A. Supplemental experiments

A.1. ERM models have poor worst-group error regardless of the degree of overparameterization

In the main text, we focused on reweighted models, trained with the reweighted objective on the full data (Sections 3-5), as well as subsampled models, trained on subsampled data with the ERM objective (Section 6). Here, we study the effect of overparameterization on ERM models, trained with the ERM objective on the full data. Consistent with prior work, we observe that ERM models obtain poor worst-group error (near or worse than random), regardless of whether the model is underparameterized or overparameterized (Sagawa et al., 2020). We also confirm that overparameterization helps average test error (see, e.g., Nakkiran et al. (2019); Belkin et al. (2019); Mei & Montanari (2019)).

Empirical results. We first consider the CelebA and Waterbirds dataset, following the experimental set-up of Section 3 but now training with the standard ERM objective (Equation (2)) instead of the reweighted objective (Equation (3)).

On these datasets, overparameterization helps the average test error (Figure 8). As model size increases past the point of zero training error, the average test error decreases. The best average test error is obtained by highly overparameterized models with zero training error—4.6% for CelebA at width 96, and 4.2% for Waterbirds at 6,000 random features.

In contrast, the worst-group error is consistently high across model sizes: it is consistently worse than random (>50%) for CelebA and nearly random (44%) for Waterbirds (Figure 8). These worst-group errors are much worse than those obtained by reweighted, underparameterized models (25.6% for CelebA and 26.6% for Waterbirds; see Section 3). Thus, while overparameterization helps ERM models achieve better test error, these models all fail to yield good worst-group error regardless of the degree of overparameterization.

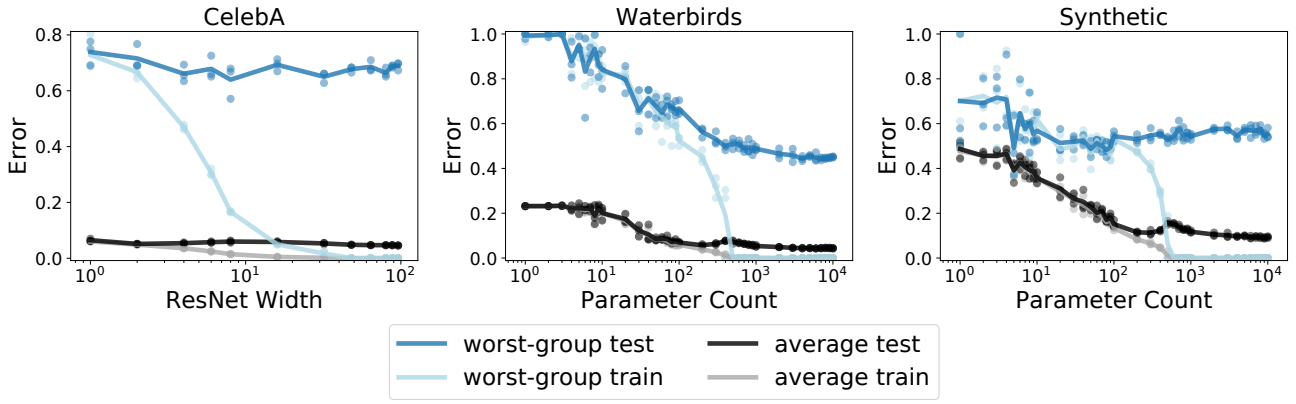

Figure 8. The effect of overparameterization on the average and worst-group error of an ERM model. Increasing model size helps average test error, but worst-group error remains poor across model sizes.

Simulation results. We also evaluate the effect of overparameterization on ERM models on the synthetic dataset introduced in Section 4. As above, ERM models fail to achieve reasonable worst-group test error across model sizes, but improve in average test error as model size increases (Figure 8). The best average test error is obtained by a highly overparameterized model with zero training error—9.0% error at 9,000 random features—while the worst-group test error is nearly random or worse ($> 48\%$) across model sizes.

A.2. Stronger L_2 regularization improves worst-group error in overparameterized reweighted models

In the main text, we studied models with default/weak or no L_2 regularization. In this section, we study the role of L_2 regularization in modulating the effect of overparameterization on worst-group error by changing the hyperparameter λ that controls L_2 regularization strength. Overall, we find that increasing L_2 regularization (to the point where models do not have zero training error) improves worst-group error but hurts average error in overparameterized reweighted models. In contrast, L_2 regularization has little effect on both worst-group and average error in the underparameterized regime.

Strong L_2 regularization improves worst-group error in overparameterized reweighted models. In the main text, we trained ResNet10 models with default, weak regularization ($\lambda = 0.0001$) on the CelebA dataset, and unregularized logistic regression on the Waterbirds and synthetic datasets. Here, we consider strongly-regularized models with $\lambda = 0.1$ for both types of models; unlike before, these models no longer achieve zero training error even when overparameterized. Figure 9 shows the results of varying model size on strongly-regularized ERM, reweighted, and subsampled models on the three datasets.

On all three datasets, with strong regularization, ERM models continue to yield poor worst-group test error across model sizes, with similar or worse worst-group test error compared to with weak/ no regularization. Conversely, strongly-regularized subsampled models continue to achieve low worst-group test error across model sizes.

Where strong regularization has a large effect is on reweighted models. With reweighting, we find that strong regularization improves worst-group error in overparameterized models: across all three datasets, the worst-group test error in the overparameterized regime is much lower for the strongly-regularized models than their weakly regularized or unregularized counterparts (Figure 3). These results are consistent with similar observations made in Sagawa et al. (2020). However, even though strongly-regularized overparameterized models outperform weakly-regularized overparameterized models, overparameterization can still hurt the worst-group error in strongly-regularized reweighted models. On the CelebA and synthetic datasets, with $\lambda = 0.1$, the best worst-group error is still obtained by an underparameterized model for the CelebA and synthetic datasets, though overparameterization seems to help worst-group error on the Waterbirds dataset at least in the range of model sizes studied.

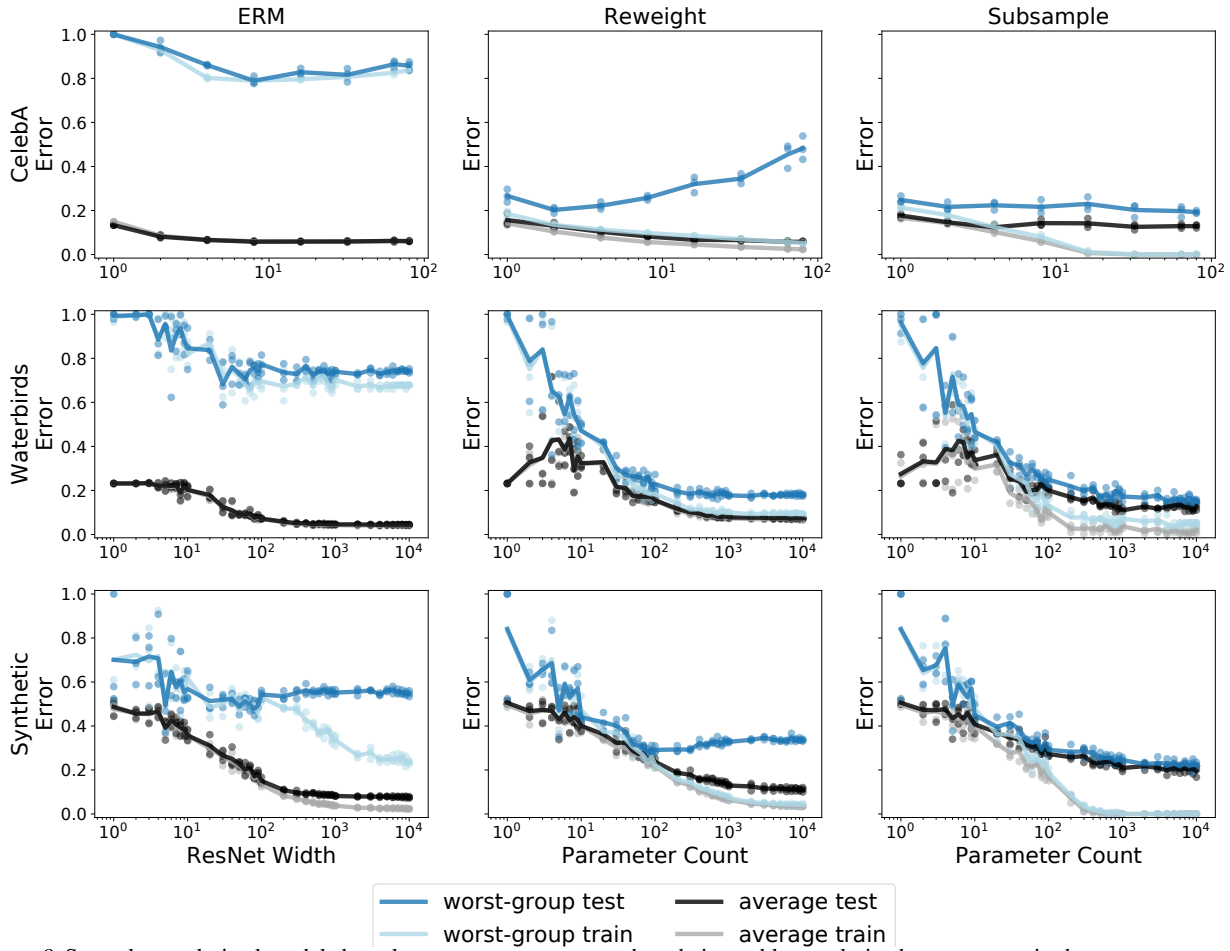

Figure 9. Strongly-regularized models have lower worst-group error than their weakly-regularized counterparts in the overparameterized regime (Figure 3). Even under strong regularization, increasing model size can hurt the worst-group error on the CelebA (top) and synthetic (bottom) datasets, although overparameterization seems to improve worst-group error in the Waterbirds datase (middle) for the range of model sizes studied.

Overparameterized models require strong regularization for worst-group test error but not average test error. Given a fixed overparameterized model size, how does its performance change with the L_2 regularization strength λ? We study this with the logistic regression model on the Waterbirds and synthetic datasets, using a model size of $m = 10,000$ random features and varying the L_2 regularization strength from $\lambda = 10^{-9}$ to $\lambda = 10^2$. [1]

Results are in Figure 10. As before, ERM models obtain poor worst-group error regardless of the regularization strength, and subsampled models are relatively insensitive to regularization, achieving reasonable worst-group error at most settings of λ.

For reweighted models, however, having the right level of regularization is critical for obtaining good worst-group test error. On both datasets, the best worst-group test error is obtained by strongly-regularized models that do not achieve zero training error. In contrast, increasing regularization strength hurts average error, with the best average test error attained by models with nearly zero regularization.

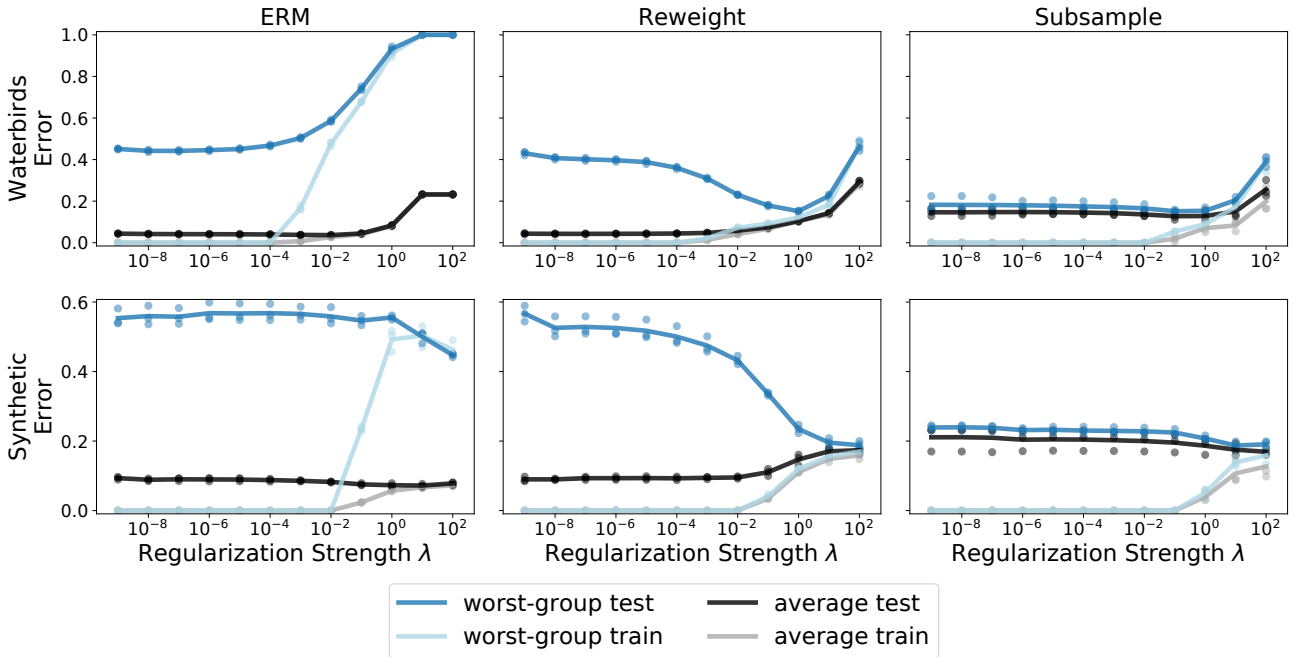

Figure 10. The effect of regularization on overparameterized random features logistic regression models ($m = 10,000$). ERM models (left) do consistently poorly while subsampled models (right) do consistently well on worst-group error. For reweighted models (middle), the best worst-group error is obtained by a strongly-regularized model that does not achieve zero training error.

L_2 regularization affects where worst-group test error plateaus as model size increases. In the above experiments, we kept either model size or regularization strength fixed, and varied the other. Here, we vary both: we consider L_2 regularization strengths $\lambda \in \{10^{-9}, 10^{-6}, 0.001, 0.1, 10\}$ and investigate the effect of increasing model size for each λ. We plot the results for Waterbirds and the synthetic dataset in Figure 11 and Figure 12 respectively.

For reweighted models, the results match what we observed above. Strengthening L_2 regularization reduces the detrimental effect of overparameterization on worst-group error. For any fixed model size in the overparameterized regime, the worst-group test error improves as λ increases up to a certain value. Worst-group test error seems to plateau at different values as model size increases, depending on the regularization strength, though we note that it is possible that further increasing model size beyond the range we studied might lead models with different regularization strengths to eventually converge. Further empirical studies as well as theoretical characterization of the interaction between regularization and overparameterization are needed to confirm this phenomenon.

Given sufficiently large λ (e.g., $\lambda = 10$ for both Waterbirds and synthetic datasets), overparameterized models seem to

[1] We did not run this experiment on the CelebA dataset for computational reasons, as doing so would have required tuning a different learning rate for each choice of regularization strength.

outperform underparameterized models, at least for the range of model sizes studied. However, we caution that this trend does not seem to hold on the CelebA dataset (Figure 9).

Finally, in contrast with its effects on overparameterized models, regularization seems to only have a modest effect on worst-group test error in the underparameterized regime.

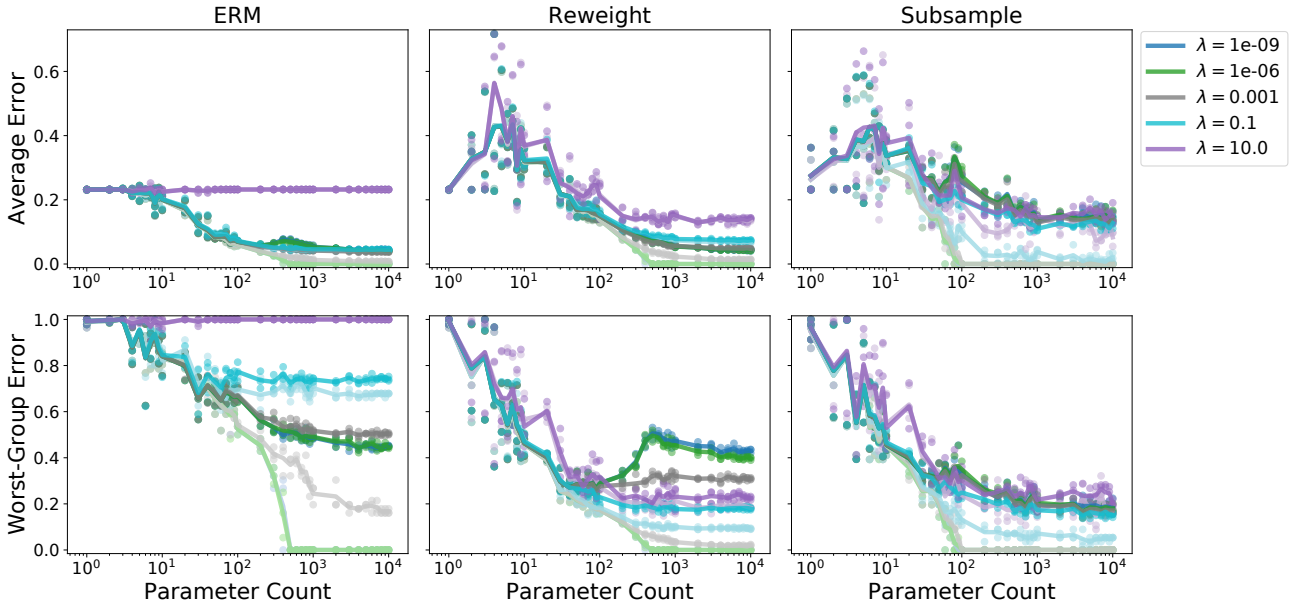

Figure 11. The effect of overparameterization on models with different L_2 regularization strengths λ on the Waterbirds dataset. Different regularization strengths are shown in different colors, with training and test errors plotted in light and dark colors, respectively.

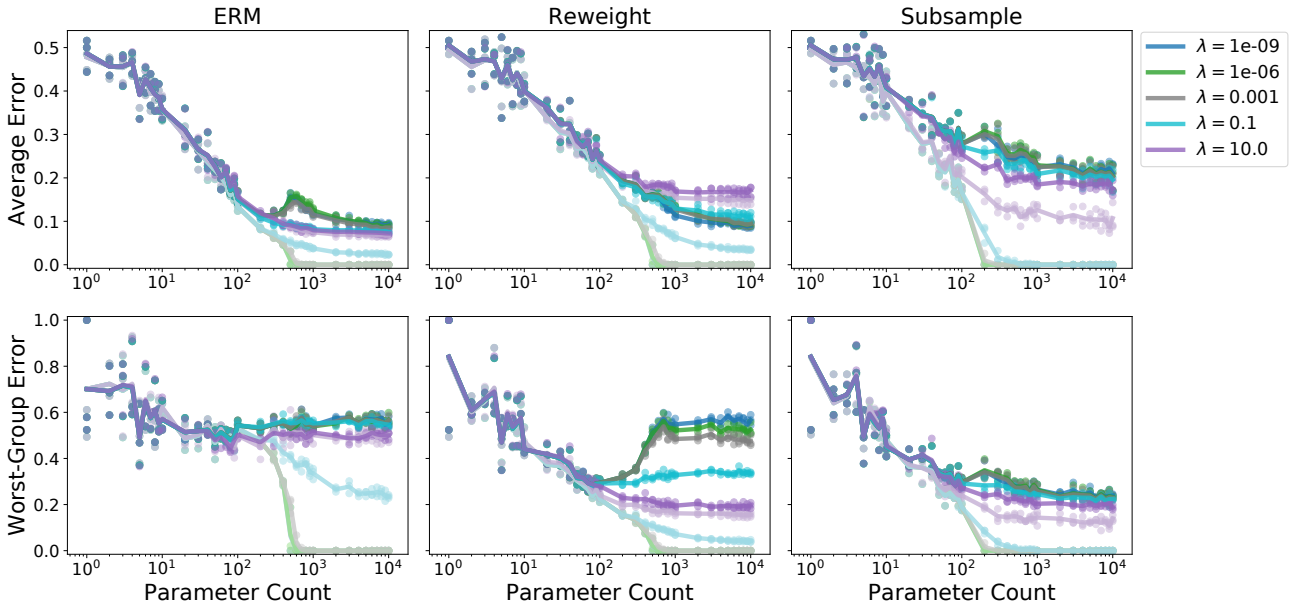

Figure 12. The effect of overparameterization on models with different L_2 regularization strengths λ on the synthetic dataset. The plotting scheme follows that of Figure 11.

A.3. Overparameterization helps average test error on the synthetic data regardless of p_{maj} and $r_{\text{s:c}}$

Figure 13 shows how the average test error changes as a function of model size under different settings of the majority fraction p_{maj} and the spurious-core ratio $r_{\text{s:c}}$ on the synthetic dataset introduced in Section 4. As expected, overparameterization helps the average test error regardless of SCR and the majority fraction.

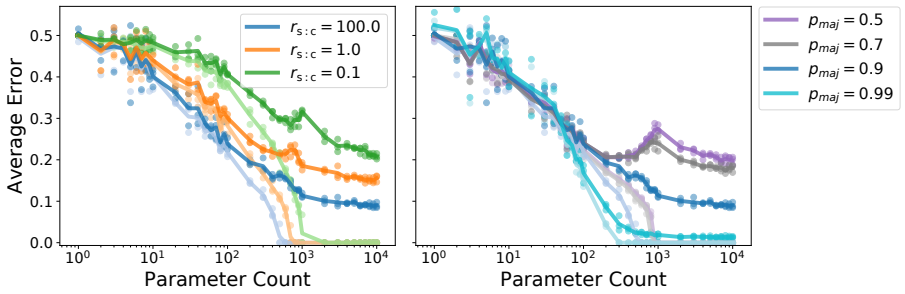

Figure 13. The effect of overparameterization on average error of a reweighted model on synthetic data. Different values of p_{maj} and $r_{\text{s:c}}$ are plotted in different colors, with training and test errors plotted in light and dark colors, respectively. Across all values of p_{maj} and $r_{\text{s:c}}$, overparameterization helps the average test error.

A.4. Comparison between implicit and explicit implicit memorization

To motivate the explicit-memorization setting, we ran some brief experiments to show that in the overparameterized regime, linear models in the explicit-memorization setting behave similarly to random projection (RP) models in the implicit-memorization setting, with σ^2_{core} and σ^2_{spu} in the latter scaled up by a factor of d (Figure 14). Recall that in the latter, $x_{\text{core}} \in \mathbb{R}^d$ is distributed as $x_{\text{core}}|y \sim \mathcal{N}(y, \sigma^2_{\text{core}} I_d)$. Roughly speaking, all the information about y is contained in the mean $\bar{x}_{\text{core}} = \frac{1}{d}\sum_j x_{\text{core},j}$, which is distributed as $\mathcal{N}(y, \sigma^2_{\text{core}} I_d/d)$. In the explicit-memorization setting, we can view $x_{\text{core}} \in \mathbb{R}$ as equivalent to $\bar{x}_{\text{core}}$ in the implicit-memorization setting (and similarly for x_{spu}), explaining the quantitative fit observed in Figure 14.

However, in the highly underparameterized regime, the RP models do poorly because of model misspecification (owing to a small number of random projections), whereas the linear models can still learn to use x_{core} and therefore do well.

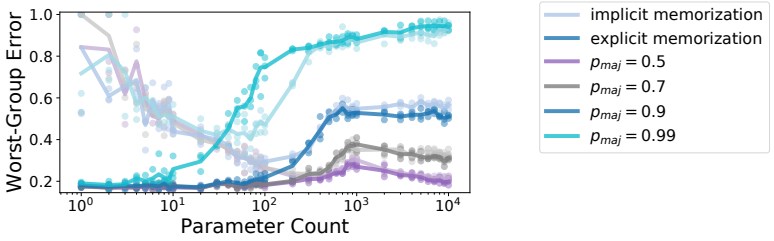

Figure 14. The effect of overparameterization on the worst-group test error for linear models in the explicit-memorization setting ($\sigma^2_{\text{core}} = 1, \sigma^2_{\text{spu}} = 0.01, \sigma^2_{\text{noise}} = 1$) and random projection models in the implicit-memorization setting ($\sigma^2_{\text{core}} = 100, \sigma^2_{\text{spu}} = 1, d = 100$). The models agree in the overparameterized regime.

A.5. Experimental details

Waterbirds and CelebA datasets. For the CelebA dataset, we use the official train-val-test split from Liu et al. (2015), with the *Blond_Hair* attribute as the target y and the *Male* as the spurious association a.

For the Waterbirds dataset, we follow the setup in Sagawa et al. (2020); for convenience, we reproduce some details of how it was constructed here. This dataset was obtained by combining bird images from the CUB dataset (Wah et al., 2011) with backgrounds from the Places dataset (Zhou et al., 2017). The CUB dataset comes with annotations of bird species. For the Waterbirds dataset, each bird was labeled was a waterbird if it was a seabird or waterfowl in the CUB dataset; otherwise, it was labeled as a landbird. Bird images were cropped using the provided segmentation masks and placed on either a land (bamboo forest or broadleaf forest) or water (ocean or natural lake) background obtained from the Places dataset.

For Waterbirds, we follow the same train-val-test split as in Sagawa et al. (2020). Note that in these validation and test sets,

landbirds and waterbirds are uniformly distributed on land and water backgrounds so that accuracy on the rare groups can be more accurately estimated. When calculating average test accuracy, we therefore first compute the average test accuracy over each group and then report a weighted average, with weights corresponding to the relative proportion of each group in the skewed training dataset.

We post-process Waterbirds by extracting feature representations taken from the last layer of a ResNet18 model pre-trained on ImageNet. We use the Pytorch `torchvision` implementation of the ResNet18 model for this. All models on the Waterbirds dataset in our paper are logistic regression models trained on top of this (fixed) feature representation.

ResNet. We used a modified ResNet10 with variable widths, following the approach in Nakkiran et al. (2019) and extending the `torchvision` implementation. We trained all ResNet10 models with stochastic gradient descent with momentum of 0.9 and a batch size of 128, with the L_2 regularization parameter λ was passed in to the optimizer as the weight decay parameter. In the experiments in the main text, we used the default setting of $\lambda = 10^{-4}$. We used a fixed learning rate instead of a learning rate schedule and selected the largest learning rate for which optimization was stable, following Sagawa et al. (2020). This resulted in learning rates of 0.01 and 0.0001 for $\lambda = 10^{-4}$ and $\lambda = 0.1$, respectively, across all training procedures. As in the original ResNet paper (He et al., 2016), we used batch normalization (Ioffe & Szegedy, 2015) and no dropout (Srivastava et al., 2014), and for simplicity, we trained all models without data augmentation.

We trained for 50 epochs for ERM and reweighted models and 500 epochs for subsampled models (due to smaller number of examples per epoch). We found that worst-group error can be unstable across epochs due to the small sample size and relatively large learning rate, so in our results we report the error averaged over the last 10 epochs.

Logistic regression. We used the logistic regression implementation from `scikit-learn`, training with the L-BFGS solver until convergence with tolerance 0.0001, and setting the regularization parameter as $C = 1/(n\lambda)$. For unregularized models, we set $\lambda = 10^{-9}$ for numerical stability.

A.6. Subsampling

Formally, given a set of groups $\mathcal{G}$ and a dataset D comprising a set of n training points with their group identities $\{(x^{(i)}, y^{(i)}, g^{(i)})\}$, the subsampling procedure involves two steps. First, we group training points based on group identities:

$$\mathrm{D}_g \stackrel{\text{def}}{=} \{(x^{(i)}, y^{(i)}) \mid g^{(i)} = g\} \text{ for each } g \in \mathcal{G}. \tag{13}$$

For each group g, we select a subset $\mathrm{D}_g^{\mathsf{ss}} \subseteq \mathrm{D}_g$ uniformly at random from D_g such that each subset has the same number of points as the smallest group in the training set. We form a new dataset D^{ss} by combining these subsets:

$$\begin{aligned} \mathrm{D}^{\mathsf{ss}} &= \bigcup_{g \in \mathcal{G}} \mathrm{D}_g^{\mathsf{ss}}, \text{ where} \\ \mathrm{D}_g^{\mathsf{ss}} \subseteq \mathrm{D}_g \text{ and } |\mathrm{D}_g^{\mathsf{ss}}| &= \min_{g \in \mathcal{G}} |\mathrm{D}_g| \end{aligned} \tag{14}$$

Note that D^{ss} is group-balanced, with $p_{\mathsf{maj}} = 0.5$. We then train a model by minimizing the average loss on D^{ss},

$$\hat{\mathcal{R}}_{\mathsf{subsample}}(w) \stackrel{\text{def}}{=} \frac{1}{|\mathrm{D}^{\mathsf{ss}}|} \sum_{(x,y) \in \mathrm{D}^{\mathsf{ss}}} \ell(w; (x, y)). \tag{15}$$

Since D^{ss} is group-balanced, the reweighted training loss (Equation 3) has the same weight on all training points and minimizing the reweighted objective on D^{ss} is equivalent to minimizing the average loss objective above.

B. Proof of Theorem 1

Here, we detail the proof of Theorem 1 presented in Section 5. We structure the proof by splitting Theorem 1 into two smaller theorems: one for the overparameterized regime (Appendix B.2), and another for the underparameterized regime (Appendix B.3).

B.1. Notation and definitions.

We denote the separate components of the weight vector $\hat{w}_{\mathsf{core}} \in \mathbb{R}, \hat{w}_{\mathsf{spu}} \in \mathbb{R}, \hat{w}_{\mathsf{noise}} \in \mathbb{R}^N$ such that

$$\hat{w} = [\hat{w}_{\mathsf{core}}, \hat{w}_{\mathsf{spu}}, \hat{w}_{\mathsf{noise}}]. \tag{16}$$

Further, by the representer theorem, we decompose $\hat{w}_{\mathsf{noise}}$ as

$$\hat{w}_{\mathsf{noise}} = \sum_{i=1}^{n} \alpha^{(i)}(\hat{w}) x^{(i)}_{\mathsf{noise}}. \tag{17}$$

Note that $\alpha^{(i)}(w)$ is equivalent to the $\alpha^{(i)}$ referred to in the main text. Recall that we define memorization of each training point $x^{(i)}$ by the weight $\alpha^{(i)}$ as follows.

Definition 2 (γ-memorization)**.** *Consider a separator $\hat{w}$ on training data $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$. For some constant $\gamma \in \mathbb{R}$, we say that a model γ-memorizes a training point if*

$$\left|\alpha^{(i)}(\hat{w})\right| > \frac{\gamma^2}{\sigma^2_{\mathsf{noise}}}. \tag{18}$$

The component $\alpha^{(i)}(\hat{w}) x^{(i)}_{\mathsf{noise}}$ serves to "memorize" $x^{(i)}$ when N is sufficiently large, as it affects the prediction on $x^{(i)}$ but not on any other training or test points (because noise vectors are nearly orthogonal when N is large). In the proof, we set the constant γ^2 appropriately (based on other parameter settings in Theorem 1) to get the required result.

Finally, let $G_{\mathsf{maj}}, G_{\mathsf{min}}$ denote the indices of training points in the majority and minority group respectively.

B.2. Overparameterized regime

In our explicit-memorization set-up, sufficiently overparameterized models provably have high worst-group error under certain settings of $\sigma^2_{\mathsf{spu}}, \sigma^2_{\mathsf{core}}, n_{\mathsf{maj}}, n_{\mathsf{min}}$ as stated in Theorem 1 (restated below as Theorem 2).

Theorem 2. *For any $p_{\mathsf{maj}} \geq \left(1 - \frac{1}{2001}\right)$, $\sigma^2_{\mathsf{core}} \geq 1$, $\sigma^2_{\mathsf{spu}} \leq \frac{1}{16 \log 100 n_{\mathsf{maj}}}$, $\sigma^2_{\mathsf{noise}} \leq \frac{n_{\mathsf{maj}}}{600^2}$ and $n_{\mathsf{min}} \geq 100$, there exists N_0 such that for all $N > N_0$ (overparametrized regime), with high probability over draws of the data,*

$$\mathit{Err}_{\mathsf{wg}}(\hat{w}^{\mathsf{mm}}) \geq 2/3, \tag{19}$$

where $\hat{w}^{\mathsf{mm}}$ is the max-margin classifier.

In Section 5, we sketched key ideas in the proof by considering special families of separators: because the minimum-norm inductive bias favors less memorization, models can prefer to learn the spurious feature and memorize the minority examples (entailing high worst-group error), instead of learning the core feature and memorizing some fraction of all training points (possibly attaining reasonable worst-group error). We now provide the full proof of Theorem 2, generalizing the above key concepts by considering *all* separators.

Proof. Recall from Section 5 that we consider the maximum-margin classifier $\hat{w}^{\mathsf{minnorm}}$:

$$\hat{w}^{\mathsf{minnorm}} = \arg\min \|w\|_2^2 \text{ s.t. } y^{(i)}(w \cdot x^{(i)}) \geq 1, \ \forall i. \tag{20}$$

In other words, $\hat{w}^{\mathsf{minnorm}}$ is the minimum-norm separator, where separator is a classifier with zero training error and required margins, satisfying $y^{(i)}(w \cdot x^{(i)}) \geq 1$ for all i. We analyze the worst-group error of the minimum-norm separator $\hat{w}^{\mathsf{minnorm}}$ as outlined below:

1. We first upper bound the fraction of *majority* examples memorized by the minimum-norm separator $\hat{w}^{\mathsf{minnorm}}$. We show that there exists a separator that can use spurious features and needs to memorize only the minority points (Lemma 1) for the parameter settings in Theorem 2 where σ_{spu} is sufficiently small. Since the norm of a separator is roughly scales with the number of points memorized ($|\alpha^{(i)}(\hat{w})| \geq \gamma^2/\sigma^2_{\mathsf{noise}}$), we have an upper bound on the number of training points memorized by $\hat{w}^{\mathsf{minnorm}}$. Since the number of majority points is much larger than the number of minority points, this says that only a small fraction of majority points could be memorized by $\hat{w}^{\mathsf{minnorm}}$.

2. Next, we observe that since the core feature is noisy as per the parameter setting in Theorem 2, if we do not use the spurious feature, a constant fraction of majority points have to be memorized if spurious features are not used. Conversely, if less than this fraction of majority points can be memorized, the separator must use spurious features. Since using spurious features leads to higher worst-group test error, this reveals a trade-off between the worst-group test error of a separator and the fraction of *majority points* that it memorizes at training time. Succinctly, smaller fraction memorized implies the use of spurious features which in turn implies higher worst-group test error. Smaller worst-group test error requires eliminating the use of spurious features which would lead to a large fraction of majority points requiring memorization in order for a classifier to be a separator. We formalize the above trade-off between the worst-group test error and fraction of majority examples to be memorized in Proposition 3.

Combining the two steps together, since $\hat{w}^{\mathsf{minnorm}}$ memorizes only a small fraction of majority points by virtue of being the minimum norm separator, $\hat{w}^{\mathsf{minnorm}}$ suffers high worst-group test error.

We now formally prove Theorem 2, invoking propositions that we prove in subsequent sections.

B.2.1. Bounding the fraction of memorized examples in the majority groups.

In the first part of the proof, we show that the minimum-norm separator $\hat{w}^{\mathsf{minnorm}}$ "memorizes" a small fraction of the majority examples. Formally, we study the quantity $\delta_{\text{maj-train}}\left(\hat{w}, \gamma^2\right)$ defined as follows.

Definition 3. *Consider a separator $\hat{w}$ on training data $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$. Let $\delta_{\textit{maj-train}}\left(\hat{w}, \gamma^2\right)$ be the fraction of training examples that $\hat{w}$ γ-memorizes in the majority groups:*

$$\delta_{\textit{maj-train}}\left(\hat{w}, \gamma^2\right) \stackrel{\text{def}}{=} \frac{1}{n_{\mathsf{maj}}} \sum_{i \in G_{\mathsf{maj}}} \mathbb{I}\left[\left|\alpha^{(i)}(\hat{w})\right| > \frac{\gamma^2}{\sigma^2_{\mathsf{noise}}}\right] \tag{21}$$

We provide an upper bound on $\delta_{\text{maj-train}}\left(\hat{w}^{\mathsf{minnorm}}, \gamma^2\right)$ (Lemma 4) by first bounding $\|\hat{w}^{\mathsf{minnorm}}\|$ and then bounding $\delta_{\text{maj-train}}\left(\hat{w}^{\mathsf{minnorm}}, \gamma^2\right)$ in terms of $\|\hat{w}^{\mathsf{minnorm}}\|$.

Bounding $\|\hat{w}^{\mathsf{minnorm}}\|$

Lemma 1. *There exists a separator $w^{\mathsf{use-spu}}$ that satisfies $y^{(i)}(w^{\mathsf{use-spu}} \cdot x^{(i)}) \geq 1,\ \forall i \in G_{\mathsf{maj}}, G_{\mathsf{min}}$. The norm of this separator gives a bound on $\|\hat{w}^{\mathsf{minnorm}}\|$ as follows. For the parameter settings under Theorem 2, with high probability, we have*

$$\|\hat{w}^{\mathsf{minnorm}}\|_2^2 \leq \|w^{\mathsf{use-spu}}\|_2^2 \leq u^2 + s^2\sigma^2_{\mathsf{noise}}(1+c_1)n_{\mathsf{min}} + \frac{s^2\sigma^2_{\mathsf{noise}}}{n^4}, \tag{22}$$

for constants $u = 1.3125, s = \frac{2.61}{\sigma^2_{\mathsf{noise}}}$.

Proof. In order to get an upper bound on $\|\hat{w}^{\mathsf{minnorm}}\|$, we compute the norm of a particular separator. Concretely, we consider a separator $w^{\mathsf{use-spu}}$ of the following form:

$$\begin{aligned}
w^{\mathsf{use-spu}}_{\mathsf{core}} &= 0 \\
w^{\mathsf{use-spu}}_{\mathsf{spu}} &= u \\
w^{\mathsf{use-spu}}_{\mathsf{noise}} &= \sum_i \alpha^{(i)}(w^{\mathsf{use-spu}}) x^{(i)}_{\mathsf{noise}} \\
\alpha^{(i)}(w^{\mathsf{use-spu}}) &= 0 \text{ for } i \in G_{\mathsf{maj}} \\
\alpha^{(i)}(w^{\mathsf{use-spu}}) &= y^{(i)} s \text{ for } i \in G_{\mathsf{min}}
\end{aligned}$$

First, because we are interested in $w^{\mathsf{use-spu}}$ that does not use the core feature and relies on the spurious feature instead, we let $w^{\mathsf{use-spu}}_{\mathsf{core}} = 0$ and $w^{\mathsf{use-spu}}_{\mathsf{spu}} = u, u \in \mathbb{R}$. We set the value u appropriately so that none of the majority points are memorized (corresponding to $\alpha^{(i)}(w^{\mathsf{use-spu}}) = 0$ for all $i \in G_{\mathsf{maj}}$). However since the spurious correlations are reversed in the minority

points and $w_{\text{core}}^{\text{use-spu}} = 0$, the minority points have to be memorized. For simplicity, we set $\alpha^{(i)}(w^{\text{use-spu}}) = y^{(i)}s$ for all $i \in G_{\text{min}}$.

Now it remains to select appropriate values of constants u and s such that $y^{(i)}(w^{\text{use-spu}} \cdot x^{(i)}) \geq 1$ is satisfied for all training examples.

For majority points, this involves setting u large enough such that the less noisy spurious feature can be used to obtain the required margin. Without loss of generality, assume $y^{(i)} = 1$. Formally, for $i \in G_{\text{maj}}$,

$$\begin{aligned} w^{\text{use-spu}} \cdot x^{(i)} &\geq x_{\text{spu}}^{(i)} u + \sum_{j \in G_{\text{min}}} s x_{\text{noise}}^{(i)} \cdot x_{\text{noise}}^{(j)} \\ &\geq 4/5u + \sum_{j \in G_{\text{min}}} s x_{\text{noise}}^{(i)} \cdot x_{\text{noise}}^{(j)}, \text{ w.h.p. from Lemma 5 with } a = y = 1 \\ &\geq 4/5u - \frac{s\sigma_{\text{noise}}^2}{n^5}, \text{ w.h.p. from Lemma 8.} \\ &\geq 4/5u - \frac{s\sigma_{\text{noise}}^2}{100}. \end{aligned}$$

The first inequality follows from the fact that σ_{spu} is small enough under the parameter settings of Theorem 2 to allow a uniform bound on $x_{\text{spu}}^{(i)}$ (Lemma 5). The second inequality follows from setting the number of random features N to be large enough so that the noise features are near orthogonal (Lemma 8). Conversely, we have

$$4/5u - \frac{s\sigma_{\text{noise}}^2}{100} \geq 1 \implies w^{\text{use-spu}} \text{ is a separator on the majority points w.h.p.} \tag{23}$$

Notice that the condition in Equation 23 requires that u be greater than 0. Since the minority points have spurious attribute $a = -y$, we need to set s to be large enough so that $w^{\text{use-spu}}$ as defined above separates the minority points. Just as before, we set $y = 1$ WLOG. For $i \in G_{\text{min}}$, we have

$$\begin{aligned} w^{\text{use-spu}} \cdot x^{(i)} &\geq x_{\text{spu}}^{(i)} u + \sum_{j \in G_{\text{min}}} s x_{\text{noise}}^{(i)} \cdot x_{\text{noise}}^{(j)} \\ &\geq -6/5u + \sum_{j \in G_{\text{min}}} s x_{\text{noise}}^{(i)} \cdot x_{\text{noise}}^{(j)}, \text{ From Lemma 5 with } a = -y = -1 \\ &\geq -6/5u + s(1 - c_1)\sigma_{\text{noise}}^2 - \frac{s\sigma_{\text{noise}}^2}{n^5}, \text{ w.h.p from Lemma 8 and Lemma 9} \\ &\geq -6/5u + s(1 - c_1)\sigma_{\text{noise}}^2 - \frac{s\sigma_{\text{noise}}^2}{100}. \end{aligned}$$

The steps are similar to the condition for majority points, with the key difference that the contribution from the noise term involves $s\|x_{\text{noise}}^{(i)}\|_2^2$ (Lemma 9).

Conversely, we have

$$-6/5u + s(1 - c_1)\sigma_{\text{noise}}^2 - \frac{s\sigma_{\text{noise}}^2}{100} \geq 1 \implies w^{\text{use-spu}} \text{ is a separator on the minority points w.h.p..} \tag{24}$$

A set of parameters that satisfies both conditions above Equation 24 and Equation 23 is the following:

$$u = 1.3125, s\sigma_{\text{noise}}^2 = 2.61.$$

We use the fact that $c_1 < 1/2000$ (From Lemma 9).

Finally, we have w.h.p,

$$\|w^{\text{use-spu}}\|_2^2 \leq u^2 + s^2\sigma_{\text{noise}}^2(1 + c_1)n_{\text{min}} + \frac{s^2\sigma_{\text{noise}}^2}{n^4}. \tag{25}$$

This follows from bounds on $\|x_{\text{noise}}^{(i)}\|_2^2$ (Lemma 9) and sum of less than n^2 terms involving $s^2 x_{\text{noise}}^{(i)} \cdot x_{\text{noise}}^{(j)}$ (using Lemma 8). □

Bounding $\delta_{\text{maj-train}}\left(\hat{w}, \gamma^2\right)$ in terms of $\|\hat{w}\|$

Lemma 2. *For a separator $\hat{w}$ with bounded $\alpha^{(i)}(\hat{w})^2 \leq \frac{10n}{\sigma_{\text{noise}}^2}$ for all $i = 1, \ldots, n$, its norm can be bounded with high probability as*

$$\|\hat{w}\|_2^2 \geq \frac{\gamma^4(1-c_1)}{\sigma_{\text{noise}}^2}\delta_{\textit{maj-train}}\left(\hat{w}, \gamma^2\right) n_{\text{maj}} - \frac{10}{\sigma_{\text{noise}}^2 n^3} \tag{26}$$

under the parameter settings of Theorem 2.

Proof. The result follows bounded norms (Lemma 9), bounded dot products (Lemma 8), and the definition of $\delta_{\text{maj-train}}\left(\hat{w}, \gamma^2\right)$ (Definition 3).

$$\|\hat{w}\|_2^2 \geq \sum_{i \in G_{\text{maj}}} \alpha^{(i)}(\hat{w})^2 \|x_{\text{noise}}^{(i)}\|_2^2 + \sum_{j \neq k} \alpha^{(j)}(\hat{w})\alpha^{(k)}(\hat{w}) x_{\text{noise}}^{(j)} \cdot x_{\text{noise}}^{(k)} \tag{27}$$

$$\geq \underbrace{\left(\frac{\gamma^4(1-c_1)}{\sigma_{\text{noise}}^2}\right)\delta_{\text{maj-train}}\left(\hat{w}, \gamma^2\right) n_{\text{maj}}}_{\text{Choosing only points with } \alpha^{(i)}(\hat{w}) \geq \gamma^2/\sigma_{\text{noise}}^2} - \underbrace{\frac{M^2}{\sigma_{\text{noise}}^2 n^4}}_{\max \alpha^{(i)}(\hat{w}) = M/\sigma_{\text{noise}}^2}, \text{ w.h.p.} \tag{28}$$

$$\geq \frac{\gamma^4(1-c_1)}{\sigma_{\text{noise}}^2}\delta_{\text{maj-train}}\left(\hat{w}, \gamma^2\right) n_{\text{maj}} - \frac{10}{\sigma_{\text{noise}}^2 n^3} \tag{29}$$

□

Bounding $\delta_{\text{maj-train}}\left(\hat{w}^{\text{minnorm}}, \gamma^2\right)$

We now apply Lemma 1 and Lemma 2 in order to bound $\delta_{\text{maj-train}}\left(\hat{w}^{\text{minnorm}}, \gamma^2\right)$, showing that the fraction of majority points that are memorized is small for appropriate choice of γ.

To invoke Lemma 2, we first show that the coefficient $\alpha^{(i)}(\hat{w}^{\text{minnorm}})$ is bounded above with high probabiltity.

Lemma 3. *Under the parameter settings of Theorem 2, with high probability, $\alpha^{(i)}(\hat{w}^{\text{minnorm}})$ is bounded above for $i = 1, \ldots, n$ as*

$$\alpha^{(i)}(\hat{w}^{\text{minnorm}})^2 \leq \frac{10n}{\sigma_{\text{noise}}^4}. \tag{30}$$

Proof. Let $\max_i \alpha^{(i)}(\hat{w}^{\text{minnorm}}) = \frac{M}{\sigma_{\text{noise}}^2}$.

$$\|\hat{w}^{\text{minnorm}}\|_2^2 \geq \|\hat{w}_{\text{noise}}^{\text{minnorm}}\|_2^2 \tag{31}$$

$$= \sum_{i \in G_{\text{min}} G_{\text{maj}}} \alpha^{(i)}(\hat{w}^{\text{minnorm}})^2 \|x_{\text{noise}}^{(i)}\|_2^2 + \sum_{i,j} \alpha^{(i)}(\hat{w}^{\text{minnorm}})\alpha^{(j)}(\hat{w}^{\text{minnorm}}) x_{\text{noise}}^{(i)} \cdot x_{\text{noise}}^{(j)} \tag{32}$$

$$\geq \frac{M^2(1-c_1)}{\sigma_{\text{noise}}^2} - \frac{M^2}{\sigma_{\text{noise}}^2 n^6} n^2 \tag{33}$$

$$\geq \frac{M^2(1-c_1)}{\sigma_{\text{noise}}^2} - \frac{M^2}{\sigma_{\text{noise}}^2 n^4}. \tag{34}$$

From the upper bound on $\|\hat{w}^{\mathsf{minnorm}}\|_2^2$ (Lemma 1), we have

$$\frac{M^2(1-c_1)}{\sigma_{\mathsf{noise}}^2} - \frac{M^2}{\sigma_{\mathsf{noise}}^2 n^4} \leq u^2 + s^2\sigma_{\mathsf{noise}}^2(1+c_1)n_{\mathsf{min}} + \frac{s^2\sigma_{\mathsf{noise}}^2}{n^4} \tag{35}$$

$$\implies M^2\left(1 - c_1 - \frac{1}{n^4}\right) \leq u^2\sigma_{\mathsf{noise}}^2 + (s\sigma_{\mathsf{noise}}^2)^2\left((1+c_1)n_{\mathsf{min}} + \frac{1}{n^4}\right) \tag{36}$$

$$\implies M^2\left(1 - c_1 - \frac{1}{n^4}\right) \leq u^2\frac{n_{\mathsf{maj}}}{360000} + (s\sigma_{\mathsf{noise}}^2)^2\left((1+c_1)n_{\mathsf{min}} + \frac{1}{n^4}\right), \tag{37}$$

From a bound on $\sigma_{\mathsf{noise}}^2$ in the parameter settings. (38)

Since $c_1 < 1/2000$, and $n \geq 2000$, setting $u = 1.3125, s\sigma_{\mathsf{noise}}^2 = 2.61$, we get $M^2 \leq 10n$. □

Now, we are ready to show that $\delta_{\text{maj-train}}\left(\hat{w}^{\mathsf{minnorm}}, \gamma^2\right)$ is small.

Lemma 4. *Under the parameter settings of Theorem 2, the following is true with high probability.*

$$\delta_{\textit{maj-train}}\left(\hat{w}^{\mathsf{minnorm}}, \frac{9}{10}\right) \leq 1/200, \tag{39}$$

Proof. Applying Lemma 2 to $\hat{w}^{\mathsf{minnorm}}$ by invoking the bounds on $\alpha^{(i)}(\hat{w}^{\mathsf{minnorm}})$ (Lemma 3),

$$\|\hat{w}^{\mathsf{minnorm}}\|_2^2 \geq \frac{\gamma^4(1-c_1)}{\sigma_{\mathsf{noise}}^2}\delta_{\text{maj-train}}\left(\hat{w}^{\mathsf{minnorm}}, \gamma^2\right) n_{\mathsf{maj}} - \frac{10}{\sigma_{\mathsf{noise}}^2 n^3} \tag{40}$$

with high probability. Putting this together with Lemma 1, we have

$$\frac{\gamma^4(1-c_1)}{\sigma_{\mathsf{noise}}^2}\delta_{\text{maj-train}}\left(\hat{w}^{\mathsf{minnorm}}, \gamma^2\right) n_{\mathsf{maj}} - \frac{10}{\sigma_{\mathsf{noise}}^2 n^3} \leq u^2 + s^2\sigma_{\mathsf{noise}}^2(1+c_1)n_{\mathsf{min}} + \frac{s^2\sigma_{\mathsf{noise}}^2}{n^4}$$

$$\implies \delta_{\text{maj-train}}\left(\hat{w}^{\mathsf{minnorm}}, \gamma^2\right) \leq \underbrace{\frac{u^2\sigma_{\mathsf{noise}}^2}{\gamma^4 n_{\mathsf{maj}}(1-c_1)}}_{\text{Very small}} + \underbrace{\left(\frac{(s\sigma_{\mathsf{noise}}^2)^2(1+c_1)}{\gamma^4(1-c_1)}\right)\frac{n_{\mathsf{min}}}{n_{\mathsf{maj}}}}_{\approx 0.0042} + \underbrace{\frac{(s\sigma_{\mathsf{noise}}^2)^2}{n^4 n_{\mathsf{maj}}}}_{\text{Very small}} + \underbrace{\frac{10}{\gamma^4(1-c_1)n^3}}_{\text{Very small}}$$

$$\implies \delta_{\text{maj-train}}\left(\hat{w}^{\mathsf{minnorm}}, \frac{9}{10}\right) \leq 1/200, \text{w.h.p},$$

where in the last step we substitute the constants $\gamma^2 = 9/10, u = 1.3125, s\sigma_{\mathsf{noise}}^2 = 2.61, n_{\mathsf{maj}}/n_{\mathsf{min}} \leq 1/2000$ and $\sigma_{\mathsf{noise}}^2 \leq n_{\mathsf{maj}}/360000$. □

B.2.2. Concentration Inequalities

Lemma 5. *With probability* $> 1 - 1/100$, *if* $\sigma_{\mathsf{spu}} \leq \frac{1}{4\sqrt{\log 100n}}$,

$$a - 1/5 \leq x_{\mathsf{spu}}^{(i)} \leq a + 1/5, \ \forall i = 1, \ldots n, \tag{41}$$

where a *is the spurious attribute.*

This follows from standard subgaussian concentration and union bound over $n = n_{\mathsf{maj}} + n_{\mathsf{min}}$ points.

Lemma 6. *For a vector* $z \in \mathbb{R}^N$ *such that* $z \in \mathcal{N}(0, \sigma^2 I)$,

$$\mathbb{P}(|\|z\|^2 - \sigma^2 N| \geq \sigma^2 t) \leq 2\exp\left(\frac{-Nt^2}{8}\right). \tag{42}$$

Lemma 7. *For two vectors* $z_i, z_j \in \mathbb{R}^N$ *such that* $z_i, z_j \sim \mathcal{N}(0, \sigma^2 I)$, *by Hoeffding's inequality, we have*

$$\mathbb{P}(|z_i \cdot z_j| \geq \sigma^2 t) \leq 2\exp\left(-\frac{t^2}{2\|z_i\|^2}\right). \tag{43}$$

Corollary 1. *Combining Lemma 6 and Lemma 7, we get*

$$\mathbb{P}(|z_i \cdot z_j| \geq \sigma^2 t) \leq 2\exp\Big(\frac{-N^3}{8}\Big) + 2\exp\Big(-\frac{t^2}{8N}\Big). \tag{44}$$

Lemma 8. *For $N = \Omega(poly(n))$, with probability greater than $1 - 1/2000$,*

$$|x^{(i)}_{\mathsf{noise}} \cdot x^{(j)}_{\mathsf{noise}}| \leq \frac{\sigma^2_{\mathsf{noise}}}{n^6} \;\; \forall x^{(i)}_{\mathsf{noise}}, x^{(j)}_{\mathsf{noise}}. \tag{45}$$

This follows from Corollary 1 and union bound over n^2 pairs of training points.

Lemma 9. *For $N = \Omega(poly(n))$, with probability greater than $1 - 1/2000$,*

$$(1 - c_1)\sigma^2 \leq \|x^{(i)}_{\mathsf{noise}}\|^2 \leq (1 + c_1)\sigma^2, \forall i. \tag{46}$$

This follows from Lemma 6 and union bound over n training points. In particular, we can set $c_1 < 1/2000$ for large enough N.

B.2.3. Small $\delta_{\text{maj-train}}\left(\hat{w}^{\mathsf{minnorm}}, \gamma^2\right)$ implies high worst-group error

In the previous section, we proved that $\delta_{\text{maj-train}}\left(\hat{w}^{\mathsf{minnorm}}, \gamma^2\right)$, the fraction of majority training samples that can have coefficient on the noise vectors greater than $\gamma^2/\sigma^2_{\mathsf{noise}}$ in the max margin separator $\hat{w}^{\mathsf{minnorm}}$ is bounded for suitable value of γ. We showed this using the fact that the norm of $\hat{w}^{\mathsf{minnorm}}$ is the smallest among all separators and the observation that the squared norm of a separator roughlty scales proportional the number of training points that have large coefficient along the noise vectors.

What does small $\delta_{\text{maj-train}}\left(\hat{w}^{\mathsf{minnorm}}, \gamma^2\right)$ imply? We now show that the bound on $\delta_{\text{maj-train}}\left(\hat{w}^{\mathsf{minnorm}}, \gamma^2\right)$ has an important consequence on the worst-group error $\text{Err}_{\mathsf{wg}}(\hat{w}^{\mathsf{minnorm}})$; low $\delta_{\text{maj-train}}\left(\hat{w}^{\mathsf{minnorm}}, \gamma\right)$ would imply high worst-group error $\text{Err}_{\mathsf{wg}}(\hat{w}^{\mathsf{minnorm}})$. We show that there is a trade-off between the worst-group test error of a separator and the fraction of *majority points* that it "memorizes" at training time. If a model that has low worst-group test error must use the core feature and not the spurious feature, and to obtain zero training error such a model would memorize a potentially large fraction of majority and minority points. In contrast, if the model instead uses only the spurious feature, then the worst-group test error would be high, but it would memorize only a small fraction of majority examples at training time; because we assume that the spurious feature is much less noisy than the core feature ($\sigma_{\mathsf{core}} \gg \sigma_{\mathsf{spu}}$), much fewer majority examples would need to be memorized. To summarize, *a large $\hat{w}_{\mathsf{spu}}$ would require smaller fraction of majority points to be memorized $\delta_{\textit{maj-train}}\left(\hat{w}, \gamma^2\right)$ but increase the worst-group test error $Err_{\mathsf{wg}}(\hat{w})$*. We formalize the above trade-off between the worst-group error and fraction of majority examples to be memorized in Proposition 3.

Proposition 3. *For the minimum norm separator $\hat{w}^{\mathsf{minnorm}}$, under the parameter settings of Theorem 2, with high probability,*

$$Err_{\mathsf{wg}}(\hat{w}^{\mathsf{minnorm}}) \geq \Phi\left(\frac{-c_3 + \hat{w}^{\mathsf{minnorm}}_{\mathsf{spu}} - \hat{w}^{\mathsf{minnorm}}_{\mathsf{core}}}{\sqrt{\hat{w}^{\mathsf{minnorm}\,2}_{\mathsf{core}}\sigma^2_{\mathsf{core}} + \hat{w}^{\mathsf{minnorm}\,2}_{\mathsf{spu}}\sigma^2_{\mathsf{spu}}}}\right) - c_4, \tag{47}$$

for some constants $c_3, c_4 < 1/1000$ and Φ the Gaussian CDF.

For any separator $\hat{w}$ that spans the training points and satisfies

$$\alpha^{(i)}(\hat{w})^2 \leq \frac{10n}{\sigma^4_{\mathsf{noise}}}, \tag{48}$$

under the parameter settings of Theorem 2, with high probability,

$$\delta_{\textit{maj-train}}\left(\hat{w}, \gamma^2\right) \geq \Phi\left(\frac{1 - (1 + c_1)\gamma^2 - c_5 - \hat{w}_{\mathsf{spu}} - \hat{w}_{\mathsf{core}}}{\sqrt{\hat{w}^2_{\mathsf{core}}\sigma^2_{\mathsf{core}} + \hat{w}^2_{\mathsf{spu}}\sigma^2_{\mathsf{spu}}}}\right) - c_6, \tag{49}$$

for some constants $c_1 < 1/2000; c_5, c_6 < 1/1000$ and Φ the Gaussian CDF.

We prove Proposition 3 in Section B.2.5.

As mentioned before, we see that the spurious component weight $\hat{w}_{\text{spu}}^{\text{minnorm}}$ has opposite effects on the two quantities; $\text{Err}_{\text{wg}}(\hat{w})$ increases with increase $\hat{w}_{\text{spu}}$, but $\delta_{\text{maj-train}}(\hat{w}, \gamma)$ decreases with increase in $\hat{w}_{\text{spu}}$. This dependence can be exploited to relate the two quantities to each other as follows.

$$\Phi^{-1}(\delta_{\text{maj-train}}\left(\hat{w}^{\text{minnorm}}, \gamma\right) + c_6) + \Phi^{-1}(\text{Err}_{\text{wg}}(\hat{w}^{\text{minnorm}}) + c_4) \geq \frac{1 - c_3 - c_5 - (1+c_1)\gamma^2 - 2\hat{w}_{\text{core}}^{\text{minnorm}}}{\sqrt{\hat{w}_{\text{core}}^2 \sigma_{\text{core}}^2 + \hat{w}_{\text{spu}}^2 \sigma_{\text{spu}}^2}}. \tag{50}$$

In other words, if the $\delta_{\text{maj-train}}\left(\hat{w}^{\text{minnorm}}, \gamma\right)$ is low, then $\text{Err}_{\text{wg}}(\hat{w}^{\text{minnorm}})$ would need to be high.

B.2.4. Worst-group error is high

Recall from part 1 that $\delta_{\text{maj-train}}\left(\hat{w}^{\text{minnorm}}, \gamma\right) < 1/200$ for appropriate choice of γ, and from part 2 the trade-off between $\delta_{\text{maj-train}}\left(\hat{w}^{\text{minnorm}}, \gamma\right)$ and $\text{Err}_{\text{wg}}(\hat{w}^{\text{minnorm}})$ (Equation (50)). As a final step, we need to bound the quantities on the RHS of Equation (50). All the constants are small, and $\gamma^2 = 9/10, \delta_{\text{maj-train}}\left(\hat{w}^{\text{minnorm}}, 9/10\right) \leq 1/200$ (Lemma 4) which allows us to write

$$\Phi^{-1}(0.006) + \Phi^{-1}(\text{Err}_{\text{wg}}(\hat{w}^{\text{minnorm}}) + c_4) \geq \frac{-2\hat{w}_{\text{core}}^{\text{minnorm}}}{\sqrt{\hat{w}_{\text{core}}^{\text{minnorm}\,2} \sigma_{\text{core}}^2 + \hat{w}_{\text{spu}}^{\text{minnorm}\,2} \sigma_{\text{spu}}^2}} \geq \frac{-2}{\sigma_{\text{core}}} \tag{51}$$

$$\implies \Phi^{-1}(\text{Err}_{\text{wg}}(\hat{w}^{\text{minnorm}}) + c_4) \geq 0.512 \tag{52}$$

$$\implies \text{Err}_{\text{wg}}(\hat{w}^{\text{minnorm}}) \geq 0.67 \tag{53}$$

We have hence proved that the minimum-norm separator $\hat{w}^{\text{minnorm}}$ incurs high worst-group error with high probability under the specified conditions.

B.2.5. Proof of Proposition 3

Proposition 3. *For the minimum norm separator $\hat{w}^{\text{minnorm}}$, under the parameter settings of Theorem 2, with high probability,*

$$\textit{Err}_{\text{wg}}(\hat{w}^{\text{minnorm}}) \geq \Phi\left(\frac{-c_3 + \hat{w}_{\text{spu}}^{\text{minnorm}} - \hat{w}_{\text{core}}^{\text{minnorm}}}{\sqrt{\hat{w}_{\text{core}}^{\text{minnorm}\,2} \sigma_{\text{core}}^2 + \hat{w}_{\text{spu}}^{\text{minnorm}\,2} \sigma_{\text{spu}}^2}}\right) - c_4, \tag{47}$$

for some constants $c_3, c_4 < 1/1000$ and Φ the Gaussian CDF.

For any separator $\hat{w}$ that spans the training points and satisfies

$$\alpha^{(i)}(\hat{w})^2 \leq \frac{10n}{\sigma_{\text{noise}}^4}, \tag{48}$$

under the parameter settings of Theorem 2, with high probability,

$$\delta_{\textit{maj-train}}\left(\hat{w}, \gamma^2\right) \geq \Phi\left(\frac{1 - (1+c_1)\gamma^2 - c_5 - \hat{w}_{\text{spu}} - \hat{w}_{\text{core}}}{\sqrt{\hat{w}_{\text{core}}^2 \sigma_{\text{core}}^2 + \hat{w}_{\text{spu}}^2 \sigma_{\text{spu}}^2}}\right) - c_6, \tag{49}$$

for some constants $c_1 < 1/2000; c_5, c_6 < 1/1000$ and Φ the Gaussian CDF.

Proof. We derive the two bounds below.

Worst-group test error

We bound the expected worst-group error $\text{Err}_{\text{wg}}(\hat{w}^{\text{minnorm}})$, which is the expected worst-group loss over the data distribution. Below, we lower bound the worst-group error $\text{Err}_{\text{wg}}(\hat{w}^{\text{minnorm}})$ by bounding the error on a particular group: minority positive

points which have label $y = 1$ and spurious attribute $a = -1$. The test error is the probability that a test example x from this group gets misclassified, i.e. $\hat{w}^{\text{minnorm}} \cdot x < 0$.

$$\begin{aligned}
\text{Err}_{\text{wg}}(\hat{w}^{\text{minnorm}}) &\geq \mathbb{P}\left(\hat{w}^{\text{minnorm}} \cdot x < 0 \mid y = 1, a = -1\right) && (54)\\
&= \mathbb{P}\left(\hat{w}_{\text{core}}^{\text{minnorm}} x_{\text{core}} + \hat{w}_{\text{spu}}^{\text{minnorm}} x_{\text{spu}} + \hat{w}_{\text{noise}}^{\text{minnorm}} \cdot x_{\text{noise}} < 0 \mid y = 1, a = -1\right) && (55)\\
&= \mathbb{P}\Big(\hat{w}_{\text{core}}^{\text{minnorm}}(1 + \sigma_{\text{core}} z_1) + \hat{w}_{\text{spu}}^{\text{minnorm}}(-1 + \sigma_{\text{spu}} z_2) + \hat{w}_{\text{noise}}^{\text{minnorm}} \cdot x_{\text{noise}} < 0\Big) && (56)
\end{aligned}$$

In the last step, we rewrite for convenience $x_{\text{core}} = y + \sigma_{\text{core}} z_1$ and $x_{\text{spu}} = a + \sigma_{\text{spu}} z_2$, where $z_1, z_2 \sim \mathcal{N}(0, 1)$.

We use the properties of high-dimensional Gaussian random vectors to bound the quantity $\hat{w}_{\text{noise}}^{\text{minnorm}} \cdot x_{\text{noise}}$. Recall that $\hat{w}_{\text{noise}}^{\text{minnorm}}$ can be written as

$$\hat{w}_{\text{noise}}^{\text{minnorm}} = \sum_{i \in G_{\text{maj}}, G_{\text{min}}} \alpha^{(i)}(\hat{w}^{\text{minnorm}}) x_{\text{noise}}^{(i)}. \tag{57}$$

From Lemma 3, we know that $\max_i \alpha^{(i)}(\hat{w}^{\text{minnorm}})^2 < \frac{10n}{\sigma_{\text{noise}}^4}$. This, along with Lemma 7 gives $|x_{\text{noise}} \cdot \hat{w}_{\text{noise}}^{\text{minnorm}}| \leq c_3$ with probability $1 - c_4$ for some small constants $c_3, c_4 < 1/1000$. Let B denote the event that this high probability event where the dot product $|x_{\text{noise}} \cdot \hat{w}_{\text{noise}}^{\text{minnorm}}| \leq c_3$. Using the fact that $\mathbb{P}(A) \geq \mathbb{P}(A \mid B) - \mathbb{P}(\neg B)$ which follows from simple algebra, we have

$$\begin{aligned}
\text{Err}_{\text{wg}}(\hat{w}^{\text{minnorm}}) &\geq \mathbb{P}\Big(\hat{w}_{\text{core}}^{\text{minnorm}}(1 + \sigma_{\text{core}} z_1) + \hat{w}_{\text{spu}}^{\text{minnorm}}(-1 + \sigma_{\text{spu}} z_2) + \hat{w}_{\text{noise}}^{\text{minnorm}} \cdot x_{\text{noise}} < 0\Big) && (58)\\
&\geq \mathbb{P}\Big(\hat{w}_{\text{core}}^{\text{minnorm}}(1 + \sigma_{\text{core}} z_1) + \hat{w}_{\text{spu}}^{\text{minnorm}}(1 - \sigma_{\text{spu}} z_2) < -c_3\Big) - c_4 && (59)\\
&= \mathbb{P}\Big(\hat{w}_{\text{core}}^{\text{minnorm}} \sigma_{\text{core}} z_1 + \hat{w}_{\text{spu}}^{\text{minnorm}} \sigma_{\text{spu}} z_2 < -c_3 + \hat{w}_{\text{spu}}^{\text{minnorm}} - \hat{w}_{\text{core}}^{\text{minnorm}}\Big) - c_4 && (60)\\
&= \Phi\left(\frac{-c_3 + \hat{w}_{\text{spu}}^{\text{minnorm}} - \hat{w}_{\text{core}}^{\text{minnorm}}}{\sqrt{\hat{w}_{\text{core}}^{\text{minnorm}\,2} \sigma_{\text{core}}^2 + \hat{w}_{\text{spu}}^{\text{minnorm}\,2} \sigma_{\text{spu}}^2}}\right) - c_4. && (61)
\end{aligned}$$

From the expression above, we see that $\text{Err}_{\text{wg}}(\hat{w}^{\text{minnorm}})$ increases as the spurious component $\hat{w}_{\text{spu}}^{\text{minnorm}}$ increases. This is because in the minority group, the spurious feature is negatively correlated with the label.

Fraction of memorized training examples in majority groups

We now compute a lower bound on $\delta_{\text{maj-train}}\left(\hat{w}^{\text{minnorm}}, \gamma^2\right)$, which is the number of majority points (where $a = y$) that are "memorized." Intuitively, we want to show that the fraction depends on $\hat{w}_{\text{spu}} - \hat{w}_{\text{core}}$. The more the core feature is used relative to the spurious feature, the larger fraction of points need to be memorized because the core feature is more noisy.

First, consider a separator $\hat{w}$ with some core and spurious components $\hat{w}_{\text{core}}$ and $\hat{w}_{\text{spu}}$. Recall that $\hat{w}_{\text{noise}} = \sum_i \alpha^{(i)}(\hat{w}) x_{\text{noise}}^{(i)}$ and $y^{(i)}(\hat{w} \cdot x^{(i)}) \geq 1$ by the definition of separators. For a given $\hat{w}_{\text{core}}$ and $\hat{w}_{\text{spu}}$, we want to bound the fraction of majority points ($a = y$) which can have $\alpha^{(i)}(\hat{w}) < \frac{\gamma^2}{\sigma_{\text{noise}}^2}$. We focus only on separators with bounded memorization, i.e. those that satisfy $\alpha^{(i)}(\hat{w})^2 \leq \frac{10n}{\sigma_{\text{noise}}^4}$. Note that from Lemma 3, w.h.p., the mininum-norm separator $\hat{w}^{\text{minnorm}}$ satifies this condition.

We bound the above by bounding a related quantity: the fraction of points that are memorized in the training distribution in expectation. We then use concentration to relate it to the fraction of the training set.

Formally, we have fixed quantities $\hat{w}_{\text{core}}$ and $\hat{w}_{\text{spu}}$. The training set is generated as per the usual data generating distribution. As before, we are interested in separators on the training set. For any majority training point, the coefficient $\alpha^{(i)}(\hat{w})$ in a separator is a random variable. Since training point i is separated, we have

$$\hat{w}_{\text{core}}(1 + \sigma_{\text{core}} z_1) + \hat{w}_{\text{spu}}(1 + \sigma_{\text{spu}} z_2) + \Big(\sum_i \alpha^{(i)}(\hat{w}) x_{\text{noise}}^{(i)}\Big)^\top x_{\text{noise}}^{(i)} \geq 1.$$

From Lemma 8, Lemma 6, and the condition on $\alpha^{(i)}(\hat{w})$, this implies with high probability that

$$\hat{w}_{\text{core}}(1 + \sigma_{\text{core}} z_1) + \hat{w}_{\text{spu}}(1 + \sigma_{\text{spu}} z_2) \geq 1 - (1 + c_1)\sigma_{\text{noise}}^2 \alpha^{(i)}(\hat{w}) - c_5,$$

for some constant $c_5 < 1/1000$. Conditioning on the high probability event just as before ($\mathbb{P}(A) \leq \mathbb{P}(A \mid B) + \mathbb{P}(\neg B)$), we get

$$\mathbb{P}(\alpha^{(i)}(\hat{w}) \leq \frac{\gamma^2}{\sigma^2_{\text{noise}}}) \leq \mathbb{P}\Big(\hat{w}_{\text{core}}\sigma_{\text{core}}z_1 + \hat{w}_{\text{spu}}\sigma_{\text{spu}}z_2 \leq -1 + (1+c_1)\gamma^2 + c_5 + \hat{w}_{\text{core}} + \hat{w}_{\text{spu}}\Big) + \delta \tag{62}$$

$$= \Phi\left(\frac{-1 + (1+c_1)\gamma^2 + c_5 + \hat{w}_{\text{spu}} + \hat{w}_{\text{core}}}{\sqrt{\hat{w}^2_{\text{core}}\sigma^2_{\text{core}} + \hat{w}^2_{\text{spu}}\sigma^2_{\text{spu}}}}\right) + \delta \tag{63}$$

$$\implies \mathbb{P}(\alpha^{(i)}(\hat{w}) \geq \frac{\gamma^2}{\sigma^2_{\text{noise}}}) \geq \Phi\left(\frac{1 - (1+c_1)\gamma^2 - c_5 - \hat{w}_{\text{spu}} - \hat{w}_{\text{core}}}{\sqrt{\hat{w}^2_{\text{core}}\sigma^2_{\text{core}} + \hat{w}^2_{\text{spu}}\sigma^2_{\text{spu}}}}\right) - \delta, \tag{64}$$

for some $\delta < 1/2000$. Finally, we connect to $\delta_{\text{maj-train}}\,(\hat{w})\,(\gamma^2)$ which is the finite sample version of the quantity $\mathbb{P}(\alpha^{(i)}(\hat{w}) \leq \frac{\gamma^2}{\sigma^2_{\text{noise}}})$. By DKW, we know that the empirical CDF converges to the population CDF. Under the conditions of Theorem 2, which lower bounds the number of majority elements, we have with high probability,

$$\delta_{\text{maj-train}}\,(\hat{w})\,(\gamma^2) \geq \Phi\left(\frac{1 - (1+c_1)\gamma^2 - c_5 - \hat{w}_{\text{spu}} - \hat{w}_{\text{core}}}{\sqrt{\hat{w}^2_{\text{core}}\sigma^2_{\text{core}} + \hat{w}^2_{\text{spu}}\sigma^2_{\text{spu}}}}\right) - c_6, \tag{65}$$

for constants $c_5, c_6 < 1/1000$.

□

□

B.2.6. Proof of Proposition 1

Proposition 1 (Norm of models using the spurious feature)**.** *When $\sigma^2_{\text{core}}, \sigma^2_{\text{spu}}$ satisfy the conditions in Theorem 1, there exists N_0 such that for all $N > N_0$, with high probability, there exists a separator $w^{\text{use-spu}} \in \mathcal{W}^{\text{use-spu}}$ such that*

$$\|w^{\text{use-spu}}\|_2^2 \leq \gamma_1^2 + \left(\frac{\gamma_2 n_{\text{min}}}{\sigma^2_{\text{noise}}}\right),$$

for some constants $\gamma_1, \gamma_2 > 0$.

Proof. The proposition follows directly from Lemma 1.

$$\begin{aligned}\|w^{\text{use-spu}}\|_2^2 &\leq u^2 + s^2\sigma^2_{\text{noise}}(1+c_1)n_{\text{min}} + \frac{s^2\sigma^2_{\text{noise}}}{n^4} \\ &\leq u^2 + s^2\sigma^2_{\text{noise}}(2+c_1)n_{\text{min}}.\end{aligned}$$

The constant $\gamma_1 = u = 1.3125$ and $\gamma_2 = s\sigma^2_{\text{noise}}(2+c_1) = 2.61(2+c_1)$ for $c_1 < 1/2000$. □

B.2.7. Proof of Proposition 2

Proposition 2 (Norm of models using the core feature)**.** *When $\sigma^2_{\text{core}}, \sigma^2_{\text{spu}}$ satisfy the conditions in Theorem 1 and $n_{\text{min}} \geq 100$, there exists N_0 such that for all $N > N_0$, with high probability, all separators $w^{\text{use-core}} \in \mathcal{W}^{\text{use-core}}$ satisfy*

$$\|w^{\text{use-core}}\|_2^2 \geq \frac{\gamma_3 n}{\sigma^2_{\text{noise}}},$$

for some constant $\gamma_3 > 0$.

Proof. To bound the norm for all $w^{\mathsf{use-core}} \in \mathcal{W}^{\mathsf{use-core}}$, we provide a lower bound on the norm of the minimum-norm separator in the set $\mathcal{W}^{\mathsf{use-core}}$:

$$\bar{w}^{\mathsf{use-core}} \stackrel{\text{def}}{=} \arg \min_{w \in \mathcal{W}^{\mathsf{use-core}}} \|w\|^2. \tag{66}$$

We bound the $\|\bar{w}^{\mathsf{use-core}}\|$ in two steps:

1. We first provide a lower bound for $\|\bar{w}^{\mathsf{use-core}}\|$ in terms of the fraction of training points memorized $\delta_{\text{train}}\left(\bar{w}^{\mathsf{use-core}}, \gamma^2\right)$ (defined formally below) in Corollary 2.
2. We then provide a lower bound for $\delta_{\text{train}}\left(\bar{w}^{\mathsf{use-core}}, \gamma^2\right)$ in Corollary 3.

We first formally define $\delta_{\text{train}}\left(\hat{w}, \gamma^2\right)$.

Definition 4. *For a separator $\hat{w}$ on training data $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$, let $\delta_{train}\left(\hat{w}, \gamma^2\right)$ be the fraction of training examples that $\hat{w}$ γ-memorizes:*

$$\delta_{train}\left(\hat{w}, \gamma^2\right) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left[\left|\alpha^{(i)}(\hat{w})\right| > \frac{\gamma^2}{\sigma_{\mathsf{noise}}^2}\right] \tag{67}$$

Bounding $\|\bar{w}^{\mathsf{use-core}}\|$ by $\delta_{\text{train}}\left(\bar{w}^{\mathsf{use-core}}, \gamma^2\right)$

Lemma 10. *For a separator $\hat{w}$ with bounded $\alpha^{(i)}(\hat{w})^2 \leq \frac{10n}{\sigma_{\mathsf{noise}}^2}$ for all $i = 1, \ldots, n$, its norm can be bounded with high probability as*

$$\|\hat{w}\|_2^2 \geq \frac{\gamma^4(1-c_1)}{\sigma_{\mathsf{noise}}^2} \delta_{train}\left(\hat{w}, \gamma^2\right) n - \frac{10}{\sigma_{\mathsf{noise}}^2 n^3} \tag{68}$$

Proof. Similarly to the proof of Lemma 2, the result follows bounded norms (Lemma 9), bounded dot products (Lemma 8), and the definition of $\delta_{\text{train}}\left(\hat{w}, \gamma^2\right)$ (Definition 4).

$$\|\hat{w}\|_2^2 \geq \sum_{i \in G_{\mathsf{maj}}} \alpha^{(i)}(\hat{w})^2 \|x_{\mathsf{noise}}^{(i)}\|_2^2 + \sum_{j \neq k} \alpha^{(j)}(\hat{w}) \alpha^{(k)}(\hat{w}) x_{\mathsf{noise}}^{(j)} \cdot x_{\mathsf{noise}}^{(k)} \tag{69}$$

$$\geq \underbrace{\left(\frac{\gamma^4(1-c_1)}{\sigma_{\mathsf{noise}}^2}\right) \delta_{\text{train}}\left(\hat{w}, \gamma^2\right) n}_{\text{Choosing only points with } \alpha^{(i)}(\hat{w}) \geq \gamma^2/\sigma_{\mathsf{noise}}^2} - \underbrace{\frac{M^2}{\sigma_{\mathsf{noise}}^2 n^4}}_{\max \alpha^{(i)}(\hat{w}) = M/\sigma_{\mathsf{noise}}^2}, \text{ w.h.p.} \tag{70}$$

$$\geq \frac{\gamma^4(1-c_1)}{\sigma_{\mathsf{noise}}^2} \delta_{\text{train}}\left(\hat{w}, \gamma^2\right) n - \frac{10}{\sigma_{\mathsf{noise}}^2 n^3} \tag{71}$$

□

Corollary 2. *With high probability,*

$$\|\bar{w}^{\mathsf{use-core}}\|_2^2 \geq \frac{\gamma^4(1-c_1)}{\sigma_{\mathsf{noise}}^2} \delta_{maj\text{-}train}\left(\bar{w}^{\mathsf{use-core}}, \gamma^2\right) n_{\mathsf{maj}} - \frac{10}{\sigma_{\mathsf{noise}}^2 n^3} \tag{72}$$

Proof. The result follows from applying Lemma 10 to $\bar{w}^{\mathsf{use-core}}$, invoking the bounds on any individual component $\alpha^{(i)}(\bar{w}^{\mathsf{use-core}})$ obtained below in Lemma 11. □

Below, we bound $\alpha^{(i)}(\bar{w}^{\mathsf{use-core}})$, where $\alpha^{(i)}(\bar{w}^{\mathsf{use-core}})$ is the component of training point i to the classifier $\bar{w}^{\mathsf{use-core}}$ via the representer theorem.

Lemma 11. *With high probability, $i = 1, \ldots, n$, $\alpha^{(i)}(\bar{w}^{\mathsf{use-core}})$ can be bounded as follows.*

$$\alpha^{(i)}(\bar{w}^{\mathsf{use-core}})^2 \leq \frac{10n}{\sigma_{\mathsf{noise}}^4}. \tag{73}$$

Proof. As a first step, we upper bound the norm of $\bar{w}^{\mathsf{use-core}}$ by the norm of another separator $w^{\mathsf{use-core}} \in \mathcal{W}^{\mathsf{use-core}}$, using the fact that $\bar{w}^{\mathsf{use-core}}$ is the minimum-norm separator in $\mathcal{W}^{\mathsf{use-core}}$. In particular, we construct a separator $w^{\mathsf{use-core}} \in \mathcal{W}^{\mathsf{use-core}}$ that "memorizes" all training points, of the following form:

$$\begin{aligned} w_{\mathsf{core}}^{\mathsf{use-core}} &= 0 \\ w_{\mathsf{spu}}^{\mathsf{use-core}} &= 0 \\ \alpha^{(i)}(w^{\mathsf{use-core}}) &= y^{(i)}\alpha \text{ for all } i = 1, \ldots, n. \end{aligned}$$

This is analogous to the construction of $w^{\mathsf{use-spu}} \in \mathcal{W}^{\mathsf{use-spu}}$ (Lemma 1), and similar calculations can be used to obtain a suitable value α to ensure that $w^{\mathsf{use-core}}$ is a separator with high probability. We provide it below for completeness. We show that the following condition is sufficient to satisfy the margin constraints $y^{(i)}w^{\mathsf{use-core}} \cdot x^{(i)} \geq 1$ for all $i = 1, \ldots, n$ with high probability:

$$\alpha\sigma_{\mathsf{noise}}^2 \geq \frac{1}{1 - c_1 - 1/n^5}. \tag{74}$$

for $c_1 < 1/2000$. We obtain the above condition by applying Lemma 8 and Lemma 9 to the margin condition.

$$w^{\mathsf{use-core}} \cdot x^{(i)} \geq 1 \tag{75}$$

$$\implies \alpha\|x_{\mathsf{noise}}^{(i)}\|^2 - \alpha\sum_{j\neq i}\left|x_{\mathsf{noise}}^{(i)} \cdot x_{\mathsf{noise}}^{(j)}\right| \geq 1 \tag{76}$$

$$\implies \alpha\sigma_{\mathsf{noise}}^2(1 - c_1) - \frac{\alpha\sigma_{\mathsf{noise}}^2}{n^5} \geq 1 \text{ with high probability} \tag{77}$$

Thus, we can construct $w^{\mathsf{use-core}}$ by setting some constant $\alpha\sigma_{\mathsf{noise}}^2 \leq 2$.

Now that we have constructed $w^{\mathsf{use-core}}$, we can bound the norm of the minimum norm separator $\bar{w}^{\mathsf{use-core}}$ by the norm of $w^{\mathsf{use-core}}$. The following is true with high probability,

$$\|\bar{w}^{\mathsf{use-core}}\|^2 \leq \|w_{\mathsf{noise}}^{\mathsf{use-core}}\|^2 \tag{78}$$

$$= \sum_{i=1}^{n}\alpha^2\|x_{\mathsf{noise}}^{(i)}\|^2 + \sum_{i\neq j}\alpha^2 x_{\mathsf{noise}}^{(i)} \cdot x_{\mathsf{noise}}^{(j)} \tag{79}$$

$$\leq \alpha^2\sigma_{\mathsf{noise}}^2(1 + c_1)n + \frac{\alpha^2\sigma_{\mathsf{noise}}^2}{n^4} \tag{80}$$

Finally, we bound $\alpha^{(i)}(\bar{w}^{\mathsf{use-core}})$ for all i by bounding $\max_i \alpha^{(i)}(\bar{w}^{\mathsf{use-core}}) = \frac{M}{\sigma_{\mathsf{noise}}^2}$. As we showed in the proof of Lemma 3, following is true with high probability:

$$\|\bar{w}^{\mathsf{use-core}}\|_2^2 \geq \frac{M^2(1 - c_1)}{\sigma_{\mathsf{noise}}^2} - \frac{M^2}{\sigma_{\mathsf{noise}}^2 n^4}. \tag{81}$$

Combined with the upper bound on $\|\bar{w}^{\mathsf{use-core}}\|_2^2$ (Equation (80)), we have

$$\frac{M^2(1 - c_1)}{\sigma_{\mathsf{noise}}^2} - \frac{M^2}{\sigma_{\mathsf{noise}}^2 n^4} \leq \|\bar{w}^{\mathsf{use-core}}\| \leq \alpha^2\sigma_{\mathsf{noise}}^2(1 + c_1)n + \frac{\alpha^2\sigma_{\mathsf{noise}}^2}{n^4} \tag{82}$$

$$\implies M^2\left(1 - c_1 - \frac{1}{n^4}\right) \leq (\alpha\sigma_{\mathsf{noise}}^2)^2\left((1 + c_1)n + \frac{1}{n^4}\right). \tag{83}$$

Since $c_1 < 1/2000$, and $n \geq 2000$, setting $\alpha\sigma_{\mathsf{noise}}^2 = 2$ yields $M^2 \leq 10n$ with high probability. □

Bounding $\delta_{\text{train}}\left(\bar{w}^{\text{use-core}}, \gamma^2\right)$

Corollary 3. *Under the parameter settings of Theorem 2, with high probability,*

$$\delta_{train}\left(\bar{w}^{\text{use-core}}, \gamma^2\right) \geq \Phi\left(\frac{1-(1+c_1)\gamma^2 - c_5 - \bar{w}_{\text{core}}^{\text{use-core}}}{\left|\bar{w}_{\text{core}}^{\text{use-core}}\sigma_{\text{core}}\right|}\right) - c_6, \tag{84}$$

for some constants $c_1 < 1/2000; c_5, c_6 < 1/1000$ where Φ is the Gaussian CDF.

Proof. The result follows from applying Proposition 3 (which computes a bound on the majority fraction of points that is $\gamma-$memorized) to $\bar{w}^{\text{use-core}}$, invoking Lemma 11, and plugging in $\bar{w}_{\text{spu}}^{\text{use-core}} = 0$. Note that when $\bar{w}_{\text{spu}}^{\text{use-core}} = 0$, $\delta_{\text{train}}\left(\bar{w}^{\text{use-core}}, \gamma^2\right) = \delta_{\text{maj-train}}\left(\bar{w}^{\text{use-core}}, \gamma^2\right)$. □

Finally, the above bound on $\delta_{\text{train}}\left(\bar{w}^{\text{use-core}}, \gamma^2\right)$ translates to a bound on the norm $\|\bar{w}^{\text{use-core}}\|$ via simple algebra. For γ that satisfies $1-(1+c_1)\gamma^2 - c_5 > 0$:

$$\delta_{\text{train}}\left(\bar{w}^{\text{use-core}}, \gamma^2\right) \geq \Phi\left(\frac{-1}{\sigma_{\text{core}}} + \frac{1-(1+c_1)\gamma^2 - c_5}{\left|\bar{w}_{\text{core}}^{\text{use-core}}\sigma_{\text{core}}\right|}\right) - c_6 \tag{85}$$

$$\geq \Phi\left(\frac{-1}{\sigma_{\text{core}}}\right) - c_6. \tag{86}$$

Plugging the above lower bound into the bound on $\|\bar{w}^{\text{use-core}}\|$ from Corollary 2, we have

$$\|\bar{w}^{\text{use-core}}\|_2^2 \geq \frac{\gamma^4(1-c_1)}{\sigma_{\text{noise}}^2}\delta_{\text{train}}\left(\bar{w}^{\text{use-core}}, \gamma^2\right) n_{\text{maj}} - \frac{10}{\sigma_{\text{noise}}^2 n^3} \tag{87}$$

$$\geq \frac{n}{\sigma_{\text{noise}}^2}\left(\Phi\left(\frac{-1}{\sigma_{\text{core}}}\right) - c_6\right)\gamma^4(1-c_1) - \frac{10}{\sigma_{\text{noise}}^2 n^3} \tag{88}$$

$$\geq \frac{n}{\sigma_{\text{noise}}^2}\underbrace{\left[\left(\Phi\left(\frac{-1}{\sigma_{\text{core}}}\right) - c_6\right)\gamma^4(1-c_1) - c_7\right]}_{\text{set to } \gamma_3} \tag{89}$$

for some $c_7 < 1/1000$. □

B.3. Underparameterized regime

So far, we have studied the overparameterized regime for the data distribution described in Section 5. In the overparameterized setting, where the dimension of noise features N is very large, logistic regression (both ERM and reweighted) leads to max-margin classifiers. We showed that for some setting of parameters $n_{\text{maj}}, n_{\text{min}}, \sigma_{\text{spu}}, \sigma_{\text{core}}$, the robust error of such max-margin classifiers can be $> 2/3$, worse than random guessing. How does the same reweighted logistic regression perform in the underparameterized regime? We focus on the setting where $N = 0$. In this setting, the data is two-dimensional, and w.h.p., the training data is not linearly separable unless $\sigma_{\text{core}} = 0$. Consequently, the learned model $\hat{w}^{\text{rw}}\mathbb{R}^2$ that minimizes the reweighted training loss is not generally a max-margin separator.

For intuition, consider the following two sets of models, which are analogous to what we considered in Equation 12 in the main text for the overparameterized regime:

$$\mathcal{W}^{\text{use-spu}} \stackrel{\text{def}}{=} \{w \in \mathbb{R}^2 \text{ such that } w_{\text{core}} = 0\}$$
$$\mathcal{W}^{\text{use-core}} \stackrel{\text{def}}{=} \{w \in \mathbb{R}^2 \text{ such that } w_{\text{spu}} = 0\}. \tag{90}$$

The first set $\mathcal{W}^{\text{use-spu}}$ comprises models that use the spurious feature but not the core feature, and the second set $\mathcal{W}^{\text{use-core}}$ comprises models that use the core feature but not the spurious feature. Models in $\mathcal{W}^{\text{use-spu}}$ that exclusively use x_{spu} will have high training loss on the minorities since the minority points cannot be memorized. Due to upweighting the minorities, these models will have high reweighted training loss. On the other hand, models in $\mathcal{W}^{\text{use-core}}$ exclusively use the core

features that are informative for the label y across all groups. Hence they obtain reasonable loss across all groups and have smaller reweighted training loss than models in $\mathcal{W}^{\mathsf{use-spu}}$.

We will show in this section that the population minimizer of the reweighted loss is indeed in $\mathcal{W}^{\mathsf{use-core}}$ and bound the asymptotic variance of the reweighted estimator, leading to the final result in Theorem 1. Our approach is to study the asypmtotic behavior of the reweighted estimator when the number of data points $n \gg d$.

Data distribution. We first recap the data generating distribution (described in Section 5). $x = [x_{\mathsf{core}}, x_{\mathsf{spu}}]$ where,

$$x_{\mathsf{core}} \mid y \sim \mathcal{N}(y, \sigma^2_{\mathsf{core}}), \quad x_{\mathsf{spu}} \mid a \sim \mathcal{N}(a, \sigma^2_{\mathsf{spu}}),$$

For p_{maj} fraction of points, we have $a = y$ (majority points) and for $1 - p_{\mathsf{maj}}$ fraction of points, we have $a = -y$ (minority points).

Reweighted logistic loss. Let p_{maj} be the fraction of the majority group points and $(1 - p_{\mathsf{maj}})$ be the fraction of minority points. In order to use standard results from the asymptotics of M-estimators, we rewrite the reweighted estimator (defined in Section 2) as the minimizer of the following loss over n training points $[x_i, y_i]_{i=1}^n$.

$$\hat{w}^{\mathsf{rw}} = \arg\min \frac{1}{n} \sum_{i=1}^{n} \ell_{\mathsf{rw}}(x_i, y_i, w) \tag{91}$$

$$\ell_{\mathsf{rw}}(x, y, w) = \frac{-1}{p_{\mathsf{maj}}} \log \left(\frac{1}{1 + \exp(-y w^\top x)} \right), \text{ For } (x, y) \text{ from majority group} \tag{92}$$

$$\ell_{\mathsf{rw}}(x, y, w) = \frac{-1}{1 - p_{\mathsf{maj}}} \log \left(\frac{1}{1 + \exp(-y w^\top x)} \right), \text{ For } (x, y) \text{ from minority group.} \tag{93}$$

We follow the standard steps of asymptotic analysis where we:

1. Compute the population minimizer $w^\star$ that satisfies $\nabla \mathrm{L}_{\mathsf{rw}}(w^\star) = 0$, where $\mathrm{L}_{\mathsf{rw}}(w^\star) = \mathbb{E}[\ell_{\mathsf{rw}}(x, y, w^\star)]$.
2. Bound the asymptotic variance $\nabla^2 \mathrm{L}_{\mathsf{rw}}(w^\star)^{-1} \operatorname{Cov}[\nabla \ell_{\mathsf{rw}}(x, y, w^\star)] \nabla^2 \mathrm{L}_{\mathsf{rw}}(w^\star)^{-1}$.

Proposition 4. *For the data distribution under study, the population minimizer $w^\star$ that satisfies $\nabla L_{\mathsf{rw}}(w^\star) = 0$ is the following.*

$$w^\star = \left[\frac{2}{\sigma^2_{\mathsf{core}}}, 0 \right]. \tag{94}$$

This is a very important property in the underparameterized regime: the population minimizer has the best possible worst-group error by only using the core feature and not the spurious feature.

Proposition 5. *The asymptotic distribution of the reweighted logistic regression estimator is as follows.*

$$\sqrt{n}(\hat{w} - w^\star) \to^d \mathcal{N}(0, V), \tag{95}$$

$$V \preceq \operatorname{diag}\left(\frac{16 \exp\left(\frac{8}{(\sigma^2_{\mathsf{core}}+8)\sigma^2_{\mathsf{core}}}\right)(\sigma^2_{\mathsf{core}} + 1)(1 + 8/\sigma^2_{\mathsf{core}})^3}{p_{\mathsf{maj}}(1 - p_{\mathsf{maj}})(\sigma^2_{\mathsf{core}} + 9)^2}, \frac{16 \exp\left(\frac{8}{(\sigma^2_{\mathsf{core}}+8)\sigma^2_{\mathsf{core}}}\right)(1 + 8/\sigma^2_{\mathsf{core}})}{p_{\mathsf{maj}}(1 - p_{\mathsf{maj}})(\sigma^2_{\mathsf{spu}} + 1)} \right). \tag{96}$$

For $\sigma_{\mathsf{core}} \geq 1$, we have

$$V \preceq \operatorname{diag}\left(\frac{C_1}{p_{\mathsf{maj}}(1 - p_{\mathsf{maj}})}, \frac{C_2}{p_{\mathsf{maj}}(1 - p_{\mathsf{maj}})} \right), \tag{97}$$

for some constants C_1, C_2.

We see that the asymptotic variance increases as p_{maj} increases. This is expected because the reweighted estimator upweights the minority points by inverse of group size. As these weights increase, the variance also increases. However, as we noted before, since the population minimizer has small worst-group error, for large enough training set size, we get small worst-group error since the asymptotic variance is finite (for fixed p_{maj}) and the estimator approaches the population minimizer.

We now prove Theorem 1 for the underparameterized regime, restated as Theorem 3 below.

Theorem 3. *In the underparameterized regime with $N = 0$, for $p_{\mathsf{maj}} = \left(1 - \frac{1}{2001}\right)$, $\sigma^2_{\mathsf{core}} = 1$, and $\sigma^2_{\mathsf{spu}} = 0$, in the asymptotic regime with $n_{\mathsf{maj}}, n_{\mathsf{min}} \to \infty$, we have*

$$\mathit{Err}_{\mathsf{wg}}(\hat{w}^{\mathsf{rw}}) < 1/4. \tag{98}$$

Proof. We now put the two Propositions 5 and 4 together. We have $\hat{w}^{\mathsf{rw}}_{\mathsf{core}} \geq 2 - \epsilon_1$ and $|\hat{w}^{\mathsf{rw}}_{\mathsf{spu}}| \leq \epsilon_2$ for $\epsilon_1, \epsilon_2 < 1/10$, i.e the estimator is very close to the population minimizer. This follows from setting $\sigma_{\mathsf{core}}, \sigma_{\mathsf{spu}}, p_{\mathsf{maj}} = \frac{n_{\mathsf{maj}}}{n_{\mathsf{maj}}+n_{\mathsf{min}}}$ to their corresponding values and setting $n = n_{\mathsf{maj}} + n_{\mathsf{min}}$ to be large enough. In order to compute the worst-group error, WLOG consider points with label $y = 1$ (labels are balanced in the population). For a point from the majority group, the probability of misclassification is as follows.

$$\Pr[\hat{w}^{\mathsf{rw}}_{\mathsf{core}} x_{\mathsf{core}} + \hat{w}^{\mathsf{rw}}_{\mathsf{spu}} x_{\mathsf{spu}} \geq 0] = \Pr[z \geq \frac{\hat{w}^{\mathsf{rw}}_{\mathsf{core}} + \hat{w}^{\mathsf{rw}}_{\mathsf{spu}}}{\sigma^2_{\mathsf{core}} \hat{w}^{\mathsf{rw}\ 2}_{\mathsf{core}} + \sigma^2_{\mathsf{spu}} \hat{w}^{\mathsf{rw}\ 2}_{\mathsf{spu}}}], \tag{99}$$

where $z \sim \mathcal{N}(0, 1)$.

Similarly, for the minority group, the probability of misclassification is

$$\Pr[z \geq \frac{\hat{w}^{\mathsf{rw}}_{\mathsf{core}} - \hat{w}^{\mathsf{rw}}_{\mathsf{spu}}}{\sigma^2_{\mathsf{core}} \hat{w}^{\mathsf{rw}\ 2}_{\mathsf{core}} + \sigma^2_{\mathsf{spu}} \hat{w}^{\mathsf{rw}\ 2}_{\mathsf{spu}}}], \text{ where } z \sim \mathcal{N}(0, 1). \tag{100}$$

Therefore, the worst-group error of $\hat{w}^{\mathsf{rw}}$ can be bounded as.

$$\mathrm{Err}_{\mathsf{wg}}(\hat{w}^{\mathsf{rw}}) \leq 1 - \Phi\left(\frac{\hat{w}^{\mathsf{rw}}_{\mathsf{core}} - |\hat{w}^{\mathsf{rw}}_{\mathsf{spu}}|}{\sigma^2_{\mathsf{core}} \hat{w}^{\mathsf{rw}\ 2}_{\mathsf{core}} + \sigma^2_{\mathsf{spu}} \hat{w}^{\mathsf{rw}\ 2}_{\mathsf{spu}}}\right), \tag{101}$$

where Φ is the Gaussian CDF. Substituting $\sigma_{\mathsf{core}} = 1, \sigma_{\mathsf{spu}} = 0, \hat{w}^{\mathsf{rw}}_{\mathsf{core}} \geq 2 - \epsilon_1, |\hat{w}^{\mathsf{rw}}_{\mathsf{spu}}| \leq \epsilon_2$ gives the required result that $\mathrm{Err}_{\mathsf{wg}}(\hat{w}^{\mathsf{rw}}) < 1/4$. In contrast, in the overparameterized regime where $N \gg n$, even for very large n, the reweighted estimator has high worst-group error, as shown in Theorem 1. □

B.3.1. Complete proofs

We now provide the proofs for Proposition 4 and Proposition 5 which mostly follow from straightforward algebra.

Proposition 4. *For the data distribution under study, the population minimizer $w^\star$ that satisfies $\nabla L_{\mathsf{rw}}(w^\star) = 0$ is the following.*

$$w^\star = \left[\frac{2}{\sigma^2_{\mathsf{core}}}, 0\right]. \tag{94}$$

Proof. For convenience, we compute expectations over the majority and minority groups separately and express the population loss L_{rw} as the weighted sum of the two terms. Recall that we denote $x = [x_{\mathsf{core}}, x_{\mathsf{spu}}]$.

$$\mathrm{L}_{\mathsf{rw}}(w) = p_{\mathsf{maj}} \mathrm{L}_{\text{rw-maj}} + (1 - p_{\mathsf{maj}}) \mathrm{L}_{\text{rw-min}} \tag{102}$$

$$\mathrm{L}_{\text{rw-maj}}(w) = \mathbb{E}_y \mathbb{E}_{x_{\mathsf{core}} \sim \mathcal{N}(y, \sigma^2_{\mathsf{core}})} \mathbb{E}_{x_{\mathsf{spu}} \sim \mathcal{N}(y, \sigma^2_{\mathsf{spu}})} [\ell_{\mathsf{rw}}(x, y, w)]. \tag{103}$$

$$\mathrm{L}_{\text{rw-min}}(w) = \mathbb{E}_y \mathbb{E}_{x_{\mathsf{core}} \sim \mathcal{N}(y, \sigma^2_{\mathsf{core}})} \mathbb{E}_{x_{\mathsf{spu}} \sim \mathcal{N}(-y, \sigma^2_{\mathsf{spu}})} [\ell_{\mathsf{rw}}(x, y, w)]. \tag{104}$$

We use the following expression for computing the population gradient.

$$\nabla \log\left(\frac{1}{1 + \exp(-y w^\top x)}\right) = \left(\frac{-y \exp(-y w^\top x)}{1 + \exp(-y w^\top x)}\right) x. \tag{105}$$

Combining the definition of the reweighted loss and population losses (Equation 91 and Equation 102) with the gradient expression above gives the following.

$$\nabla \mathrm{L}_{\text{rw-maj}}(w) = \mathbb{E}_y \mathbb{E}_{x_{\text{core}} \sim \mathcal{N}(y, \sigma^2_{\text{core}})} \mathbb{E}_{x_{\text{spu}} \sim \mathcal{N}(y, \sigma^2_{\text{spu}})} \left[\frac{1}{p_{\text{maj}}} \left(\frac{-y \exp(-y w^\top x)}{1 + \exp(-y w^\top x)} \right) x \right]. \tag{106}$$

$$\nabla \mathrm{L}_{\text{rw-min}}(w) = \mathbb{E}_y \mathbb{E}_{x_{\text{core}} \sim \mathcal{N}(y, \sigma^2_{\text{core}})} \mathbb{E}_{x_{\text{spu}} \sim \mathcal{N}(-y, \sigma^2_{\text{spu}})} \left[\frac{1}{1 - p_{\text{maj}}} \left(\frac{-y \exp(-y w^\top x)}{1 + \exp(-y w^\top x)} \right) x \right]. \tag{107}$$

Now we compute $\nabla \mathrm{L}_{\text{rw}}(w^\star) = p_{\text{maj}} \nabla \mathrm{L}_{\text{rw-maj}}(w^\star) + (1 - p_{\text{maj}}) \nabla \mathrm{L}_{\text{rw-min}}(w^\star)$. First we compute wrt the spurious attribute $\nabla_{\text{spu}} \mathrm{L}_{\text{rw}}(w^\star)$. For convenience, let $c = \frac{2}{\sigma^2_{\text{core}}}$.

$$\begin{aligned}
\nabla_{\text{spu}} \mathrm{L}_{\text{rw-maj}}(w^\star) &= \mathbb{E}_y \mathbb{E}_{x_{\text{core}} \sim \mathcal{N}(y, \sigma^2_{\text{core}})} \mathbb{E}_{x_{\text{spu}} \sim \mathcal{N}(y, \sigma^2_{\text{spu}})} \left[\frac{1}{p_{\text{maj}}} \left(\frac{-y \exp(-y c x_{\text{core}})}{1 + \exp(-y c x_{\text{core}})} \right) x_{\text{spu}} \right] \\
&= \frac{1}{2} \mathbb{E}_{x_{\text{core}} \sim \mathcal{N}(1, \sigma^2_{\text{core}})} \mathbb{E}_{x_{\text{spu}} \sim \mathcal{N}(1, \sigma^2_{\text{spu}})} \left[\frac{1}{p_{\text{maj}}} \left(\frac{-\exp(-c x_{\text{core}})}{1 + \exp(-c x_{\text{core}})} \right) x_{\text{spu}} \right] \\
&\quad + \frac{1}{2} \mathbb{E}_{x_{\text{core}} \sim \mathcal{N}(-1, \sigma^2_{\text{core}})} \mathbb{E}_{x_{\text{spu}} \sim \mathcal{N}(-1, \sigma^2_{\text{spu}})} \left[\frac{1}{p_{\text{maj}}} \left(\frac{\exp(c x_{\text{core}})}{1 + \exp(c x_{\text{core}})} \right) x_{\text{spu}} \right] \\
&= \frac{1}{2} \mathbb{E}_{x_{\text{core}} \sim \mathcal{N}(1, \sigma^2_{\text{core}})} \left[\frac{1}{p_{\text{maj}}} \left(\frac{-\exp(-c x_{\text{core}})}{1 + \exp(-c x_{\text{core}})} \right) \right] - \frac{1}{2} \mathbb{E}_{x_{\text{core}} \sim \mathcal{N}(-1, \sigma^2_{\text{core}})} \left[\frac{1}{p_{\text{maj}}} \left(\frac{\exp(c x_{\text{core}})}{1 + \exp(c x_{\text{core}})} \right) \right] \\
&= \frac{1}{2} \mathbb{E}_{x_{\text{core}} \sim \mathcal{N}(1, \sigma^2_{\text{core}})} \left[\frac{1}{p_{\text{maj}}} \left(\frac{-\exp(-c x_{\text{core}})}{1 + \exp(-c x_{\text{core}})} \right) \right] \underbrace{- \frac{1}{2} \mathbb{E}_{x_{\text{core}} \sim \mathcal{N}(1, \sigma^2_{\text{core}})} \left[\frac{1}{p_{\text{maj}}} \left(\frac{\exp(-c x_{\text{core}})}{1 + \exp(-c x_{\text{core}})} \right) \right]}_{\text{Replacing } x_{\text{core}} \sim \mathcal{N}(-1, \sigma^2_{\text{core}}) \text{ with } -x_{\text{core}} \sim \mathcal{N}(1, \sigma^2_{\text{core}})} \\
&= \mathbb{E}_{x_{\text{core}} \sim \mathcal{N}(1, \sigma^2_{\text{core}})} \left[\frac{1}{p_{\text{maj}}} \left(\frac{-\exp(-c x_{\text{core}})}{1 + \exp(-c x_{\text{core}})} \right) \right] \\
\nabla_{\text{spu}} \mathrm{L}_{\text{rw-min}}(w^\star) &= \mathbb{E}_y \mathbb{E}_{x_{\text{core}} \sim \mathcal{N}(y, \sigma^2_{\text{core}})} \mathbb{E}_{x_{\text{spu}} \sim \mathcal{N}(-y, \sigma^2_{\text{spu}})} \left[\frac{1}{1 - p_{\text{maj}}} \left(\frac{-y \exp(-y c x_{\text{core}})}{1 + \exp(-y c x_{\text{core}})} \right) x_{\text{spu}} \right] \\
&= \frac{1}{2} \mathbb{E}_{x_{\text{core}} \sim \mathcal{N}(1, \sigma^2_{\text{core}})} \left[\frac{1}{p_{\text{maj}}} \left(\frac{\exp(-c x_{\text{core}})}{1 + \exp(-c x_{\text{core}})} \right) \right] + \frac{1}{2} \mathbb{E}_{x_{\text{core}} \sim \mathcal{N}(-1, \sigma^2_{\text{core}})} \left[\frac{1}{p_{\text{maj}}} \left(\frac{\exp(c x_{\text{core}})}{1 + \exp(c x_{\text{core}})} \right) \right] \\
&= \mathbb{E}_{x_{\text{core}} \sim \mathcal{N}(1, \sigma^2_{\text{core}})} \left[\frac{1}{1 - p_{\text{maj}}} \left(\frac{\exp(-c x_{\text{core}})}{1 + \exp(-c x_{\text{core}})} \right) \right]
\end{aligned}$$

Now we take the weighted combination of $\nabla_{\text{spu}} \mathrm{L}_{\text{rw-maj}}(w^\star)$ and $\nabla_{\text{spu}} \mathrm{L}_{\text{rw-min}}(w^\star)$, based on the fraction of the majority and minority samples in the population, which makes the two terms cancel out.

$$\nabla_{\text{spu}} \mathrm{L}_{\text{rw}} = p_{\text{maj}} \nabla_{\text{spu}} \mathrm{L}_{\text{rw-maj}}(w^\star) + (1 - p_{\text{maj}}) \nabla_{\text{spu}} \mathrm{L}_{\text{rw-min}}(w^\star) = 0. \tag{108}$$

Now we compute $\nabla_{\mathsf{core}} \mathrm{L}_{\mathsf{rw}}(w^\star)$.

$$\begin{aligned}
\nabla_{\mathsf{core}} \mathrm{L}_{\text{rw-maj}}(w^\star) &= \mathbb{E}_y \mathbb{E}_{x_{\mathsf{core}} \sim \mathcal{N}(y, \sigma_{\mathsf{core}}^2)} \mathbb{E}_{x_{\mathsf{spu}} \sim \mathcal{N}(y, \sigma_{\mathsf{spu}}^2)} \left[\frac{1}{p_{\mathsf{maj}}} \left(\frac{-y \exp(-ycx_{\mathsf{core}})}{1 + \exp(-ycx_{\mathsf{core}})} \right) x_{\mathsf{core}} \right] \\
&= \frac{1}{2} \mathbb{E}_{x_{\mathsf{core}} \sim \mathcal{N}(1, \sigma_{\mathsf{core}}^2)} \left[\frac{1}{p_{\mathsf{maj}}} \left(\frac{-\exp(-cx_{\mathsf{core}})}{1 + \exp(-cx_{\mathsf{core}})} \right) x_{\mathsf{core}} \right] + \frac{1}{2} \mathbb{E}_{x_{\mathsf{core}} \sim \mathcal{N}(-1, \sigma_{\mathsf{core}}^2)} \left[\frac{1}{p_{\mathsf{maj}}} \left(\frac{\exp(cx_{\mathsf{core}})}{1 + \exp(cx_{\mathsf{core}})} \right) x_{\mathsf{core}} \right] \\
&= \frac{1}{2} \mathbb{E}_{x_{\mathsf{core}} \sim \mathcal{N}(1, \sigma_{\mathsf{core}}^2)} \left[\frac{1}{p_{\mathsf{maj}}} \left(\frac{-\exp(-cx_{\mathsf{core}})}{1 + \exp(-cx_{\mathsf{core}})} \right) x_{\mathsf{core}} \right] + \frac{1}{2} \mathbb{E}_{x_{\mathsf{core}} \sim \mathcal{N}(-1, \sigma_{\mathsf{core}}^2)} \left[\frac{1}{p_{\mathsf{maj}}} \left(\frac{1}{1 + \exp(-cx_{\mathsf{core}})} \right) x_{\mathsf{core}} \right] \\
&= \frac{1}{2p_{\mathsf{maj}}} \frac{1}{\sigma_{\mathsf{core}} \sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{\exp(-cx_{\mathsf{core}}) \exp\left(\frac{-(x-1)^2}{2\sigma_{\mathsf{core}}^2} \right) - \exp\left(\frac{-(x+1)^2}{2\sigma_{\mathsf{core}}^2} \right)}{1 + \exp(-cx_{\mathsf{core}})} x_{\mathsf{core}} \, dx_{\mathsf{core}} \\
&= \frac{1}{2p_{\mathsf{maj}}} \frac{1}{\sigma_{\mathsf{core}} \sqrt{2\pi}} \int_{-\infty}^{\infty} 0 \, dx_{\mathsf{core}}, \text{ Substituting } c = \frac{2}{\sigma_{\mathsf{core}}^2} \\
&= 0.
\end{aligned}$$

Similarly, we get $\nabla_{\mathsf{core}} \mathrm{L}_{\text{rw-min}}(w^\star) = 0$ and hence proved that $\nabla_{\mathsf{core}} \mathrm{L}_{\mathsf{rw}}(w^\star) = 0$. □

Lemma 12. *The following is true.*

$$\mathrm{Cov}[\nabla \ell_{\mathsf{rw}}(x, y, w^\star)] \preceq \mathrm{diag}\left(\frac{\sigma_{\mathsf{core}}^2 + 1}{p_{\mathsf{maj}}(1 - p_{\mathsf{maj}})}, \frac{\sigma_{\mathsf{spu}}^2 + 1}{p_{\mathsf{maj}}(1 - p_{\mathsf{maj}})} \right). \tag{109}$$

We now compute the asymptotic variance which involves computing $\nabla^2 L(w^\star)$ and $\mathrm{Cov}[\nabla \ell_{\mathsf{rw}}(w^\star)]$.

Proof. First, we show that the off-diagonal entries of $\mathrm{Cov}[\ell_{\mathsf{rw}}(x, y, w^\star)]$ are zero.

$$\begin{aligned}
&\mathbb{E}[\nabla_{\mathsf{core}} \ell_{\mathsf{rw}}(x, y, w^\star) \nabla_{\mathsf{spu}} \ell_{\mathsf{rw}}(x, y, w^\star)] - \mathbb{E}[\nabla_{\mathsf{core}} \ell_{\mathsf{rw}}(x, y, w^\star)] \mathbb{E}[\nabla_{\mathsf{spu}} \ell_{\mathsf{rw}}(x, y, w^\star)] \\
&= \mathbb{E}[\nabla_{\mathsf{core}} \ell_{\mathsf{rw}}(x, y, w^\star) \nabla_{\mathsf{spu}} \ell_{\mathsf{rw}}(x, y, w^\star)] \\
&= p_{\mathsf{maj}} \mathbb{E}_y \mathbb{E}_{x_{\mathsf{core}} \sim \mathcal{N}(y, \sigma_{\mathsf{core}}^2)} \mathbb{E}_{x_{\mathsf{spu}} \sim \mathcal{N}(y, \sigma_{\mathsf{spu}}^2)} \left[\frac{1}{p_{\mathsf{maj}}^2} \left(\frac{-y \exp(-ycx_{\mathsf{core}})}{1 + \exp(-ycx_{\mathsf{core}})} \right)^2 x_{\mathsf{core}} x_{\mathsf{spu}} \right] \\
&+ (1 - p_{\mathsf{maj}}) \mathbb{E}_y \mathbb{E}_{x_{\mathsf{core}} \sim \mathcal{N}(y, \sigma_{\mathsf{core}}^2)} \mathbb{E}_{x_{\mathsf{spu}} \sim \mathcal{N}(-y, \sigma_{\mathsf{spu}}^2)} \left[\frac{1}{(1 - p_{\mathsf{maj}})^2} \left(\frac{-y \exp(-ycx_{\mathsf{core}})}{1 + \exp(-ycx_{\mathsf{core}})} \right)^2 x_{\mathsf{core}} x_{\mathsf{spu}} \right] \\
&= \mathbb{E}_y \mathbb{E}_{x_{\mathsf{core}} \sim \mathcal{N}(y, \sigma_{\mathsf{core}}^2)} \left[\frac{1}{p_{\mathsf{maj}}} \left(\frac{-y \exp(-ycx_{\mathsf{core}})}{1 + \exp(-ycx_{\mathsf{core}})} \right)^2 y \right] \\
&- \mathbb{E}_y \mathbb{E}_{x_{\mathsf{core}} \sim \mathcal{N}(y, \sigma_{\mathsf{core}}^2)} \left[\frac{1}{1 - p_{\mathsf{maj}}} \left(\frac{-y \exp(-ycx_{\mathsf{core}})}{1 + \exp(-ycx_{\mathsf{core}})} \right)^2 y \right] \\
&= \frac{1 - 2p_{\mathsf{maj}}}{2p_{\mathsf{maj}}(1 - p_{\mathsf{maj}})} \mathbb{E}_{x_{\mathsf{core}} \sim \mathcal{N}(1, \sigma_{\mathsf{core}}^2)} \left[\left(\frac{\exp(-cx_{\mathsf{core}})}{1 + \exp(-cx_{\mathsf{core}})} \right)^2 \right] - \frac{1 - 2p_{\mathsf{maj}}}{2p_{\mathsf{maj}}(1 - p_{\mathsf{maj}})} \mathbb{E}_{x_{\mathsf{core}} \sim \mathcal{N}(-1, \sigma_{\mathsf{core}}^2)} \left[\left(\frac{\exp(cx_{\mathsf{core}})}{1 + \exp(cx_{\mathsf{core}})} \right)^2 \right] \\
&= \frac{1 - 2p_{\mathsf{maj}}}{2p_{\mathsf{maj}}(1 - p_{\mathsf{maj}})} \mathbb{E}_{x_{\mathsf{core}} \sim \mathcal{N}(1, \sigma_{\mathsf{core}}^2)} \left[\left(\frac{\exp(-cx_{\mathsf{core}})}{1 + \exp(-cx_{\mathsf{core}})} \right)^2 \right] - \frac{1 - 2p_{\mathsf{maj}}}{2p_{\mathsf{maj}}(1 - p_{\mathsf{maj}})} \mathbb{E}_{x_{\mathsf{core}} \sim \mathcal{N}(1, \sigma_{\mathsf{core}}^2)} \left[\left(\frac{\exp(-cx_{\mathsf{core}})}{1 + \exp(-cx_{\mathsf{core}})} \right)^2 \right] = 0.
\end{aligned}$$

Now, we bound the diagonal elements.

$$
\begin{aligned}
&\mathbb{E}[\nabla_{\mathsf{core}}(\ell_{\mathsf{rw}}(x,y,w^\star))^2] - (\mathbb{E}[\nabla_{\mathsf{core}}\ell_{\mathsf{rw}}(x,y,w^\star)])^2 \\
&= \mathbb{E}[\nabla_{\mathsf{core}}(\ell_{\mathsf{rw}}(x,y,w^\star))^2] \\
&= p_{\mathsf{maj}}\mathbb{E}_y\mathbb{E}_{x_{\mathsf{core}}\sim\mathcal{N}(y,\sigma^2_{\mathsf{core}})}\left[\frac{1}{p^2_{\mathsf{maj}}}\left(\frac{-y\exp(-ycx_{\mathsf{core}})}{1+\exp(-ycx_{\mathsf{core}})}\right)^2 x^2_{\mathsf{core}}\right] \\
&+ (1-p_{\mathsf{maj}})\mathbb{E}_y\mathbb{E}_{x_{\mathsf{core}}\sim\mathcal{N}(y,\sigma^2_{\mathsf{core}})}\left[\frac{1}{(1-p_{\mathsf{maj}})^2}\left(\frac{-y\exp(-ycx_{\mathsf{core}})}{1+\exp(-ycx_{\mathsf{core}})}\right)^2 x^2_{\mathsf{core}}\right] \\
&= \frac{1}{p_{\mathsf{maj}}(1-p_{\mathsf{maj}})}\mathbb{E}_y\mathbb{E}_{x_{\mathsf{core}}\sim\mathcal{N}(y,\sigma^2_{\mathsf{core}})}\left[\left(\frac{-y\exp(-ycx_{\mathsf{core}})}{1+\exp(-ycx_{\mathsf{core}})}\right)^2 x^2_{\mathsf{core}}\right] \\
&= \frac{1}{2p_{\mathsf{maj}}(1-p_{\mathsf{maj}})}\mathbb{E}_{x_{\mathsf{core}}\sim\mathcal{N}(1,\sigma^2_{\mathsf{core}})}\left[\left(\frac{-\exp(-cx_{\mathsf{core}})}{1+\exp(-cx_{\mathsf{core}})}\right)^2 x^2_{\mathsf{core}}\right] + \frac{1}{2p_{\mathsf{maj}}(1-p_{\mathsf{maj}})}\mathbb{E}_{x_{\mathsf{core}}\sim\mathcal{N}(-1,\sigma^2_{\mathsf{core}})}\left[\left(\frac{-\exp(cx_{\mathsf{core}})}{1+\exp(cx_{\mathsf{core}})}\right)^2 x^2_{\mathsf{core}}\right] \\
&= \frac{1}{p_{\mathsf{maj}}(1-p_{\mathsf{maj}})}\mathbb{E}_{x_{\mathsf{core}}\sim\mathcal{N}(1,\sigma^2_{\mathsf{core}})}\left[\left(\frac{-\exp(-cx_{\mathsf{core}})}{1+\exp(-cx_{\mathsf{core}})}\right)^2 x^2_{\mathsf{core}}\right] \\
&\leq \frac{1}{p_{\mathsf{maj}}(1-p_{\mathsf{maj}})}\mathbb{E}_{x_{\mathsf{core}}\sim\mathcal{N}(1,\sigma^2_{\mathsf{core}})}[x^2_{\mathsf{core}}] = \frac{\sigma^2_{\mathsf{core}}+1}{p_{\mathsf{maj}}(1-p_{\mathsf{maj}})}.
\end{aligned}
$$

Finally,

$$
\begin{aligned}
&\mathbb{E}[\nabla_{\mathsf{spu}}(\ell_{\mathsf{rw}}(x,y,w^\star))^2] - (\mathbb{E}[\nabla_{\mathsf{spu}}\ell_{\mathsf{rw}}(x,y,w^\star)])^2 \\
&= \mathbb{E}[\nabla_{\mathsf{spu}}(\ell_{\mathsf{rw}}(x,y,w^\star))^2] \\
&= p_{\mathsf{maj}}\mathbb{E}_y\mathbb{E}_{x_{\mathsf{core}}\sim\mathcal{N}(y,\sigma^2_{\mathsf{core}})}\mathbb{E}_{x_{\mathsf{spu}}\sim\mathcal{N}(y,\sigma^2_{\mathsf{spu}})}\left[\frac{1}{p^2_{\mathsf{maj}}}\left(\frac{-y\exp(-ycx_{\mathsf{core}})}{1+\exp(-ycx_{\mathsf{core}})}\right)^2 x^2_{\mathsf{spu}}\right] \\
&+ (1-p_{\mathsf{maj}})\mathbb{E}_y\mathbb{E}_{x_{\mathsf{core}}\sim\mathcal{N}(y,\sigma^2_{\mathsf{core}})}\mathbb{E}_{x_{\mathsf{spu}}\sim\mathcal{N}(-y,\sigma^2_{\mathsf{spu}})}\left[\frac{1}{(1-p_{\mathsf{maj}})^2}\left(\frac{-y\exp(-ycx_{\mathsf{core}})}{1+\exp(-ycx_{\mathsf{core}})}\right)^2 x^2_{\mathsf{spu}}\right] \\
&\leq \frac{1}{p_{\mathsf{maj}}}\mathbb{E}_y\mathbb{E}_{x_{\mathsf{spu}}\sim\mathcal{N}(y,\sigma^2_{\mathsf{spu}})}[x^2_{\mathsf{spu}}] + \frac{1}{1-p_{\mathsf{maj}}}\mathbb{E}_y\mathbb{E}_{x_{\mathsf{spu}}\sim\mathcal{N}(-y,\sigma^2_{\mathsf{spu}})}[x^2_{\mathsf{spu}}] = \frac{\sigma^2_{\mathsf{spu}}+1}{p_{\mathsf{maj}}(1-p_{\mathsf{maj}})}.
\end{aligned}
$$

□

Lemma 13. *The following is true.*

$$
\nabla^2 L_{\mathsf{rw}}(x,y,w^\star)] \succeq \operatorname{diag}\left(\frac{\exp\left(\frac{-4}{(\sigma^2_{\mathsf{core}}+8)\sigma^2_{\mathsf{core}}}\right)(\sigma^2_{\mathsf{core}}+9)}{4(1+8/\sigma^2_{\mathsf{core}})^{3/2}}, \frac{\exp\left(\frac{-4}{(\sigma^2_{\mathsf{core}}+8)\sigma^2_{\mathsf{core}}}\right)(\sigma^2_{\mathsf{spu}}+1)}{4\sqrt{1+8/\sigma^2_{\mathsf{core}}}}\right). \tag{110}
$$

Proof. We use the following expression for computing the population gradient.

$$
\nabla^2\log\left(\frac{1}{1+\exp(-yw^\top x)}\right) = \nabla\left(\frac{-y\exp(-yw^\top x)}{1+\exp(-yw^\top x)}\right)x \quad = \nabla\left(\frac{-y}{1+\exp(yw^\top x)}\right)x = \left(\frac{\exp(yw^\top x)}{(1+\exp(yw^\top x))^2}\right)xx^\top. \tag{111}
$$

Recall the definition of the population majority and minority losses (Equation 102).

$$\nabla^2 \mathrm{L}_{\text{rw-maj}}(w) = \mathbb{E}_y \mathbb{E}_{x_{\text{core}} \sim \mathcal{N}(y, \sigma^2_{\text{core}})} \mathbb{E}_{x_{\text{spu}} \sim \mathcal{N}(y, \sigma^2_{\text{spu}})} \left[\frac{1}{p_{\text{maj}}} \left(\frac{\exp(yw^\top x)}{(1+\exp(yw^\top x))^2} \right) xx^\top \right]. \tag{112}$$

$$\nabla^2 \mathrm{L}_{\text{rw-min}}(w) = \mathbb{E}_y \mathbb{E}_{x_{\text{core}} \sim \mathcal{N}(y, \sigma^2_{\text{core}})} \mathbb{E}_{x_{\text{spu}} \sim \mathcal{N}(-y, \sigma^2_{\text{spu}})} \left[\frac{1}{1-p_{\text{maj}}} \left(\frac{\exp(yw^\top x)}{(1+\exp(yw^\top x))^2} \right) xx^\top \right]. \tag{113}$$

Like previously, we first compute the off-diagonal entries.

$$\begin{aligned}
[\nabla^2 \mathrm{L}_{\text{rw-maj}}(w^\star)]_{\text{spu, core}} &= \frac{1}{p_{\text{maj}}} \mathbb{E}_y \mathbb{E}_{x_{\text{core}} \sim \mathcal{N}(y, \sigma^2_{\text{core}})} \mathbb{E}_{x_{\text{spu}} \sim \mathcal{N}(y, \sigma^2_{\text{spu}})} \left[\left(\frac{\exp(yw^{\star\top} x)}{(1+\exp(yw^{\star\top} x))^2} \right) x_{\text{core}} x_{\text{spu}} \right] \\
&\quad + \frac{1}{p_{\text{maj}}} \mathbb{E}_y \mathbb{E}_{x_{\text{core}} \sim \mathcal{N}(y, \sigma^2_{\text{core}})} \mathbb{E}_{x_{\text{spu}} \sim \mathcal{N}(-y, \sigma^2_{\text{spu}})} \left[\left(\frac{\exp(yw^{\star\top} x)}{(1+\exp(yw^{\star\top} x))^2} \right) x_{\text{core}} x_{\text{spu}} \right] \\
&= \frac{1}{p_{\text{maj}}} \mathbb{E}_y \mathbb{E}_{x_{\text{core}} \sim \mathcal{N}(y, \sigma^2_{\text{core}})} \mathbb{E}_{x_{\text{spu}} \sim \mathcal{N}(y, \sigma^2_{\text{spu}})} \left[\left(\frac{\exp(yw^{\star\top} x)}{(1+\exp(yw^{\star\top} x))^2} \right) x_{\text{core}} x_{\text{spu}} \right] \\
&\quad - \frac{1}{p_{\text{maj}}} \mathbb{E}_y \mathbb{E}_{x_{\text{core}} \sim \mathcal{N}(y, \sigma^2_{\text{core}})} \mathbb{E}_{x_{\text{spu}} \sim \mathcal{N}(y, \sigma^2_{\text{spu}})} \left[\left(\frac{\exp(yw^{\star\top} x)}{(1+\exp(yw^{\star\top} x))^2} \right) x_{\text{core}} x_{\text{spu}} \right] \\
&= 0 \\
[\nabla^2 \mathrm{L}_{\text{rw-min}}(w^\star)]_{\text{spu, core}} &= 0, \text{ Similar calculation as above} \\
[\nabla^2 \mathrm{L}_{\text{rw}}(w^\star)]_{\text{spu, core}} &= 0.
\end{aligned}$$

Now, we bound the diagonal entries. Recall that $w^\star_{\text{spu}} = 0$ and $w^\star_{\text{core}} = c$ where $c = \frac{2}{\sigma^2_{\text{core}}}$.

$$\begin{aligned}
[\nabla^2 \mathrm{L}_{\text{rw-maj}}(w^\star)]_{\text{core, core}} &= \frac{1}{p_{\text{maj}}} \mathbb{E}_y \mathbb{E}_{x_{\text{core}} \sim \mathcal{N}(y, \sigma^2_{\text{core}})} \left[\left(\frac{\exp(ycx_{\text{core}})}{(1+\exp(ycx_{\text{core}}))^2} \right) x^2_{\text{core}} \right] \\
&= \frac{1}{2p_{\text{maj}}} \mathbb{E}_{x_{\text{core}} \sim \mathcal{N}(1, \sigma^2_{\text{core}})} \left[\left(\frac{\exp(cx_{\text{core}})}{(1+\exp(cx_{\text{core}}))^2} \right) x^2_{\text{core}} \right] + \frac{1}{2p_{\text{maj}}} \mathbb{E}_{x_{\text{core}} \sim \mathcal{N}(-1, \sigma^2_{\text{core}})} \left[\left(\frac{\exp(-cx_{\text{core}})}{(1+\exp(-cx_{\text{core}}))^2} \right) x^2_{\text{core}} \right] \\
&= \frac{1}{p_{\text{maj}}} \mathbb{E}_{x_{\text{core}} \sim \mathcal{N}(1, \sigma^2_{\text{core}})} \left[\left(\frac{\exp(cx_{\text{core}})}{(1+\exp(cx_{\text{core}}))^2} \right) x^2_{\text{core}} \right] \\
&\geq \frac{1}{p_{\text{maj}}} \frac{1}{4} \mathbb{E}_{x_{\text{core}} \sim \mathcal{N}(1, \sigma^2_{\text{core}})} \left[\exp(-c^2 x^2_{\text{core}}) x^2_{\text{core}} \right] \\
&= \frac{1}{p_{\text{maj}}} \frac{1}{4\sigma_{\text{core}}\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(-c^2 x^2_{\text{core}}) \exp\left(\frac{-(x_{\text{core}}-1)^2}{2\sigma^2_{\text{core}}} \right) x^2_{\text{core}} \, dx_{\text{core}} \\
&= \frac{1}{p_{\text{maj}}} \frac{1}{4\sigma_{\text{core}}\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\Big(-\frac{8x^2_{\text{core}}/\sigma^2_{\text{core}}}{2\sigma^2_{\text{core}}} \Big) \exp\left(\frac{-(x_{\text{core}}-1)^2}{2\sigma^2_{\text{core}}} \right) x^2_{\text{core}} \, dx_{\text{core}} \\
&= \frac{1}{p_{\text{maj}}} \frac{\exp\left(\frac{-8}{(\sigma^2_{\text{core}}+8)\sigma^2_{\text{core}}} \right)}{4\sigma_{\text{core}}\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\Big(\frac{-(\sqrt{1+8/\sigma^2_{\text{core}}} x_{\text{core}} - \frac{1}{\sqrt{1+8/\sigma^2_{\text{core}}}})^2}{2\sigma^2_{\text{core}}} \Big) x^2_{\text{core}} \, dx_{\text{core}} \\
&= \frac{1}{p_{\text{maj}}} \frac{\exp\left(\frac{-8}{(\sigma^2_{\text{core}}+8)\sigma^2_{\text{core}}} \right)(\sigma^2_{\text{core}}+9)}{4(1+8/\sigma^2_{\text{core}})^{5/2}}. \\
[\nabla^2 \mathrm{L}_{\text{rw-min}}(w^\star)]_{\text{core, core}} &= \frac{1}{1-p_{\text{maj}}} \frac{\exp\left(\frac{-8}{(\sigma^2_{\text{core}}+8)\sigma^2_{\text{core}}} \right)(\sigma^2_{\text{core}}+9)}{4(1+8/\sigma^2_{\text{core}})^{5/2}}, \text{ By symmetry.} \\
[\nabla^2 \mathrm{L}_{\text{rw}}(w^\star)]_{\text{core, core}} &= p_{\text{maj}} [\nabla^2 \mathrm{L}_{\text{rw-maj}}(w^\star)]_{\text{core, core}} + (1-p_{\text{maj}}) [\nabla^2 \mathrm{L}_{\text{rw-min}}(w^\star)]_{\text{core, core}} \\
&= \frac{\exp\left(\frac{-8}{(\sigma^2_{\text{core}}+8)\sigma^2_{\text{core}}} \right)(\sigma^2_{\text{core}}+9)}{4(1+8/\sigma^2_{\text{core}})^{5/2}}.
\end{aligned}$$

Finally, we calculate $[\nabla^2 \mathrm{L}_{\text{rw-maj}}(w^\star)]_{\text{spu, spu}}$ as follows.

$$
\begin{aligned}
[\nabla^2 \mathrm{L}_{\text{rw-maj}}(w^\star)]_{\text{spu, spu}} &= \frac{1}{p_{\text{maj}}} \mathbb{E}_y \mathbb{E}_{x_{\text{core}} \sim \mathcal{N}(y, \sigma_{\text{core}}^2)} \mathbb{E}_{x_{\text{spu}} \sim \mathcal{N}(y, \sigma_{\text{spu}}^2)} \left[\left(\frac{\exp(ycx_{\text{core}})}{(1+\exp(ycx_{\text{core}}))^2} \right) x_{\text{spu}}^2 \right] \\
&= \frac{1}{2p_{\text{maj}}} \mathbb{E}_{x_{\text{core}} \sim \mathcal{N}(1, \sigma_{\text{core}}^2)} \left[\left(\frac{\exp(cx_{\text{core}})}{(1+\exp(cx_{\text{core}}))^2} \right) \right] (\sigma_{\text{spu}}^2 + 1) \\
&+ \frac{1}{2p_{\text{maj}}} \mathbb{E}_{x_{\text{core}} \sim \mathcal{N}(-1, \sigma_{\text{core}}^2)} \left[\left(\frac{\exp(-cx_{\text{core}})}{(1+\exp(-cx_{\text{core}}))^2} \right) \right] (\sigma_{\text{spu}}^2 + 1) \\
&\geq \frac{1}{4p_{\text{maj}}} \mathbb{E}_{x_{\text{core}} \sim \mathcal{N}(1, \sigma_{\text{core}}^2)} [\exp(-c^2 x_{\text{core}}^2)] (\sigma_{\text{spu}}^2 + 1) \\
&= \frac{1}{4p_{\text{maj}}} \frac{\exp\left(\frac{-4}{(\sigma_{\text{core}}^2+8)\sigma_{\text{core}}^2} \right)}{\sqrt{1 + 8/\sigma_{\text{core}}^2}} (\sigma_{\text{spu}}^2 + 1) \\
[\nabla^2 \mathrm{L}_{\text{rw-min}}(w^\star)]_{\text{spu, spu}} &= \frac{1}{4(1-p_{\text{maj}})} \frac{\exp\left(\frac{-4}{(\sigma_{\text{core}}^2+8)\sigma_{\text{core}}^2} \right)}{\sqrt{1 + 8/\sigma_{\text{core}}^2}} (\sigma_{\text{spu}}^2 + 1), \text{ By symmetry.} \\
[\nabla^2 \mathrm{L}_{\text{rw}}(w^\star)]_{\text{spu, spu}} &= \frac{\exp\left(\frac{-4}{(\sigma_{\text{core}}^2+8)\sigma_{\text{core}}^2} \right)(\sigma_{\text{spu}}^2 + 1)}{4\sqrt{1 + 8/\sigma_{\text{core}}^2}}.
\end{aligned}
$$

□

Proposition 5. *The asymptotic distribution of the reweighted logistic regression estimator is as follows.*

$$
\sqrt{n}(\hat{w} - w^\star) \to^d \mathcal{N}(0, V), \tag{95}
$$

$$
V \preceq \operatorname{diag}\left(\frac{16 \exp\left(\frac{8}{(\sigma_{\text{core}}^2+8)\sigma_{\text{core}}^2} \right)(\sigma_{\text{core}}^2 + 1)(1 + 8/\sigma_{\text{core}}^2)^3}{p_{\text{maj}}(1-p_{\text{maj}})(\sigma_{\text{core}}^2 + 9)^2}, \frac{16 \exp\left(\frac{8}{(\sigma_{\text{core}}^2+8)\sigma_{\text{core}}^2} \right)(1 + 8/\sigma_{\text{core}}^2)}{p_{\text{maj}}(1-p_{\text{maj}})(\sigma_{\text{spu}}^2 + 1)} \right). \tag{96}
$$

For $\sigma_{\text{core}} \geq 1$, we have

$$
V \preceq \operatorname{diag}\left(\frac{C_1}{p_{\text{maj}}(1-p_{\text{maj}})}, \frac{C_2}{p_{\text{maj}}(1-p_{\text{maj}})} \right), \tag{97}
$$

for some constants C_1, C_2.

Proof. By asymptotic normality, we have $\sqrt{n}(\hat{w} - w^\star) \to \mathcal{N}(0, \nabla^2 L(w^\star)^{-1} \operatorname{Cov}[\nabla \ell(x, y, w^\star)] \nabla^2 L(w^\star)^{-1})$. Combining Lemma 12 and Lemma 13, we get the expression in Equation 96. Each term is decreasing in σ_{core}, and hence we get the final result by substituting $\sigma_{\text{core}}^2 = 1$ to obtain the constants C_1, C_2 (and noting that $\sigma_{\text{spu}}^2 \geq 0$). □