On Representation Learning Under Class Imbalance

Ravid Shwartz-Ziv[*]
New York University
ravid.shwartz.ziv@nyu.edu

Micah Goldblum[*]
New York University
goldblum@nyu.edu

Yucen Lily Li
New York University
yucenli@gmail.com

C. Bayan Bruss
Capital One
bayan.bruss@capitalone.com

Andrew Gordon Wilson
New York University
andrewgw@cims.nyu.edu

Abstract

Unlike carefully curated academic benchmarks, real-world datasets are often highly class-imbalanced, especially in safety-critical scenarios. Through extensive empirical investigation, we study a number of foundational learning behaviors for various models such as neural networks, gradient-boosted decision trees, and SVMs under class imbalance across a range of domains. Motivated by our observation that re-balancing class-imbalanced training data is ineffective, we show that several simple techniques for improving representation learning are effective in this setting: (1) self-supervised pre-training is insensitive to imbalance and can be used for feature learning before fine-tuning on labels; (2) Bayesian inference is effective because neural networks are especially underspecified under class imbalance; (3) flatness-seeking regularization pulls decision boundaries away from minority samples, especially when we seek minima that are particularly flat on the minority samples' loss.

1 Introduction

Data collection scenarios in real life include common and rare events. Machine learning systems are routinely trained and deployed on class-imbalanced data where relatively few samples are associated with minority classes. Nonetheless, the vast majority of works exclusively consider class balanced benchmarks [LeCun, 1998, Krizhevsky, 2009, Deng et al., 2009]. In this work, we explore various machine learning approaches of what makes learning under class imbalance so difficult and the associated implications for best practices in such scenarios. Many of the widely referenced methods for remedying class-imbalance problems rely on modifying how the training data is sampled [Chawla et al., 2002] and have been shown to be ineffective for neural networks [Buda et al., 2018]. To tease out exactly why oversampling is ineffective, we begin by studying the relationship between imbalances seen at train and test time, and we investigate whether poor generalization under class-imbalance can really be explained by failures of optimization. We find that while minority samples are hard to fit, this optimization phenomenon has little explanatory power regarding generalization. Furthermore, both rebalancing training data to include more minority samples as well as gathering more majority samples can negatively affect generalization. Following our investigation, we show why different methods are particularly well-suited in class-imbalanced settings; Self-supervised learning algorithms are less sensitive to the proportion of samples in various classes, so we can learn better feature representations before fine-tuning even on the same data. By looking on the singular values of the Hessian, we observe that neural networks trained on imbalanced datasets are significantly more

[*] Authors contributed equally.

ML Safety Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

underdetermined by the data. To overcome this problem, we show that Bayesian Neural Networks (BNNs), which can represent our uncertainty [Wilson and Izmailov, 2020, Shwartz-Ziv et al., 2022] achieve better results. Finally, while neural network decision boundaries tend to hug minority samples in order to expand the margins from majority data points which occur more frequently in training data, we can counteract this behavior with Sharpness-Aware Minimization (SAM) [Foret et al., 2020] which increase loss function flatness corresponding the minority samples. In summary, our work questions the motivation of sampling methods and proposes new directions to improve representation in class-imbalanced settings.

2 Experimental Setup

Class-imbalance ratio: The ratio between the number of samples in the rarest class to the number of samples in the most frequent class. In this paper, we will construct and investigate both *train* and *test* sets with varying imbalance ratios. **Datasets:** For experiments with neural networks, we use CIFAR-10 [Krizhevsky, 2009] as well as a binary variant in which we simply use two of the CIFAR-10 classes. For tabular data, we use the Adult dataset and Forest Cover dataset from the UCI Machine Learning Repository [Dua and Graff, 2017] **Models:** We use ResNet-34 [He et al., 2016] on CIFAR-10. We use XGBoost [Chen and Guestrin, 2016] and SVM on tabular datasets. For each evaluation in our experiments, we run five seeds and report the mean along with one standard error. Appendix B.1 contains additional details, and Appendix B.2 contains experiments on additional models and datasets.

3 The Role of Imbalanced Data in Generalization and Optimization

We investigate the impact of training set imbalances on generalization across various testing scenarios, and find that rebalancing training data actually harms generalization. We also found that fitting minority samples is difficult because of severe class imbalance, but fixing it doesn't help generalization.

3.1 The Relationship Between Train-Time and Test-Time Imbalance

In many cases, both training and test data are typically imbalanced. Therefore, we try to disentangle training and testing balances. To answer what is the optimal train set balance for a given test set, we train on datasets with a wide range of imbalance ratios and evaluate each trained model on a variety of testing ratios. We illustrate three scenarios in Figure 1 (left): (1) identical training and testing ratios, (2) balanced training, and (3) the training ratio with the lowest test error. We see that training on data with the same imbalance as testing data is typically superior.

To determine the optimal training distribution, we also check the best train ratio versus the best test ratio Figure 1 (middle), and found that it is best to train with a very similar ratio to that of the test dataset. Interestingly, when the optimal training distribution is not exactly the same as the testing distribution, it is very close to. Even in cases where the best ratio for training is more balanced, there is minimal difference in test error between the best ratio for training and the test-equivalent ratio for training.

3.2 When More Data Degrades Performance

In the previous section, we fixed the total number of training samples and saw that models perform best when trained on a similar data distribution to their testing data. However, in practice, a practitioner likely will not have precise control over the data they collect. Will collecting additional samples always help performance? If not, practitioners must be cautious when collecting new data that its balance matches their testing data or else the additional data could result in even worse models. Instead of fixing the total number of samples and varying their class ratio, we now fix the number of samples from the minority class and vary the number of others.

In Figure 1 (right), we see that increasing the number of samples from the majority class, initially boosts performance on a balanced test set. Nevertheless, the performance reaches an optimum before the ever increasing imbalance in training data eventually degrades test accuracy. Thus, adding training data can be helpful, even without considering the balance of the additional data, but if we add enough

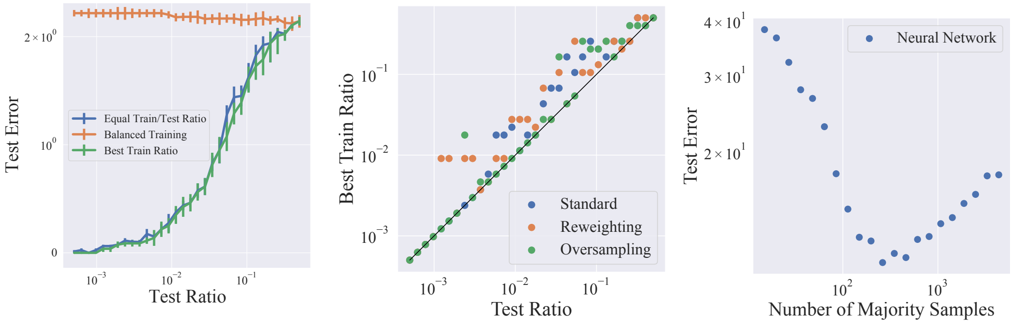


Figure 1: **The Role of Imbalanced Data in Generalization and Optimization** - CIFAR-10. **Left:** Test accuracy as a function of the test dataset ratio for different training setups . **Middle:** Optimal train imbalance ratio as a function of test imbalance ratio for various models. **Right:** The potentially destructive effects of adding majority class data.

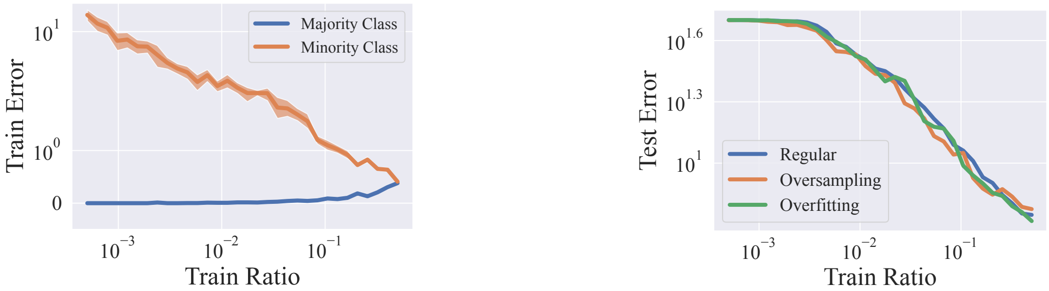


Figure 2: **Left:** Standard training routines fail to fit minority samples. **Right:** Overfitting with a low learning rate or oversampling does not improve neural network generalization - CIFAR-10

samples, we need to be careful not to cause too sharp a mismatch between training and testing distributions. Notably, the optimal training set ratio is nearly balanced, matching the test set, even when we are allowed to gather extra samples from one class without having to forego samples from another.

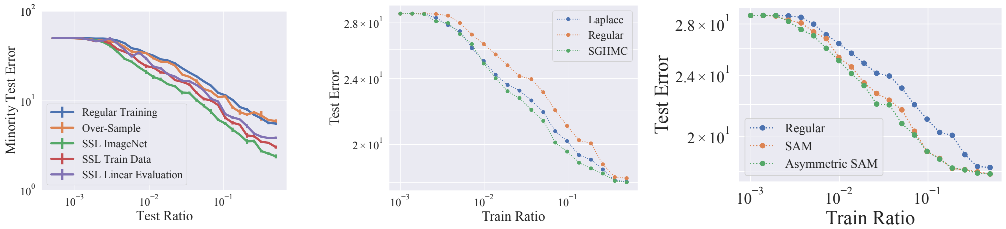
3.3 Minority Data is Hard to Fit, but Fitting It Does Not Help

Next, we examine optimization on imbalanced data. Our goal is to determine if poor optimization can explain low performance under class imbalance. In imbalanced training, the vast majority of the gradient signal during training comes from majority class samples, making it hard to fit minority samples [Figure 2 (left)]. Various methods are used to rebalance imbalanced training data artificially, and we confirm in Appendix B.2 that such sampling methods do indeed help fit minority samples. But is the inability to fit minority samples a culprit responsible for worse test accuracy?

We train our model using two additional methods: (1) oversampling the minority class, and (2) overfitting the training examples using a large number of epochs and a small learning rate. Both methods successfully fit minority and majority classes. Nonetheless, they fail to improve neural network test performance [Figure 2 (left)]. We include additional details and results for neural networks and XGBoost, as well as results on imbalanced test sets, in Appendix B.2.

4 Combining Self-Supervised Learning with Supervised Fine-Tuning for Robust Feature Learning

Self-supervised learning (SSL) pre-trained networks often exhibit high transferable representations [Grill et al., 2020]. However, many use-cases for deep learning are not accompanied by massive pre-training datasets. We thus propose a two-step procedure in which we first perform SSL pre-training and then supervised fine-tuning, all on the same imbalanced dataset. By first learning a feature extractor via SSL, which is insensitive to class imbalance, we can improve the quality of features and as a result generalization too. We use SimCLR [Chen et al., 2020] for self-supervised pre-training and try two fine-tuning routines: (1) fine-tune all layers in an end-to-end fashion and (2) train only a fully-connected layer on top of the fixed SSL feature extractor. We train on CIFAR-10 using a wide range of class-imbalance ratios reporting accuracy on a balanced test set, and we compare to three baselines including standard supervised learning, oversampling, and SimCLR ImageNet pre-training. In Figure 3a, we see that our two-step procedure which uses SSL pre-training on the small imbalanced CIFAR-10 dataset achieves almost as high performance as ImageNet pre-training and far superior performance to standard supervised learning and oversampling, even when the training set is relatively balanced.



(a) **Self-Supervised Training** (b) **Bayesian inference** (c) **SAM**

Figure 3: **Different techniques are effective for imbalance data**

5 Underspecification and Bayesian Inference

Expressive models such as neural networks are capable of representing numerous functions compatible with the training data. Bayesian Neural Networks approach this problem and it is particularly effective when the model is *underspecified* by the data [Wilson and Izmailov, 2020]. Several works have found that the number of high singular values of the loss function Hessian is equal to the number of classes [Sagun et al., 2017, Papyan, 2020]. We thus measure the 10 leading singular values on models trained on balanced or imbalanced CIFAR-10 data. We observe in Figure 10a that imbalanced data leads to *lower* singular values. We also perturb the parameter vector in random directions in Figure 10b using filter-normalization [Huang et al., 2019], and see that the loss increases slightly slower on imbalanced data, indicating a flatter minimum which indicates that there are many solutions, which are all consistent with the training samples. Next, by using Bayesian inference methods, we need not commit to only a single solution. We try two such Bayesian inference procedures; The Laplace approximation [Daxberger et al., 2021] and SGHMC [Chen et al., 2014]. We see in Figure 3b that such methods confer especially large boosts in accuracy on class-imbalanced training data with virtually no additional training cost.

6 Flatness-Seeking Regularization Pulls Decision Boundaries Away

Sharpness-Aware Minimization (SAM) [Foret et al., 2020] is an optimizer for finding flat minima of the loss function which often generalize better than those found by SGD. Huang et al. [2019] connect flat minima to wide margin decision boundaries. By plotting the decision boundaries of a small multi-layer perceptron on a toy 2D dataset in Figure 9, we see small margins surrounding minority class samples, and SAM expands these margins supporting the intuition of Huang et al. [2019]. Following this observation, we employ SAM on our CIFAR-10 setup and find that SAM especially improves generalization on class-imbalanced training data in Figure 3c. In light of this result, we now further increase flatness specifically on minority class loss terms by increasing the ascent step size in SAM's inner loop. We see in Figure 3c that this adaptation can yield even greater performance boosts.

7 Discussion

In this work, we examined the effects of class-imbalanced data on optimization and generalization. Following our above study, we suggest three existing methods for improving performance on class-imbalanced data without ad-hoc interventions specific to imbalance.

References

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. In *International conference on machine learning*, pages 1683–1691. PMLR, 2014.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless bayesian deep learning. *Advances in Neural Information Processing Systems*, 34:20089–20103, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2020.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

W Ronny Huang, Zeyad Emam, Micah Goldblum, Liam Fowl, Justin K Terry, Furong Huang, and Tom Goldstein. Understanding generalization through visualizations. *arXiv preprint arXiv:1906.03291*, 2019.

Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. 2009. URL `https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf`.

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Vardan Papyan. Traces of class/cross-class structure pervade deep learning spectra. *Journal of Machine Learning Research*, 21(252):1–64, 2020.

Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

Ravid Shwartz-Ziv, Micah Goldblum, Hossein Souri, Sanyam Kapoor, Chen Zhu, Yann LeCun, and Andrew Gordon Wilson. Pre-train your loss: Easy bayesian transfer learning with informative priors. *arXiv preprint arXiv:2205.10279*, 2022.

Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems*, 33:4697–4708, 2020.

A X-Risk Sheet

Individual question responses do not decisively imply relevance or irrelevance to existential risk reduction. Do not check a box if it is not applicable.

A.1 Long-Term Impact on Advanced AI Systems

In this section, please analyze how this work shapes the process that will lead to advanced AI systems and how it steers the process in a safer direction.

1. **Overview.** How is this work intended to reduce existential risks from advanced AI systems?
   Answer: By focusing on class-balanced benchmarks, the community is ignoring critical use-cases where data is highly imbalanced. Learning can fail altogether in such use-cases.
2. **Direct Effects.** If this work directly reduces existential risks, what are the main hazards, vulnerabilities, or failure modes that it directly affects?
   Answer: This work proposes learning pipelines which are robust to class imbalance and seeks to understand exactly why imbalance can cause failure. In doing so, we mitigate the risks associated with class imbalance.
3. **Diffuse Effects.** If this work reduces existential risks indirectly or diffusely, what are the main contributing factors that it affects?
   Answer: Our work proposes methods which are easy-to-use and do not require ad-hoc interventions. In doing so, we hope to make systems which are robust to class imbalance more accessible and expand their use.
4. **What's at Stake?** What is a future scenario in which this research direction could prevent the sudden, large-scale loss of life? If not applicable, what is a future scenario in which this research direction be highly beneficial?
   Answer: Safety-critical applications of ML are often accompanied by massive imbalance. For example, most luggage which passes through airport security does not contain bombs. The ability of our systems to accurately perform inference on minority samples, where mistakes can be fatal, is crucial for preventing loss of life.
5. **Result Fragility.** Do the findings rest on strong theoretical assumptions; are they not demonstrated using leading-edge tasks or models; or are the findings highly sensitive to hyperparameters?
   Answer: Our findings do rest on strong theoretical assumptions, and we conduct experiments on widely deployed models. We do only use small datasets, so we encourage future works to apply such methods on an industrial scale.
6. **Problem Difficulty.** Is it implausible that any practical system could ever markedly outperform humans at this task?
   Answer: We believe that our systems can already outperform humans on the tasks on which we conduct experiments.
7. **Human Unreliability.** Does this approach strongly depend on handcrafted features, expert supervision, or human reliability?
   Answer: No, it does not.

8. **Competitive Pressures.** Does work towards this approach strongly trade off against raw intelligence, other general capabilities, or economic utility?
   Answer: No. Improving performance on imbalanced data is aligned with general capabilities and therefore does not require such a trade-off.

A.2 Safety-Capabilities Balance

In this section, please analyze how this work relates to general capabilities and how it affects the balance between safety and hazards from general capabilities.

1. **Overview.** How does this improve safety more than it improves general capabilities?
   Answer: In our work, safety and general capabilities are aligned. By developing methods for improving learning on highly imbalanced data, we are improving model accuracy while also reducing risk posed by failure in this setting.
2. **Red Teaming.** What is a way in which this hastens general capabilities or the onset of x-risks?
   Answer: Improved algorithms on imbalanced data may be used for military purposes such as automatic detection in conflict areas, potentially enabling the use of automated weapons.
3. **General Tasks.** Does this work advance progress on tasks that have been previously considered the subject of usual capabilities research?
   Answer: Yes, since avoiding failure in learning under class imbalance may be viewed either in terms of reliability or usual capabilities.
4. **General Goals.** Does this improve or facilitate research towards general prediction, classification, state estimation, efficiency, scalability, generation, data compression, executing clear instructions, helpfulness, informativeness, reasoning, planning, researching, optimization, (self-)supervised learning, sequential decision making, recursive self-improvement, open-ended goals, models accessing the Internet, or similar capabilities?
   Answer: Yes
5. **Correlation With General Aptitude.** Is the analyzed capability known to be highly predicted by general cognitive ability or educational attainment?
6. **Safety via Capabilities.** Does this advance safety along with, or as a consequence of, advancing other capabilities or the study of AI?
   Answer: Yes, avoiding failures under class imbalance improves accuracy, but improving accuracy on minority samples has the effect of avoiding catastrophic risks associated with such failures.

A.3 Elaborations and Other Considerations

1. **Other.** What clarifications or uncertainties about this work and x-risk are worth mentioning?
   Answer: The focus of our work is on understanding the nature of learning under class imbalance, and we recommend simple remedies. However, in safety-critical scenarios, one should consider specially tailored solutions designed for class imbalance.

B Appendix

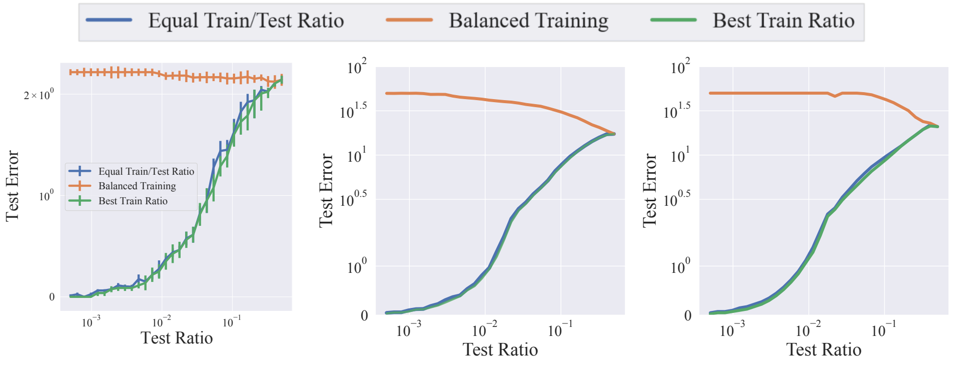
B.1 Model Details

XGBoost: ‘XGBClassifier’ from XGBoost version 1.6.2

- ‘n_estimators’ = 100
- ‘subsample’ = 0.5
- ‘eta’ = 0.3
- ‘max_depth’ = 6

SVM: ‘LinearSVM’ from sklearn version 1.1.2

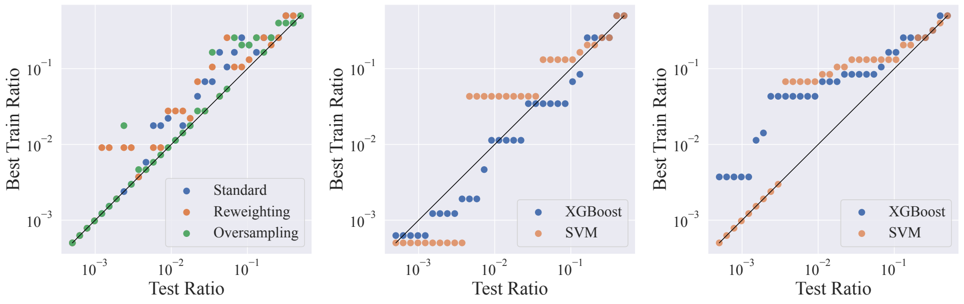
- ‘dual’ = false
- ‘max_iter’ = 1000



(a) CIFAR-10 Dataset (b) XGBoost on Adult Dataset (c) SVM on Forest Cover Dataset

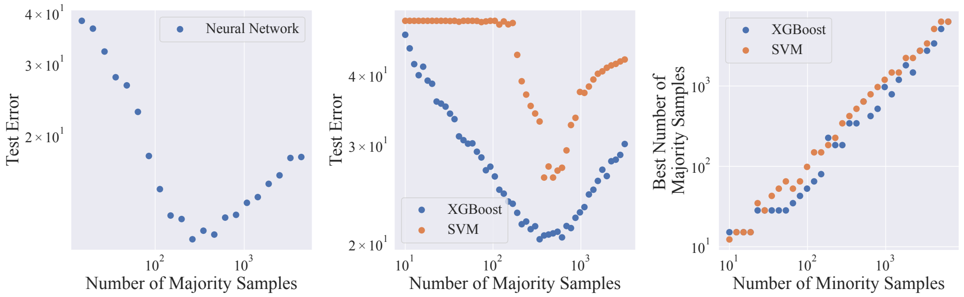
Figure 4: **Training on imbalanced data is optimal for imbalanced testing scenarios.** Test accuracy as a function of the test dataset ratio for different training setups. Error bars correspond to one standard error over 5 trials.

B.2 Supplemental Figures



(a) CIFAR-10 Dataset (b) Adult Dataset (c) Forest Cover Dataset

Figure 5: **The optimal train dataset ratio is very close to the test dataset ratio.** Optimal train imbalance ratio as a function of test imbalance ratio for various datasets and models.

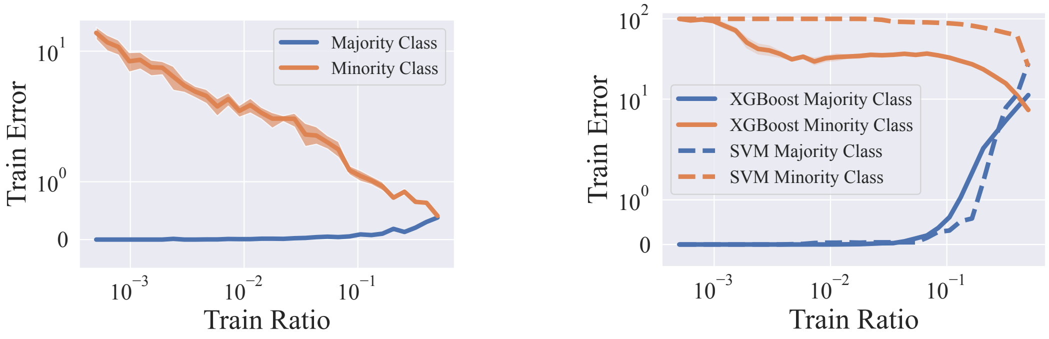


(a) Binary CIFAR-10

(b) Adult Dataset

(c) Adult Dataset

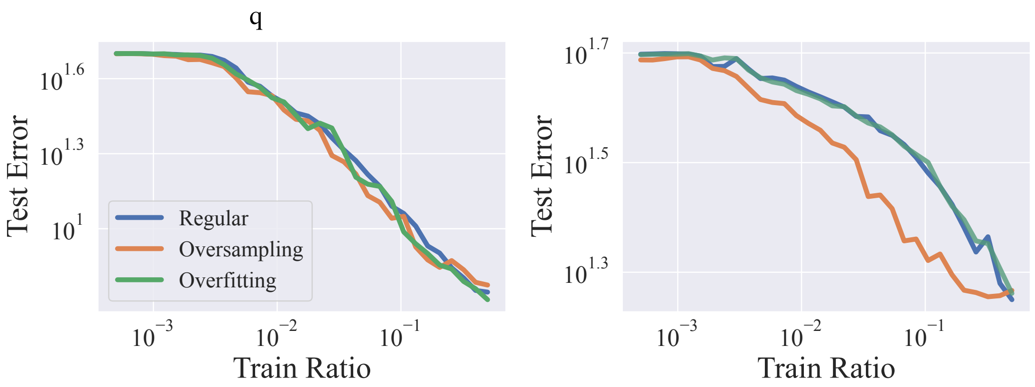
Figure 6: **The potentially destructive effects of adding majority class data**. In (a) and (b), we fix the number of minority samples to be 500 and vary the number of majority samples. In (c), we plot the number of majority samples that gives us the lowest test error against the number of minority samples. We see that increasing the number of majority samples degrades performance. Error reported on balanced test set.



(a) ResNet on CIFAR-10

(b) XGBoost and SVM on Adult

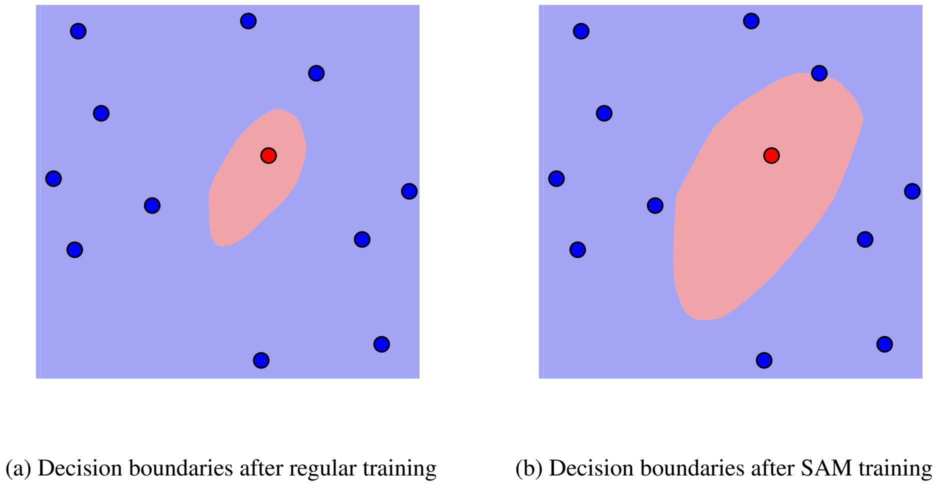
Figure 7: **Standard training routines fail to fit minority samples.** Error bars correspond to one standard error over 5 trials



(a) ResNet Test Error on Binary CIFAR-10

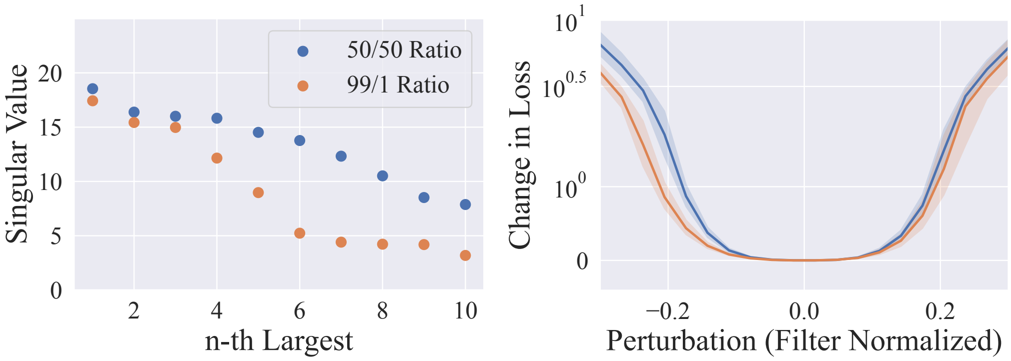
(b) XGBoost Test Error on Adult

Figure 8: **Overfitting with a low learning rate or oversampling does not improve neural network generalization**. Oversampling is helpful for XGBoost. Error reported on a balanced test set.


(a) Decision boundaries after regular training

(b) Decision boundaries after SAM training

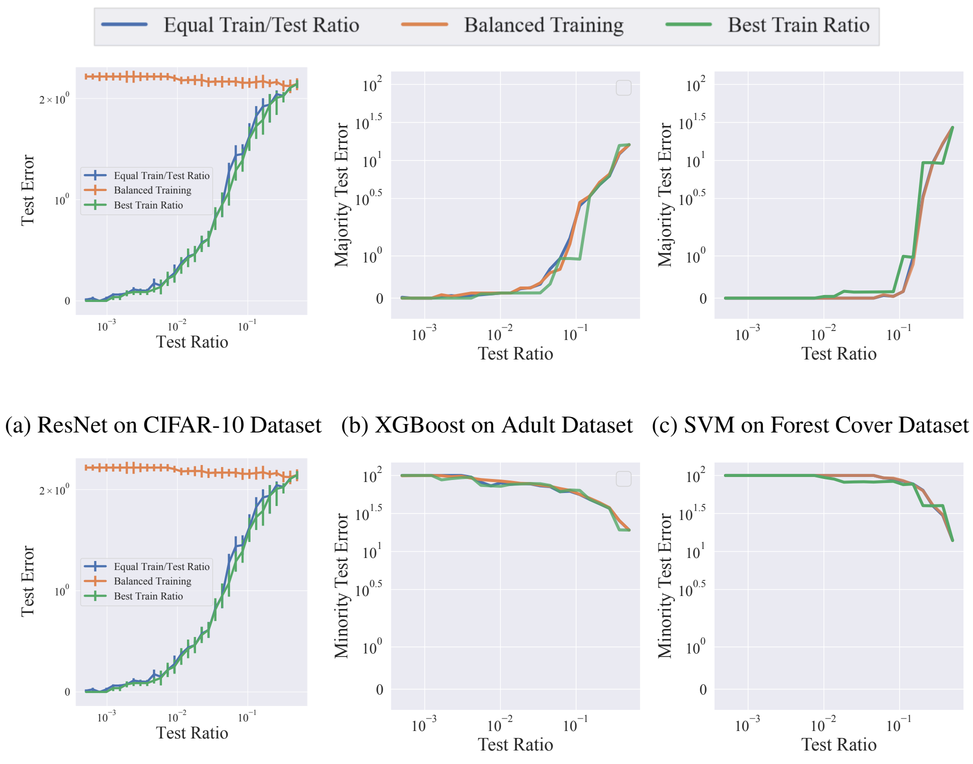
Figure 9: **Flatness seeking regularization pulls decision boundaries away from minority samples.** Experiments conducted on toy 2D dataset paired with MLP architecture.



(a) Ten largest singular values of the Hessian.

(b) Loss minima are flatter under imbalanced data.

Figure 10: **Loss minima are flatter under imbalanced data**. Models trained on CIFAR-10.



(a) ResNet on CIFAR-10 Dataset (b) XGBoost on Adult Dataset (c) SVM on Forest Cover Dataset

(d) ResNet on CIFAR-10 Dataset (e) XGBoost on Adult Dataset (f) SVM on Forest Cover Dataset

Figure 11: Test error split by majority and minority classes for balanced test sets. We see similar trends across all models and datasets.

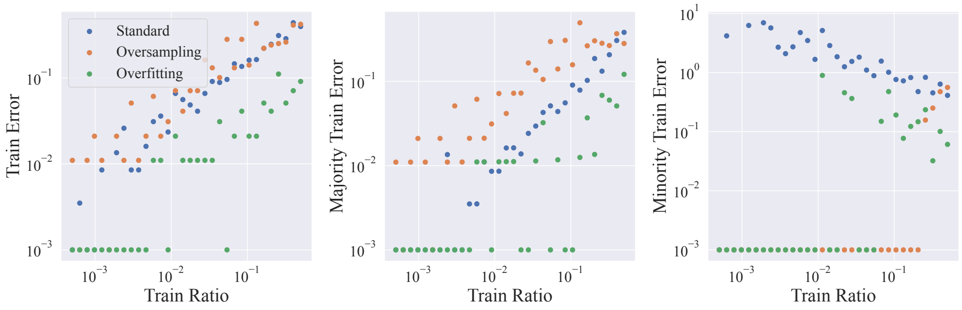


Figure 12: ResNet train error on CIFAR-10

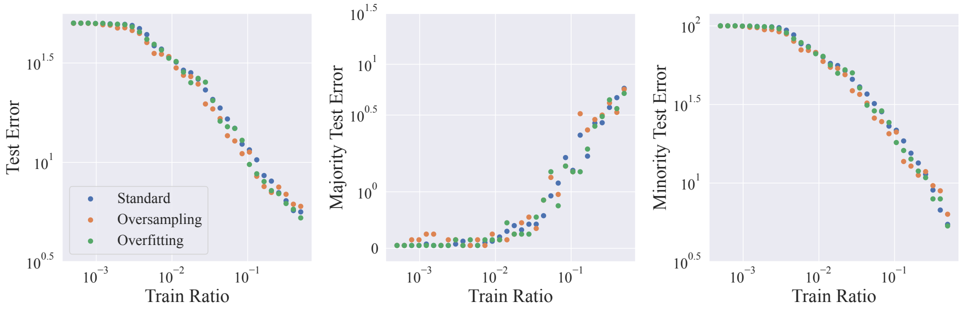


Figure 13: ResNet test error on imbalanced test sets from CIFAR-10

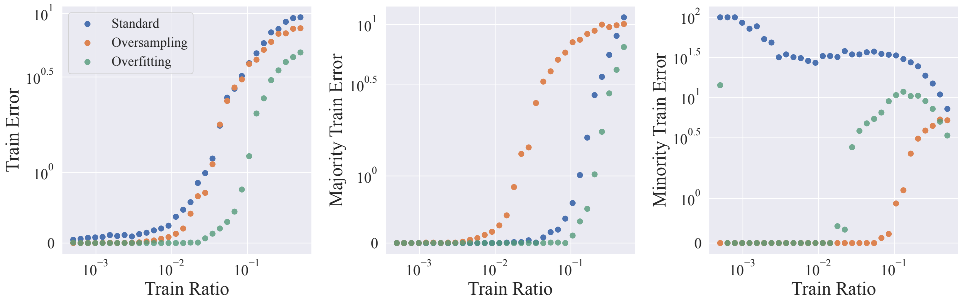


Figure 14: XGBoost train error on Adult

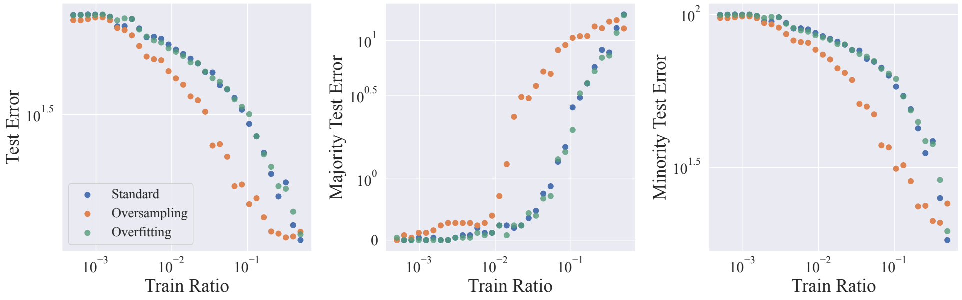


Figure 15: XGBoost test error on balanced test sets from Adult

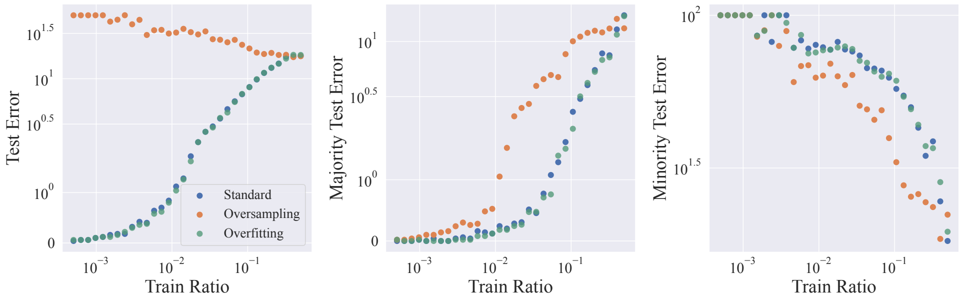


Figure 16: XGBoost test error on imbalanced test sets (train ratio and test ratio are equal) from Adult