# Catastrophic Failures of Neural Active Learning on Heteroskedastic Distributions

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Models which can actively seek out the best quality training data hold the promise of more accurate, adaptable, and efficient machine learning. State-of-the-art techniques tend to prefer examples which are the most difficult to classify. While this works well on homogeneous datasets, we find that it can lead to catastrophic failures when performing active learning on multiple distributions which have different degrees of label noise (heteroskedasticity). Most active learning algorithms strongly prefer to draw from the distribution with more noise, even if its examples have no informative structure (such as solid color images). We find that active learning which encourages diversity and model uncertainty in the selected examples can significantly mitigate these failures. We hope these observations are immediately useful to practitioners and can lead to the construction of more realistic and challenging active learning benchmarks.

## 1 Introduction

In an active learning setup, a model has access to a pool of labeled data and a pool of unlabeled data. After training on the available labeled data, some selection rule is applied to identify a batch of $k$ unlabeled samples to be labeled and integrated into the training set before repeating the process.Under this paradigm, data are considered to be abundant but label acquisition is costly. The goal of an active learning algorithm is to identify unlabeled samples that, once labeled at used to fit model parameters, will elicit the most performant hypothesis possible given a fixed labeling budget.

In an effort to fulfill this objective, high-quality selection criteria generally favor a diverse set of examples where the model has a high degree of uncertainty. That is, we want to favor the selection of items that might least to the most significant change from the model's current state, but we need to ensure we do not waste our labeling budget by selecting items that are similar to each other.

Preferring examples with high uncertainty often works well on homogeneous datasets, but can lead to catastrophic failure when training on a mixture of distinct distributions with different degrees of noise, as the active learning algorithm prefers the noisier distribution over the cleaner distribution. We refer to these as *Heteroskedastic Distributions*.

In this paper, we demonstrate that modern active learning algorithms designed for use with deep neural networks are significantly damaged when subjected to training data that's been corrupted by heteroskedastic noise. These approaches typically rely on notions of model improvement that are unable to disambiguate aleatoric from epistemic uncertainty, overselecting samples for which the model is unconfident but which are unlikely to best improve the current hypothesis. We show that this inefficiency can be partially reduced in two ways: by more heavily favor diversity and encouraging examples with high divergence between a conventionally trained model and an exponential moving average (EMA) of its iterates.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

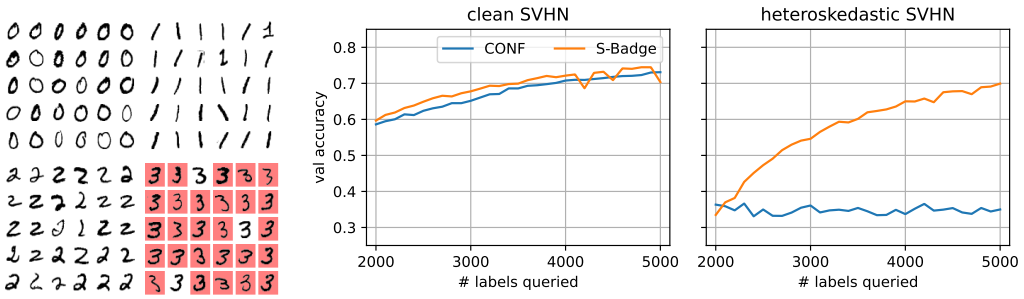


Figure 1: Overview of our setting and results. *Left:* We consider neural active learning settings with high hetereoskedasticity; as as extreme example, the labels from one class are replaced with uniform noise (incorrect labels in red). Active learning algorithms have a strong tendency to prefer this random-label class over other classes. *Right:* Our proposed S-Badge algorithm, which encourages both model uncertainty and diversity in sampled points, performs similarly to least-confidence sampling on the original SVHN data, but greatly outperforms it on the heteroskedastic distributions.

Empirically, with this EMA divergence, which we measure on the penultimate layer representation, seems to capture information about the model's sensitivity to that point. Importantly, we observe that this score is nearly zero on the noisy samples considered in this paper. When used to scale the representations used in for the Badge algorithm, we find we can significantly reduce the frequency with which noisy samples are selected, which we refer to as the S-Badge algorithm.

### 1.1 Related work

**Neural active learning.** Active learning is an extremely well-researched area, with the richest theory developed for the convex setting [1, 2]. More recently, however, there have been several attempts to tractably generalize active learning to the deep regime. Such approaches can be thought or as identifying batches of samples that cater more to either the model's predictive uncertainty or to the diversity of the selection.

In the former approach, a batch of points are selected in order of the model's uncertainty about their label. Many of these methods query samples that are nearest the decision boundary, an approach that's theoretically well understood in the linear regime when the batch size is 1 [3, 4, 5]. Some deep learning-specific approaches have also been developed, including using the variance of dropout samples to quantify uncertainty [6], and adversarial examples have been used to approximate the distance between an unlabeled sample and the decision boundary. In the deep setting however, where models are typically retrained from scratch after every round of selection, a larger batch size is usually necessary for efficiency purposes.

For large batch sizes, algorithms that cater to diversity are usually more effective. In deep learning, several methods take the representation obtained at the penultimate layer of the network, and aim to identify a batch of samples that might summarize this space well [7, 8]. [9]. Other methods promote diversity by minimizing an upper bound on some notion of model's loss on unseen data [10, 11, 12, 13, 14]. This approach has also been taken to trade-off between diversity and uncertainty in deep active learning [15, 16].

**Data poisoning and distributional robustness.** A related body of work seeks to obtain models and training procedures which are robust against *worst-case* perturbations to the data distribution. For recent treatments of this topic and further references, see [17, 18]. A few recent works have considered data poisoning in the active learning setting [19, 20], with defenses focusing on modifying the setting rather than the algorithm. Overall, though the settings are compatible, the aim of this work is to directly address the empirical performance of deep active learning with low-quality labels, rather than a more pessimistic min-max robustness formulation.

**Heteroskedasticity in deep learning.** The issues of class imbalance and heteroskedasticity are of interest in the supervised learning setting [21, 22, 23], in which various methods have been proposed

to make training more robust to these distributions. Our work seeks to initiate the study of the orthogonal (but analogous) issue in sample *selection*.

## 2 Heteroskedastic Benchmarks for Neural Active Learning

We introduce three new benchmarks for active learning on heteroskedastic distributions. In all cases, we introduce an additional set of $N$ examples with purely random labels, which the model is trained on with it shuffled into the original data. In all cases, whether an example is one of these special noisy samples is not given to the model, but it is always reasonably easily predictable from the example's features. This distinguishes our benchmarks from IID label noise, which is not predictable based on the example's features.

- *Noisy-Blank*: We introduce $N$ examples which are solid black ($x = 0$) with a random example $y \sim U(1, 10)$.
- *Noisy-Diverse*: We increase the difficulty by introducing $K = 100$ different types of examples, where each type is a random solid color and has a label randomly drawn from three different choices (unique to that type). $N$ of these examples are sampled. This benchmark is designed to make the heteroskedastic distribution more diverse while still keeping the noisy examples simple.
- *Noisy-Class*: In our hardest setting, we take the examples in the dataset with a particular class $y = 1$ and assign these examples uniformly random labels $y \sim U(1, 10)$. We then randomly repeat these examples to give $N$ examples. In this case, the randomly labeled examples are challenging but still possible to identify.

## 3 Methods and Experiments

Here we experiment with several active learning algorithms, noising strategies, and model architectures. In all experiments, we use the same experimental settings from [16], starting with 2000 points, and query samples in batches of 100 points, until we got to a total of 5000 labeled points. We avoid warm-starting and retrain from a fresh initialization after each round of selection [24]. We ran with both a small 1-layer MLP with 512 hidden units as well as a ResNet18. We also added the additional noisy-labeled examples in all cases to make 80% of the examples from the noisy distribution and 20% from the original distribution.

### 3.1 Baselines

We consider four baseline algorithms commonly used in the literature. Confidence sampling [25] and Margin sampling [26] are uncertainty-based strategies: Confidence sampling selects the $k$ unlabeled points for which the most likely label has the smallest amount of probability mass, and Margin sampling selects the $k$ points for which the difference in probability mass in the two most likely labels is smallest. The Coreset algorithm is a diversity-based approach that aims to select a batch of representative points, as measured in penultimate layer space of the current state of the model [7]. BADGE is a hybridized approach, meant to strike a balance between uncertainty and diversity. BADGE represents data in a hallucinated gradient space before performing diverse selection using k-means++ [16].

### 3.2 S-Badge: Increasing Sampling where Representations Change Across Training Iterations

We conjecture that while examples with noisy labels will have high loss and high predictive uncertainty, the model's predictions will converge quickly and undergo little change later in training. By encouraging the selection process in active learning to not select these examples, we hope to improve performance in the heteroskedastic setting.

In addition to the main (or online) model ($F_\theta$), we introduce an EMA model ($F_\beta$) into the example-selection pipeline. The EMA model has the same architecture as the online model but uses a different set of parameters $\beta$, which are exponentially moving averages of $\theta$. That is,

$$\beta = \alpha \cdot \beta + (1 - \alpha) \cdot \theta \tag{1}$$

Table 1: Classification accuracy on SVHN with 5000 actively queried examples, with different heteroskedastic distribution corruptions. In parenthesis, we report the percentage of the examples (over the course of training) which the active learning selected from the non-corrupted original examples (100% being best, and 0% being worst).

| ResNet | Clean | Noisy-Blank | | Noisy-Diverse | | Noisy-Class | |
|---|---|---|---|---|---|---|---|
| Random | 78.6% | 64.0% | (20.7%) | 50.8% | (19.6%) | 25.3% | (20.5%) |
| Confidence | 68.5% | 57.9% | (0.00%) | 55.9% | (42.4%) | 23.4% | (14.2%) |
| Margin | 73.6% | 72.9% | (82.6%) | 60.8% | (44.5%) | 30.6% | (15.9%) |
| Badge | 76.3% | 75.2% | (99.0%) | 57.0% | (29.2%) | 32.9% | (30.1%) |
| S-Badge | 76.0% | 74.7% | (99.0%) | 67.1% | (50.0%) | 30.1% | (36.9%) |
| MLP | | | | | | | |
| Random | 70.6% | 49.5% | (20.3%) | 46.5% | (20.2%) | 41.5% | (19.9%) |
| Confidence | 73.1% | 35.0% | (0.67%) | 36.2% | (12.9%) | 40.9% | (14.5%) |
| Margin | 75.0% | 65.2% | (83.8%) | 60.5% | (60.0%) | 46.7% | (19.4%) |
| Badge | 74.7% | 69.1% | (99.0%) | 48.7% | (19.8%) | 43.8% | (24.1%) |
| S-Badge | 70.3% | 70.0% | (99.0%) | 52.7% | (22.5%) | 47.7% | (28.2%) |

where $\alpha$ is set to a high value of 0.999.

We conjecture that the state difference between the EMA model and the online model can be a helpful signal for querying unlabelled examples. To this end, we re-weight the gradients in the Badge algorithm $\frac{d\mathcal{L}}{dW}$ by the average value of this hidden state difference before running K-means++ seeding. Where the hidden state difference is small, this reweighted gradient will be close to zero, and few of these examples are likely to be selected.

### 3.3 Results

The results in Table 1 show consistent deterioration in test accuracy when selecting from heteroskedastic distributions. We found that Badge and S-Badge, through their use of diversity in the selection process, were nearly perfect in solving the *Noisy-Blank* task. However on *Noisy-Diverse*, we found that S-Badge often significantly outperformed Badge. We also found that max-margin sampling was a surprisingly effective baseline compared to least-confidence sampling.

### 3.4 Analysis

One of our more striking experimental results (Table 1) is that selecting examples with the lowest prediction confidence can fail catastrophically on heteroskedastic distributions. In (Appendix A) we provide a theory explaining why training only on high loss examples (which would have low confidence given a well-calibrated model) can lead to poor performance on heteroskedastic distributions.

## 4 Conclusion

Neural Active Learning is an active area of research, with a plethora of new techniques competing to achieve better results. Our work seeks to throw this research program a curve ball, by showing that techniques which are competitive on homogeneous datasets with little label noise can fail catastrophically when presented with diverse heteroskedastic distributions. Active learning techniques which actively seek diversity fail less catastrophically, but still struggle when the noisy examples are themselves somewhat diverse. Despite some techniques failing less than others, our results suggest significant room for improvement and future research on this task.

# References


[1] Burr Settles. Active learning literature survey. *University of Wisconsin, Madison*, 2010.

[2] Sanjoy Dasgupta. Two faces of active learning. *Theoretical computer science*, 2011.

[3] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2001.

[4] Greg Schohn and David Cohn. Less is more: Active learning with support vector machines. In *International Conference on Machine Learning*, 2000.

[5] Gokhan Tur, Dilek Hakkani-Tür, and Robert E Schapire. Combining active and semi-supervised learning for spoken language understanding. *Speech Communication*, 2005.

[6] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, 2017.

[7] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

[8] Yonatan Geifman and Ran El-Yaniv. Deep active learning over the long tail. *arXiv:1711.00941*, 2017.

[9] Daniel Gissin and Shai Shalev-Shwartz. Discriminative active learning. *arXiv:1907.06347*, 2019.

[10] Yuhong Guo and Dale Schuurmans. Discriminative batch mode active learning. In *Neural Information Processing Systems*, 2008.

[11] Zheng Wang and Jieping Ye. Querying discriminative and representative samples for batch mode active learning. *Transactions on Knowledge Discovery from Data*, 2015.

[12] Yuxin Chen and Andreas Krause. Near-optimal batch mode active learning and adaptive submodular optimization. In *International Conference on Machine Learning*, 2013.

[13] Kai Wei, Rishabh Iyer, and Jeff Bilmes. Submodularity in data subset selection and active learning. In *International Conference on Machine Learning*, 2015.

[14] Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In *Advances in Neural Information Processing Systems*, 2019.

[15] Jordan T Ash, Surbhi Goel, Akshay Krishnamurthy, and Sham Kakade. Gone fishing: Neural active learning with fisher embeddings. *arXiv preprint arXiv:2106.09675*, 2021.

[16] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *International Conference on Learning Representations*, 2020.

[17] Jacob Steinhardt, Pang Wei Koh, and Percy Liang. Certified defenses for data poisoning attacks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 3520–3532, 2017.

[18] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2019.

[19] Jing Lin, Ryan Luley, and Kaiqi Xiong. Active learning under malicious mislabeling and poisoning attacks. *arXiv preprint arXiv:2101.00157*, 2021.

[20] Jose Rodrigo Sanchez Vicarte, Gang Wang, and Christopher W. Fletcher. Double-cross attacks: Subverting active learning systems. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 1593–1610. USENIX Association, August 2021.

[21] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *arXiv preprint arXiv:1906.07413*, 2019.

[22] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. *arXiv preprint arXiv:1902.07379*, 2019.

[23] Kaidi Cao, Yining Chen, Junwei Lu, Nikos Arechiga, Adrien Gaidon, and Tengyu Ma. Heteroskedastic and imbalanced deep learning with adaptive regularization. *arXiv preprint arXiv:2006.15766*, 2020.

[24] Jordan T Ash and Ryan P Adams. On warm-starting neural network training. *Advances in Neural Information Processing Systems*, 2020.

[25] Dan Wang and Yi Shang. A new active labeling method for deep learning. In *International Joint Conference on Neural Networks*, 2014.

[26] Dan Roth and Kevin Small. Margin-based active learning for structured output spaces. In *European Conference on Machine Learning*, 2006.

[27] Kenji Kawaguchi and Haihao Lu. Ordered sgd: A new stochastic optimization framework for empirical risk minimization. In *International Conference on Artificial Intelligence and Statistics*, pages 669–679, 2020.

[28] Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pages 297–299. PMLR, 2018.

[29] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[30] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012.

[31] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6240–6249, 2017.

[32] Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. Generalization in deep learning. *In Mathematics of Deep Learning, Cambridge University Press, to appear. Prepint available as: MIT-CSAIL-TR-2018-014, Massachusetts Institute of Technology*, 2018.

# A Generalization bound for biased query batches

## A.1 Notation

Let $\mathcal{D} = ((x_i, y_i))_{i=1}^n$ be a training dataset of $n$ samples where $x_i \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$ is the input vector and $y_i \in \mathcal{Y} \subseteq \mathbb{R}^{d_y}$ is the target output vector for the $i$-th sample. A standard objective function is

$$L(\theta; \mathcal{D}) := \frac{1}{n} \sum_{i=1}^{n} L_i(\theta; \mathcal{D}), \tag{2}$$

where $\theta \in \mathbb{R}^{d_\theta}$ is the parameter vector of the prediction model $f(\cdot\,; \theta) : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$, and $L_i(\theta; \mathcal{D}) := \ell(f(x_i; \theta), y_i)$ with the function $\ell : \mathbb{R}^{d_y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ is the loss of the $i$-th sample.

Similarly to the notation of order statistics, we first introduce the notation of ordered indexes: given a model parameter $\theta$, let $L_{(1)}(\theta; \mathcal{D}) \geq L_{(2)}(\theta; \mathcal{D}) \geq \cdots \geq L_{(n)}(\theta; \mathcal{D})$ be the decreasing values of the individual losses $L_1(\theta; \mathcal{D}), \ldots, L_n(\theta; \mathcal{D})$, where $(j) \in \{1, \ldots, n\}$ (for all $j \in \{1, \ldots, n\}$). That is, $\{(1), \ldots, (n)\}$ as a perturbation of $\{1, \ldots, n\}$ defines the order of sample indexes by loss values. Whenever we encounter ties on the values, we employ an arbitrary fixed tie-breaking rule in order to ensure the uniqueness of such an order.

Denote $r_i(\theta; \mathcal{D}) = \sum_{j=1}^n \mathbb{1}\{i = (j)\}\gamma_j$ where $(j)$ depends on $(\theta, \mathcal{D})$. Given an arbitrary set $\Theta \subseteq \mathbb{R}^{d_\theta}$, we define $\mathfrak{R}_n(\Theta)$ as the (standard) Rademacher complexity of the set $\{(x, y) \mapsto \ell(f(x; \theta), y) : \theta \in \Theta\}$:

$$\mathfrak{R}_n(\Theta) = \mathbb{E}_{\overline{\mathcal{D}}, \xi} \left[ \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \xi_i \ell(f(\bar{x}_i; \theta), \bar{y}_i) \right],$$

where $\overline{\mathcal{D}} = ((\bar{x}_i, \bar{y}_i))_{i=1}^n$, and $\xi_1, \ldots, \xi_n$ are independent uniform random variables taking values in $\{-1, 1\}$ (i.e., Rademacher variables). Given a tuple $(\ell, f, \Theta, \mathcal{X}, \mathcal{Y})$, define $M$ as the least upper bound on the difference of individual loss values: $|\ell(f(x; \theta), y) - \ell(f(x'; \theta), y')| \leq M$ for all $\theta \in \Theta$ and all $(x, y), (x', y') \in \mathcal{X} \times \mathcal{Y}$. For example, $M = 1$ if $\ell$ is the 0-1 loss function.

$$\hat{\mathfrak{R}}_n(\Theta) = \mathbb{E}_{\xi} \left[ \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \xi_i \ell(f(x_i; \theta), y_i) \right],$$

## A.2 Preliminaries

The previous paper [27] proves the following theoretical result. The stochastic optimization method that uses a gradient estimator that is purposely biased toward those samples with the current top-$q$ losses (i.e., ordered SGD) implicitly minimizes a new objective function of

$$L_q(\theta; \mathcal{D}) = \frac{1}{q} \sum_{j=1}^{n} \gamma_j L_{(j)}(\theta; \mathcal{D}),$$

for any $\mathcal{D}$ (including $g(\mathcal{D})$), in the sense that such a gradient estimator is an unbiased estimator of a (sub-) gradient of $L_q(\theta; \mathcal{D})$, instead of $L(\theta; \mathcal{D})$. Accordingly, the top-$q$-biased stochastic optimization method converges in terms of $L_q$ instead of $L$.

Building up on this result, we consider generalization properties of the top-$q$-biased stochastic optimization with the presence of additional label noises in training data. We want to minimize the expected loss,

$$\mathbb{E}_{(x,y)\sim\mathcal{P}}[\ell(f(x; \theta), y)]$$

by minimizing the training loss

$$L_q(\theta; g(\mathcal{D})),$$

where $g(\mathcal{D}) = ((g_i^x(x_i), g_i^y(y_i)))_{i=1}^n$ is potentially corrupted by arbitrary noise and corruption effects within arbitrary fixed functions $g_i^x$ and $g_i^y$ for $i = 1 \ldots, n$, where $(x_i, y_i) \sim \mathcal{P}$. Thus, we want to analyze the generalization gap:

$$\mathbb{E}_{(x,y)\sim\mathcal{P}}[\ell(f(x; \theta), y)] - L_q(\theta; g(\mathcal{D}))$$

The previous paper [27] showed the benefit of the top-$q$-biased stochastic optimization method in terms of generalization when $g_i^x$ and $g_i^y$ are identity functions and thus when the distributions are the same for both expected loss and training loss. In contrast, in our setting, the distributions are different for expected loss and training loss with potential noise corruptions through $g_i^x$ and $g_i^y$.

### A.3 Analysis

Theorem 1 presents a generalization bound for the top-$q$-biased stochastic optimization:

**Theorem 1.** *Let $\Theta$ be a fixed subset of $\mathbb{R}^{d_\theta}$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over an iid draw of $n$ examples $\mathcal{D} = ((x_i, y_i))_{i=1}^n$, the following holds for all $\theta \in \Theta$:*

$$\mathbb{E}_{(x,y)}[\ell(f(x;\theta),y)] \leq L_q(\theta; g(\mathcal{D})) + 2\hat{\mathfrak{R}}_n(\Theta) + M\left(2 + \frac{s}{q}\right)\sqrt{\frac{\ln(2/\delta)}{2n}} - \mathcal{Q}_{n,q}(\Theta, g), \quad (3)$$

*where* $\mathcal{Q}_{n,q}(\Theta, g) := \mathbb{E}_{\bar{\mathcal{D}}}[\inf_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n (\frac{r_i(\theta; g(\bar{\mathcal{D}}))n}{q}\ell(f(g_i^x(\bar{x}_i);\theta), g_i^y(\bar{y}_i)) - \ell(f(\bar{x}_i;\theta),\bar{y}_i))]$.

The expected error $\mathbb{E}_{(x,y)}[\ell(f(x;\theta),y)]$ in the left-hand side of Equation (3) is a standard objective for generalization, whereas the right-hand side contains the data corruption function $g$. Here, we typically have $\mathfrak{R}_n(\Theta) = O(1/\sqrt{n})$ in terms of $n$. For example, consider the standard feedforward deep neural networks of the form $f(x) = (\omega_T \circ \sigma_{T-1} \circ \omega_{T-1} \circ \sigma_{T-2} \cdots \sigma_1 \circ \omega_1)(x)$ where $T$ is the number of layers, $\omega_l(a) = W_l a$ with $\|W_l\|_F \leq M_l$, and $\sigma_l$ is an element-wise nonlinear activation function that is 1-Lipschitz and positive homogeneous (e.g., ReLU). Then, if $\|x\| \leq B$ for all $x \in \mathcal{X}$, using Theorem 1 of [28], we have that

$$\hat{\mathfrak{R}}_n(\Theta) \leq \frac{B(\sqrt{2\log(2)T}+1)(\prod_{l=1}^T M_l)}{\sqrt{n}}.$$

In Theorem 1, we can see that a label noise corruption $g$ can lead to the failure of the top-$q$-biased stochastic optimization via increasing the training loss $L_q(\theta; g(\mathcal{D}))$ and decreasing the top-$q$-biased factor $\mathcal{Q}_{n,q}(\Theta, g)$. Here, if there is no corruption $g$ (i.e., if $g_i^x$ and $g_i^y$ are identity functions), then we have that $\mathcal{Q}_{n,q}(\Theta, g) \geq 0$ because $\mathcal{Q}_{n,q}(\Theta, g) = \mathbb{E}_{\bar{\mathcal{D}}}[\inf_{\theta\in\Theta} L_q(\theta;\bar{\mathcal{D}}) - L(\theta;\bar{\mathcal{D}})] \geq 0$ due to $L_q(\theta;\bar{\mathcal{D}}) - L(\theta;\bar{\mathcal{D}}) \geq 0$ for any $\theta$ and $\bar{\mathcal{D}}$ when $g_i^x$ and $g_i^y$ are identity functions. Thus, the top-$q$-biased factor $\mathcal{Q}_{n,q}(\Theta, g)$ can explain the improvement of the generalization of the top-$q$-biased stochastic optimization over the standard unbiased stochastic optimization. However, with the presence of the corruption $g$, $\frac{r_i(\theta; g(\bar{\mathcal{D}}))n}{q}\ell(f(g_i^x(\bar{x}_i);\theta), g_i^y(\bar{y}_i))$ can be smaller than $\ell(f(\bar{x}_i;\theta),\bar{y}_i)$ *by fitting the corrupted noise*, resulting $\mathcal{Q}_{n,q}(\Theta, g) < 0$. This leads to a significant failure in the following sense: the generalization gap $(\mathbb{E}_{(x,y)}[\ell(f(x;\theta),y)] - L_q(\theta; g(\mathcal{D})))$ goes to zero as $n$ approach infinity if $\mathcal{Q}_{n,q}(\Theta, g) \geq 0$ with no data corruption, but the generalization gap no longer goes to zero as as $n$ approach infinity if $\mathcal{Q}_{n,q}(\Theta, g) < 0$ with data corruption.

To see this, let us look at the asymptotic case when $n \to \infty$. Let $\Theta$ be constrained such that $\mathfrak{R}_n(\Theta) \to 0$ as $n \to \infty$, which has been shown to be satisfied for various models and sets $\Theta$, including the standard deep neural networks above [29, 30, 31, 32, 28]. The third term in the right-hand side of Equation (3) disappear as $n \to \infty$. Thus, if there is no corruption (i.e., if $g_i^x$ and $g_i^y$ are identity functions), it holds with high probability that $\mathbb{E}_{(x,y)}[\ell(f(x;\theta),y)] \leq L_q(\theta; g(\mathcal{D})) - \mathcal{Q}_{n,q}(\Theta, g) \leq L_q(\theta; g(\mathcal{D}))$, where $L_q(\theta; g(\mathcal{D}))$ is minimized by the top-$q$-biased stochastic optimization. From this viewpoint, the top-$q$-biased stochastic optimization minimizes the expected error for generalization when $n \to \infty$, if there is no corruption. However, if there is corruption, $\mathbb{E}_{(x,y)}[\ell(f(x;\theta),y)] \leq L_q(\theta; g(\mathcal{D})) - \mathcal{Q}_{n,q}(\Theta, g) \not\leq L_q(\theta; g(\mathcal{D}))$, and hence $\mathbb{E}_{(x,y)}[\ell(f(x;\theta),y)] - L_q(\theta; g(\mathcal{D})) \not\to 0$ even in the asymptotic case.

### A.4 Proof of Theorem 1

We first notice that the following proposition from [27] still holds with the corrupted data with the same proof $g(\mathcal{D})$:

**Proposition 1.** *For any $j \in \{1, \ldots, n\}$, $\gamma_j \leq \frac{s}{n}$.*

We use this proposition in the following proof of Theorem 1 to bound the effect of replacing one sample in a dataset.

*Proof of Theorem 1.* We find an upper bound on $\sup_{\theta\in\Theta} \mathbb{E}_{(x,y)}[\ell(f(x;\theta),y)] - L_q(\theta; g(\mathcal{D}))$ based on McDiarmid's inequality. Define

$$\Phi(\mathcal{D}) = \sup_{\theta\in\Theta} \mathbb{E}_{(x,y)}[\ell(f(x;\theta),y)] - L_q(\theta; g(\mathcal{D})).$$

Our proof plan is to provide the upper bound on $\Phi(\mathcal{D})$ by using McDiarmid's inequality. To apply McDiarmid's inequality to $\Phi(\mathcal{D})$, we first show that $\Phi(\mathcal{D})$ satisfies the remaining condition of McDiarmid's inequality on the effect of changing one sample. Let $\mathcal{D}$ and $\mathcal{D}'$ be two datasets differing by exactly one point of an arbitrary index $i_0$; i.e., $\mathcal{D}_i = \mathcal{D}'_i$ for all $i \neq i_0$ and $\mathcal{D}_{i_0} \neq \mathcal{D}'_{i_0}$. Since $(j)$ depends on $g(\mathcal{D})$, we sometimes write $(j;\mathcal{D}) = (j)$ to stress the dependence on $\mathcal{D}$ under $g$. Then, we provide an upper bound on $\Phi(\mathcal{D}') - \Phi(\mathcal{D})$ as follows:

$$\begin{aligned}
\Phi(\mathcal{D}') - \Phi(\mathcal{D}) &\leq \sup_{\theta\in\Theta} L_q(\theta; g(\mathcal{D})) - L_q(\theta; g(\mathcal{D}')). \\
&= \sup_{\theta\in\Theta} \frac{1}{q} \sum_{j=1}^{n} \gamma_j (L_{(j;\mathcal{D})}(\theta; g(\mathcal{D})) - L_{(j;\mathcal{D}')}(\theta; g(\mathcal{D}'))) \\
&\leq \sup_{\theta\in\Theta} \frac{1}{q} \sum_{j=1}^{n} |\gamma_j| |L_{(j;\mathcal{D})}(\theta; g(\mathcal{D})) - L_{(j;\mathcal{D}')}(\theta; g(\mathcal{D}'))| \\
&\leq \sup_{\theta\in\Theta} \frac{1}{q} \frac{s}{n} \sum_{j=1}^{n} |L_{(j;\mathcal{D})}(\theta; g(\mathcal{D})) - L_{(j;\mathcal{D}')}(\theta; g(\mathcal{D}'))|
\end{aligned}$$

where the first line follows the property of the supremum, $\sup(a) - \sup(b) \leq \sup(a-b)$, the second line follows the definition of $L_q$ where $(j;\mathcal{D}) \neq (j;\mathcal{D}')$, and the last line follows Proposition 1 ($|\gamma_j| \leq \frac{s}{n}$).

We now bound the last term $\sum_{j=1}^{n} |L_{(j;\mathcal{D})}(\theta; g(\mathcal{D})) - L_{(j;\mathcal{D}')}(\theta; g(\mathcal{D}'))|$. This requires a careful examination because $|L_{(j;\mathcal{D})}(\theta; g(\mathcal{D})) - L_{(j;\mathcal{D}')}(\theta; g(\mathcal{D}'))| \neq 0$ for more than one index $j$ (although $\mathcal{D}$ and $\mathcal{D}'$ differ only by exactly one point). This is because it is possible to have $(j;\mathcal{D}) \neq (j;\mathcal{D}')$ for many indexes $j$ where $(j;\mathcal{D})$ in $L_{(j;\mathcal{D})}(\theta; g(\mathcal{D}))$ and $(j;\mathcal{D}')$ in $L_{(j;\mathcal{D}')}(\theta; g(\mathcal{D}'))$. To analyze this effect, we now conduct case analysis. Define $l(i;\mathcal{D})$ such that $(j) = i$ where $j = l(i;\mathcal{D})$; i.e., $L_i(\theta; g(\mathcal{D})) = L_{(l(i;\mathcal{D}))}(\theta; g(\mathcal{D}))$.

Consider the case where $l(i_0;\mathcal{D}') \geq l(i_0;\mathcal{D})$. Let $j_1 = l(i_0;\mathcal{D})$ and $j_2 = l(i_0;\mathcal{D}')$. Then,

$$\begin{aligned}
&\sum_{j=1}^{n} |L_{(j)}(\theta; g(\mathcal{D})) - L_{(j)}(\theta; g(\mathcal{D}'))| \\
&= \sum_{j=j_1}^{j_2-1} |L_{(j)}(\theta; g(\mathcal{D})) - L_{(j)}(\theta; g(\mathcal{D}'))| + |L_{(j_2)}(\theta; g(\mathcal{D})) - L_{(j_2)}(\theta; g(\mathcal{D}'))| \\
&= \sum_{j=j_1}^{j_2-1} |L_{(j)}(\theta; g(\mathcal{D})) - L_{(j+1)}(\theta; g(\mathcal{D}))| + |L_{(j_2)}(\theta; g(\mathcal{D})) - L_{(j_2)}(\theta; g(\mathcal{D}'))| \\
&= \sum_{j=j_1}^{j_2-1} (L_{(j)}(\theta; g(\mathcal{D})) - L_{(j+1)}(\theta; g(\mathcal{D}))) + L_{(j_2)}(\theta; g(\mathcal{D})) - L_{(j_2)}(\theta; g(\mathcal{D}')) \\
&= L_{(j_1)}(\theta; g(\mathcal{D})) - L_{(j_2)}(\theta; g(\mathcal{D}')) \\
&\leq M,
\end{aligned}$$

where the first line uses the fact that $j_2 = l(i_0;\mathcal{D}') \geq l(i_0;\mathcal{D}) = j_1$ where $i_0$ is the index of samples differing in $\mathcal{D}$ and $\mathcal{D}'$. The second line follows the equality $(j;\mathcal{D}') = (j+1;\mathcal{D})$ from $j_1$ to $j_2 - 1$ in this case. The third line follows the definition of the ordering of the indexes. The fourth line follows the cancellations of the terms from the third line.

Consider the case where $l(i_0;\mathcal{D}') < l(i_0;\mathcal{D})$. Let $j_1 = l(i_0;\mathcal{D}')$ and $j_2 = l(i_0;\mathcal{D})$. Then,

$$\sum_{j=1}^{n} |L_{(j)}(\theta; g(\mathcal{D})) - L_{(j)}(\theta; g(\mathcal{D}'))|$$

$$
\begin{aligned}
&= |L_{(j_1)}(\theta; g(\mathcal{D})) - L_{(j_1)}(\theta; g(\mathcal{D}'))| + \sum_{j=j_1+1}^{j_2} |L_{(j)}(\theta; g(\mathcal{D})) - L_{(j)}(\theta; g(\mathcal{D}'))| \\
&= |L_{(j_1)}(\theta; g(\mathcal{D})) - L_{(j_1)}(\theta; g(\mathcal{D}'))| + \sum_{j=j_1+1}^{j_2} |L_{(j)}(\theta; g(\mathcal{D})) - L_{(j-1)}(\theta; g(\mathcal{D}))| \\
&= L_{(j_1)}(\theta; g(\mathcal{D})) - L_{(j_1)}(\theta; g(\mathcal{D}')) + \sum_{j=j_1+1}^{j_2} (L_{(j)}(\theta; g(\mathcal{D})) - L_{(j-1)}(\theta; g(\mathcal{D}))) \\
&= L_{(j_1)}(\theta; g(\mathcal{D}')) - L_{(j_2)}(\theta; g(\mathcal{D})) \\
&\leq M.
\end{aligned}
$$

where the first line uses the fact that $j_1 = l(i_0; \mathcal{D}') < l(i_0; \mathcal{D}) = j_2$ where $i_0$ is the index of samples differing in $\mathcal{D}$ and $\mathcal{D}'$. The second line follows the equality $(j; \mathcal{D}') = (j-1; \mathcal{D})$ from $j_1 + 1$ to $j_2$ in this case. The third line follows the definition of the ordering of the indexes. The fourth line follows the cancellations of the terms from the third line.

Therefore, in both cases of $l(i_0; \mathcal{D}') \geq l(i_0; \mathcal{D})$ and $l(i_0; \mathcal{D}') < l(i_0; \mathcal{D})$, we have that

$$
\Phi(\mathcal{D}') - \Phi(\mathcal{D}) \leq \frac{s}{q}\frac{M}{n}.
$$

Similarly, $\Phi(\mathcal{D}) - \Phi(\mathcal{D}') \leq \frac{s}{q}\frac{M}{n}$, and hence $|\Phi(\mathcal{D}) - \Phi(\mathcal{D}')| \leq \frac{s}{q}\frac{M}{n}$. Thus, by McDiarmid's inequality, for any $\delta > 0$, with probability at least $1 - \delta$,

$$
\Phi(\mathcal{D}) \leq \mathbb{E}_{\bar{\mathcal{D}}}[\Phi(\bar{\mathcal{D}})] + \frac{Ms}{q}\sqrt{\frac{\ln(1/\delta)}{2n}}.
$$

Moreover, since

$$
\sum_{i=1}^{n} r_i(\theta; g(\mathcal{D}))L_i(\theta; g(\mathcal{D})) = \sum_{j=1}^{n} \gamma_j \sum_{i=1}^{n} \mathbb{1}\{i = (j; \mathcal{D})\}L_i(\theta; g(\mathcal{D})) = \sum_{j=1}^{n} \gamma_j L_{(j)}(\theta; g(\mathcal{D})),
$$

we have that

$$
L_q(\theta; g(\mathcal{D})) = \frac{1}{q}\sum_{i=1}^{n} r_i(\theta; g(\mathcal{D}))L_i(\theta; g(\mathcal{D})).
$$

Therefore,

$$
\begin{aligned}
&\mathbb{E}_{\bar{\mathcal{D}}}[\Phi(\bar{\mathcal{D}})] \\
&= \mathbb{E}_{\bar{\mathcal{D}}}\left[\sup_{\theta\in\Theta} \mathbb{E}_{(\bar{x}', \bar{y}')}[\ell(f(\bar{x}'; \theta), \bar{y}')] - L(\theta; \bar{\mathcal{D}}) + L(\theta; \bar{\mathcal{D}}) - L_q(\theta; g(\bar{\mathcal{D}}))\right] \\
&\leq \mathbb{E}_{\bar{\mathcal{D}}}\left[\sup_{\theta\in\Theta} \mathbb{E}_{(\bar{x}', \bar{y}')}[\ell(f(\bar{x}'; \theta), \bar{y}')] - L(\theta; \bar{\mathcal{D}})\right] - \mathcal{Q}_{n,q}(\Theta, g) \\
&\leq \mathbb{E}_{\bar{\mathcal{D}}, \bar{\mathcal{D}}'}\left[\sup_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n}(\ell(f(\bar{x}_i'; \theta), \bar{y}_i') - \ell(f(\bar{x}_i; \theta), \bar{y}_i))\right] - \mathcal{Q}_{n,q}(\Theta, g) \\
&\leq \mathbb{E}_{\xi, \bar{\mathcal{D}}, \bar{\mathcal{D}}'}\left[\sup_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n}\xi_i(\ell(f(\bar{x}_i'; \theta), \bar{y}_i') - \ell(f(\bar{x}_i; \theta), \bar{y}_i))\right] - \mathcal{Q}_{n,q}(\Theta, g) \\
&\leq 2\mathfrak{R}_n(\Theta) - \mathcal{Q}_{n,q}(\Theta, g).
\end{aligned}
$$

where the third line and the last line follow the subadditivity of supremum, the forth line follows the Jensen's inequality and the convexity of the supremum, the fifth line follows that for each $\xi_i \in \{-1, +1\}$, the distribution of each term $\xi_i(\ell(f(\bar{x}_i'; \theta), \bar{y}_i') - \ell(f(\bar{x}_i; \theta), \bar{y}_i))$ is the distribution of $(\ell(f(\bar{x}_i'; \theta), \bar{y}_i') - \ell(f(\bar{x}_i; \theta), \bar{y}_i))$ since $\bar{\mathcal{D}}$ and $\bar{\mathcal{D}}'$ are drawn iid with the same distribution. Therefore, for any $\delta > 0$, with probability at least $1 - \delta$,

$$
\Phi(\mathcal{D}) \leq 2\mathfrak{R}_n(\Theta) - \mathcal{Q}_{n,q}(\Theta, g) + \frac{Ms}{q}\sqrt{\frac{\ln(1/\delta)}{2n}}.
$$

Finally, since changing one data point in $\mathcal{D}$ changes $\hat{\mathfrak{R}}_n(\Theta)$ by at most $M/m$, McDiarmid's inequality implies that for any $\delta > 0$, with probability at least $1 - \delta$,

$$\mathfrak{R}_n(\Theta) \leq \hat{\mathfrak{R}}_n(\Theta) + M\sqrt{\frac{\ln(1/\delta)}{2n}}.$$

By taking union bound, we obtain the statement of this theorem.

□