}} 0. \tag{49}$$

Ties between candidates converging to g_c do not affect this conclusion. For finitely many limiting choices of a count-dependent rule, take the minimum loss gap and the maximum estimation error across those choices.

It remains to show that the bound itself vanishes. Since $\widehat{f} \in \mathcal{F}_n$ and $\|g_c\|_\Sigma = c\gamma$, the triangle inequality gives

$$\varepsilon_{\mathrm{fit}} \le 2\|\widehat{f} - g_c\|_\Sigma \xrightarrow{\mathbb{P}} 0.$$

Also $g_c(0,\nu) = 0$ and $\kappa \operatorname{Cov}(g_c(Z), f_{\mathrm{ref}}(Z) \mid Y, E) = 0$ by Eq. (47). Using $|p + p_\kappa - 1| \le 1$ and Cauchy–Schwarz yields

$$\varepsilon_{\mathrm{ref}} \le (\gamma_c + \kappa\|f_{\mathrm{ref}}\|_\Sigma)\|\widehat{f} - g_c\|_\Sigma \xrightarrow{\mathbb{P}} 0.$$

Thus all three errors vanish. The model parameters, κ, and c are fixed, so C_{model} is fixed. Theorem 4.2 therefore gives

$$\mathrm{WGA}(\widehat{f}) \xrightarrow{\mathbb{P}} \Phi\left(\sqrt{\mu^\top \Sigma_{\mathrm{task}}^{-1} \mu}\right) = \max_{f \in \mathcal{F}} \mathrm{WGA}(f).$$

This conclusion applies to the final predictor selected from the whole pool, whether suppression or relearning supplies the successful candidate.

C.5.2 Applications to the two correction regimes

We apply the validation and convergence arguments to the two configurations in Appendix C.2, with the model, sampling, and correction settings specified there.

Population validation selection. First take $\epsilon = 0$. The reference is Z_c, the source is a positive multiple of Z_c, and the suppression endpoint is $c_s Z_1$. Every suppression interpolation has form $cZ_1 + vZ_c$, with $0 \le c \le c_s$ and $v \ge 0$. Its conditional covariance with the reference is $v \ge 0$. The reference condition in Eq. (32) and the Gaussian comparison in Eq. (38) imply

$$R_\tau(cZ_1 + vZ_c) \ge \frac{Q_{\sqrt{c^2+v^2}}(cm + \nu v) + Q_{\sqrt{c^2+v^2}}(cm - \nu v)}{2} \ge Q_c(cm) = q(c) \ge q(c_s),$$

where $Q_a(b) = \mathbb{E}\sigma(-b - aG)$ for $G \sim \mathcal{N}(0,1)$. For the middle inequality, a symmetric shift increases Q_a at a nonnegative mean by Eq. (39). Increasing Gaussian variance also increases it: $\partial_a Q_a(b) = aQ_a''(b) \ge 0$ for $b \ge 0$, because $-Q_a'$ is even and decreasing on the positive axis. The zero-variance case follows by continuity.

The useful task feature is independent of the reference, so $R_\tau(c_h Z_1) = q(c_h) < q(c_s)$. This gives a strict validation gap between relearning and the entire suppression path at $\epsilon = 0$. The split, the moment-balanced source, the strongly convex constrained fit, and the validation risks are continuous for small ϵ. Continuity is uniform over the compact interpolation interval. The gap therefore persists for sufficiently small positive ϵ. Every relearning round remains $c_h Z_1$, so population validation selects a globally optimal head in the regime where suppression cannot attain it.

In the suppression regime of Eq. (33), full suppression fits $c_s Z_1$, as shown in Appendix C.2.

The preceding inequality shows that its validation score $q(c_s)$ is strictly smaller than those of the other suppression interpolations: their task coefficients are $\alpha c_s < c_s$.

Using the large-B_0 coefficient bounds established in Appendix C.2, the ordinary and weighted soft errors of the relearning heads tend, respectively, to $\mathbb{P}(V < 0)$ and the reference-weighted probability of $V < 0$.

Decreasing weights increase this event's conditional probability within each context component. Writing $a = \Phi(-\nu) < 1/2$, their average is therefore at least

$$\frac{(p + p_\kappa)a + (2 - p - p_\kappa)(1 - a)}{2} \ge \frac{1}{2}.$$

The last inequality uses $p + p_\kappa \le 1$. The smooth maximum is at least the average, whereas $q(c_s) < 1/2$. For sufficiently large finite B_0, all relearning rounds have strictly worse validation score and WGA than full suppression. Hence validation selects the globally optimal suppression head.

Empirical selection in the two regimes. Fix either regime, its feature map and heads, all settings, and the finite round count. Let the per-class fitting and validation counts grow.

The nonzero rank-one directions, unique nondegenerate splits, and positive penalties in Appendix C.2 satisfy the regularity requirements for the fitted-pool consistency argument in Appendix C.5.1. That argument gives convergence of the empirical feature scales, directions, regions, weighted moments, and strongly convex fits to their population counterparts.

In the relearning regime, every population regional error is $q(c_h)$. Induction over the finite rounds therefore gives convergence of empirical region shares to equal shares, and of all relearned heads to $c_h Z_1$. In the suppression regime, the same induction gives convergence to the corresponding population rounds.

The uniform bound in Appendix C.4 controls the selected heads, including their dependence on validation data.

The strict population validation gaps established above imply selection of the successful branch with probability tending to one. Within the suppression pool, full suppression uniquely minimizes population validation risk; for a continuous interpolation interval the same conclusion follows from compactness and uniqueness. Thus $\widehat{f}$ converges in coefficients to $c_s Z_1$ in the suppression regime and to $c_h Z_1$ in the relearning regime.

Continuity of WGA at these nonzero limits gives

$$\mathrm{WGA}(\widehat{f}) \xrightarrow{\mathbb{P}} \Phi(m) = \max_{f \in \mathcal{F}} \mathrm{WGA}(f).$$

Removing either correction leaves a strictly suboptimal population pool in one of these two distributions. The general consistency argument for the selected classifier is given in Appendix C.5.1.