Proceedings of Machine Learning Research vol 195:1–23, 2023 36th Annual Conference on Learning Theory

Distribution-Independent Regression for Generalized Linear Models with Oblivious Corruptions

Ilias Diakonikolas ILIAS@CS.WISC.EDU
University of Wisconsin-Madison

Sushrut Karmalkar SKARMALKAR@WISC.EDU
University of Wisconsin-Madison

Jongho Park JONGHO.PARK@WISC.EDU
KRAFTON Inc.

Christos Tzamos CTZAMOS@GMAIL.COM
University of Wisconsin-Madison[*]

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We demonstrate the first algorithms for the problem of regression for generalized linear models (GLMs) in the presence of additive oblivious noise. We assume we have sample access to examples (x, y) where y is a noisy measurement of $g(w^* \cdot x)$. In particular, $y = g(w^* \cdot x) + \xi + \epsilon$ where ξ is the oblivious noise drawn independently of x, satisfying $\Pr[\xi = 0] \geq o(1)$, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Our goal is to accurately recover a function $g(w \cdot x)$ with arbitrarily small error when compared to the true values $g(w^* \cdot x)$, rather than the noisy measurements y.

We present an algorithm that tackles the problem in its most general distribution-independent setting, where the solution may not be identifiable. The algorithm is designed to return the solution if it is identifiable, and otherwise return a small list of candidates, one of which is close to the true solution. Furthermore, we characterize a necessary and sufficient condition for identifiability, which holds in broad settings. The problem is identifiable when the quantile at which $\xi + \epsilon = 0$ is known, or when the family of hypotheses does not contain candidates that are nearly equal to a translated $g(w^* \cdot x) + A$ for some real number A, while also having large error when compared to $g(w^* \cdot x)$.

This is the first result for GLM regression which can handle more than half the samples being arbitrarily corrupted. Prior work focused largely on the setting of linear regression with oblivious noise, and giving algorithms under more restrictive assumptions.

Keywords: Oblivious noise, Regression, Generalized Linear Models

1. Introduction

Learning neural networks is a fundamental challenge in machine learning with various practical applications. Generalized Linear Models (GLMs) are the most fundamental building blocks of larger neural networks. These correspond to a linear function $w^* \cdot x$ composed with a (typically non-linear) activation function $g(\cdot)$. The problem of learning GLMs has received extensive attention in the past, especially for the case of ReLU activations. The simplest scenario is the "realizable setting", i.e., when the labels exactly match the target function, and can be solved efficiently with practical algorithms, such as gradient descent (see, e.g., Soltanolkotabi (2017)). In many real-world settings, noise comes from various sources, ranging from rare events and mistakes to skewed and corrupted

[*] The names of the authors are arranged in alphabetical order.

measurements, making even simple regression problems computationally challenging. In contrast to the realizable setting, when even a small amount of data is adversarially labeled, computational hardness results are known even for approximate recovery (Hardt and Moitra, 2013; Manurangsi and Reichman, 2018; Diakonikolas et al., 2022a) and under well-behaved distributions (Goel et al., 2019; Diakonikolas et al., 2020b; Goel et al., 2020; Diakonikolas et al., 2021a, 2023). To investigate more realistic noise models, Chen et al. (2020b) and Diakonikolas et al. (2021b) study linear and ReLU regression in the Massart noise model, where an adversary has access to a *random* subset of *at most half* the samples and can perturb the labels arbitrarily after observing the uncorrupted samples. By tackling regression in an intermediate ("semi-random") noise model — lying between the clean realizable and the adversarially labeled models — these works recover w^* under only mild assumptions on the distribution. Interestingly, without any distributional assumptions, computational limitations have recently been established even in the Massart noise model (Diakonikolas and Kane, 2022; Nasser and Tiegel, 2022; Diakonikolas et al., 2022c,b).

In this paper, we consider the problem of GLM regression under the *oblivious noise model* (see Definition 1), which is another intermediate model that allows the adversary to corrupt almost all the labels yet limits their capability by requiring the oblivious noise be determined independently of the samples. The only assumption on this additive and independent noise is that it takes the value 0 with *vanishingly small* probability $\alpha > 0$. The oblivious noise model is a strong noise model that (information-theoretically) allows for *arbitrarily accurate* recovery of the target function. This stands in stark contrast to Massart noise, where it is impossible to recover the target function if more than half of the labels are corrupted. On the other hand, oblivious noise allows for recovery even when noise overwhelms, i.e., as $\alpha \to 0$.

We formally define the problem of learning GLMs in the presence of additive oblivious noise below. As is the case with prior work on GLM regression (see, e.g., Kakade et al. (2011)), we make the standard assumptions that the data distribution is supported in the unit ball (i.e., $\|x\|_2 \leq 1$) and that the parameter space of weight vectors is bounded (i.e, $\|w^*\|_2 \leq R$).

Definition 1 (GLM-Regression with Oblivious Noise) *We say that $(x, y) \sim$ GLM-Ob(g, σ, w^*) if $x \in \mathbb{R}^d$ is drawn from some distribution supported in the unit ball and $y = g(w^* \cdot x) + \xi + \epsilon$, where ϵ and ξ are drawn independently of x and satisfy $\Pr[\xi = 0] \geq \alpha = o(1)$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. We assume that $\|w^*\|_2 \leq R$ and that $g(\cdot)$ is 1-Lipschitz and monotonically non-decreasing.*

In recent years, there has been increased focus on the problem of linear regression in the presence of oblivious noise (Pesme and Flammarion, 2020; Dalalyan and Thompson, 2019; Suggala et al., 2019; Tsakonas et al., 2014; Bhatia et al., 2015). This line of work has culminated in consistent estimators when the fraction of clean data is $\alpha = d^{-c}$, where c is a small constant (d'Orsi et al., 2021b). In addition to linear regression, the oblivious noise model has also been studied for the problems of PCA, sparse recovery (Pesme and Flammarion, 2020; d'Orsi et al., 2021a), and in the online setting (Dalalyan and Thompson, 2019). See Section 1.3 for a detailed summary of related work.

However, prior algorithms and analyses often contain somewhat restrictive assumptions and exploit symmetry that only arises for the special case of linear functions. In this work, we address the following shortcomings of previous work:

1. **Assumptions on ξ and marginal distribution**: Prior work either assumed that the oblivious noise was symmetric or made strong distributional assumptions on the x's, such as mean-zero

Gaussian or sub-Gaussian tails. We allow the distribution to be arbitrary (while being supported on the unit ball) and make no additional assumptions on the oblivious noise.

2. **Linear functions**: One useful technique to center an instance of the problem for linear functions is to take pairwise differences of the data to induce symmetry. This trick does not work for GLMs, since taking pairwise differences does not preserve the function class we are trying to learn. Similarly, existing approaches do not generalize beyond linear functions. Our algorithm works for a large variety of generalized models, including (but not restricted to) ReLUs and sigmoids.

As our main result, we demonstrate an efficient algorithm to efficiently recover $g(w^* \cdot x)$ if the distribution satisfies an efficient identifiability condition (see Definition 2) and $\alpha = d^{-c}$ for any constant $c > 0$. If the condition of Definition 2 does not hold, our algorithm returns a list of candidates, each of which is an approximate translation of $g(w^* \cdot x)$ and one of which is guaranteed to be as close to $g(w^* \cdot x)$ as we would like. In fact, if the condition does not hold, it is information-theoretically impossible to learn a unique function that explains the data.

1.1. Our Results

We start by noting that, at the level of generality we consider, the learning problem we study is not identifiable, i.e., multiple candidates in our hypothesis class might explain the data equally well. As our first contribution, we identify a necessary and sufficient condition characterizing when a unique solution is identifiable. We describe the efficient identifiability condition below.

Definition 2 (Efficient Unique Identifiability) *We say u and v are Δ-separated if*

$$\mathop{\mathbf{E}}_{x}\left[|g(u \cdot x) - g(v \cdot x)|\right] > \Delta.$$

For any $\tau > 0$, an instance of the problem given in Definition 1 is (Δ, τ)-identifiable if any two Δ-separated u, v satisfy $\Pr_x\left[|g(u \cdot x) - g(v \cdot x) - A| > \tau\right] > \tau$ for all $A \in \mathbb{R}$.

Let $\mathbf{E}_x\left[|g(w \cdot x) - g(w^* \cdot x)|\right]$ denote the "excess loss" of w. Throughout the paper, we refer to Δ as the upper bound on the "excess loss" we would like to achieve. When the problem is (Δ, τ)-identifiable, the parameter τ describes the anti-concentration on the clean label difference $g(w \cdot x) - g(w^* \cdot x)$ centered around A.

Essentially, if there is a weight vector w that is Δ-separated from w^*, (Δ, τ)-identifiability ensures that $g(w \cdot x)$ is not close to a translation of $g(w^* \cdot x)$. On the other hand, if $g(w \cdot x)$ is approximately a translation of $g(w^* \cdot x)$ for most x, the following lower bound shows that the adversary can design oblivious noise distributions so that $g(w \cdot x)$ and $g(w^* \cdot x)$ are indistinguishable.

Theorem 3 (Necessity of Efficient Unique Identifiability) *Suppose that GLM-Ob(g, σ, w^*) is not (Δ, τ)-identifiable, i.e., there exist $u, v \in \mathbb{R}^d$ and $A \in \mathbb{R}$ such that u, v are Δ-separated and satisfy $\Pr_x\left[|g(u \cdot x) - g(v \cdot x) - A| > \tau\right] \le \tau$. Then any algorithm that distinguishes between u and v with probability at least $1 - \delta$ requires $m = \Omega(\min(\sigma, 1)\ln(1/\delta)/\tau)$ samples.*

Note that any algorithm that solves the oblivious regression problem must be able to differentiate between w^* and any Δ-separated candidate. Theorem 3 explains the necessity of the efficient identifiability condition for such differentiation. If no $\tau > 0$ satisfies the condition, then Theorem 3

implies that no algorithm with finite sample complexity can find a unique solution to oblivious regression. The result also shows that any (Δ, τ)-identifiable instance requires a sample complexity dependent on $1/\tau^*$, where the instance is (Δ, τ)-identifiable for all $\tau \leq \tau^*$.

Our main result is an efficient algorithm that performs GLM regression for any Lipschitz monotone activation function $g(\cdot)$. Our algorithm is qualitatively instance optimal – whenever the problem instance GLM-Ob(g, σ, w^*) is (Δ, τ)-identifiable, the algorithm returns a single candidate achieving excess loss of 4Δ with respect to $g(w^* \cdot x)$. If not (Δ, τ)-identifiable, then our algorithm returns a list of candidates, one element of which achieves excess loss of 4Δ.

Theorem 4 (Main Theorem) *There is an algorithm that takes as input the desired accuracy $\Delta > 0$, an upper bound R on $\|w^*\|$, τ, α and σ, draws $m = \mathrm{poly}(d, R, \sigma, \alpha^{-1}, \Delta^{-1})$ samples from GLM-Ob(g, σ, w^*), runs in time $\mathrm{poly}(d, R, \sigma, \alpha^{-1}, \Delta^{-1})$ and returns a $\mathrm{poly}(R, \sigma, \alpha^{-1}, \Delta^{-1})$-sized list of candidates, one of which achieves excess loss smaller than Δ, i.e., there is an element $\widehat{w} \in \mathbb{R}^d$ satisfying $\mathbf{E}_x[|g(\widehat{w} \cdot x) - g(w^* \cdot x)|] \leq \Delta$.*

Moreover, if the problem instance is (Δ, τ)-identifiable (as in Definition 2), then there is an algorithm which, takes as input $\Delta, R, \alpha, \sigma$ and τ as input, draws $\mathrm{poly}(d, R, \sigma, \alpha^{-1}, \Delta^{-1}, \tau^{-1})$ samples, runs in time $\mathrm{poly}(d, R, \sigma, \alpha^{-1}, \Delta^{-1}, \tau^{-1})$ and returns a single candidate.

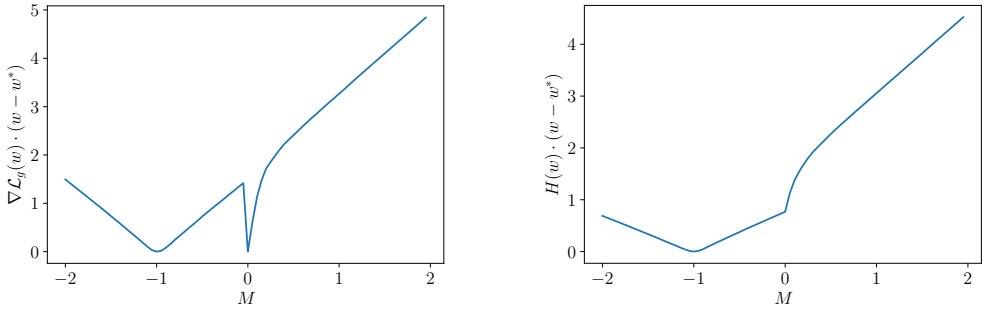
Our results hold for polynomially bounded x and w^* as well, by running the algorithm after scaling the x's and reparameterizing Δ. To see this, observe that we recover a $\widehat{w}$ such that $\mathbb{E}[|g(\widehat{w} \cdot x) - g(w^* \cdot x)|] \leq O(\Delta)$ for any choice of Δ when $\|x\| \leq 1$ and $\|w\| \leq R$ for polynomially bounded R. Suppose instead of the setting for the theorem, we have $\|x\| \leq A$ and $\|w\| \leq R$. We can then divide the x's by A and interpret $y(x) = g(w \cdot x) = g(Aw \cdot (x/A))$. We can then apply Theorem 4 with the upper bound on w set to AR and recover $\widehat{w}$, getting, $\mathbf{E}[|g((\widehat{w}/A) \cdot x) - g(w^* \cdot x)|] = \mathbf{E}[|g(\widehat{w} \cdot (x/A)) - g(w^* \cdot x)|] \leq O(\Delta)$. Prior work on linear regression with oblivious noise either assumed that the oblivious noise was symmetric or that the mean of the underlying distribution was zero. Our result holds in a significantly more general setting, even for the special case of linear regression, since we make no assumptions on the quantile of the oblivious noise or the mean of the underlying distribution.

At a high-level, we prove Theorem 4 in three steps: (1) We create an oracle that, given a sufficiently close estimate of $\Pr[\xi \leq 0]$, generates a hyperplane that separates vectors achieving large loss with respect to $g(w^* \cdot x)$ from those achieving small loss, (2) We use online gradient descent to produce a list of candidate solutions, one of which is close to the actual solution, (3) We apply a unique tournament-style pruning procedure that eliminates all candidates far away from w^*.

Since we do not have a good estimate of $\Pr[\xi \leq 0]$, we run steps (1) and (2) for each candidate value of $1 - 2\Pr[\xi \leq 0]$ chosen from a uniform partition of $[-1, 1]$ and then perform (3) on the union of all these candidates.

1.2. Technical Overview

For simplicity of exposition, we will analyze the problem without additive Gaussian noise and when the oblivious noise ξ is symmetric. This is the typical scenario for linear regression with oblivious noise in the context of general distributions. Inspired by the fact that the median of a dataset can be expressed as the ℓ_1 minimizer of the dataset, a natural idea is to minimize the ℓ_1 loss $\mathcal{L}_g(w) := \frac{1}{m}\sum_{i=1}^m |y_i - g(x_i \cdot w)|$. This simple approach has been used in the context of linear regression with oblivious noise (Nasrabadi et al., 2011) and also ReLU-regression for Massart



(*a*) Hyperplane based on $\nabla \mathcal{L}_{\mathrm{ReLU}}(w)$ (*b*) Our separating hyperplane $H(w)$

Figure 1: We set $g = \mathrm{ReLU}$ and $w^* = -1$ and plot according to the line $w = -Mw^*$. Here, $H(w) := \mathbf{E}_{x,y}[\mathrm{sign}(g(w \cdot x) - y)\, x]$. Observe that $\nabla \mathcal{L}_{\mathrm{ReLU}}(w) \cdot (w - w^*) \to 0$ as $M \to 0^+$ even when $w^* \neq 0$. On the other hand, $H(w)$ does not suffer from this. Here ξ takes the value $-3/20$ w.p. 0.3, 0 w.p. 0.1, and $3/20$ w.p. 0.6; $\epsilon \sim \mathcal{N}(0, 0.25)$, and x is drawn uniformly from the 2-dimensional unit ball.

noise (Diakonikolas et al., 2021b). Unfortunately, if the activation function g is not linear, the loss $\mathcal{L}_g(w)$ is not convex. Let $\mathcal{L}^*_g(w) := \frac{1}{m}\sum_{i=1}^m |g(w^* \cdot x_i) - g(w \cdot x_i)|$ denote the clean loss. To solve the problem of optimizing a nonconvex function, instead of using gradient-based methods, we can create an oracle that produces a separating hyperplane between points achieving a large clean loss and those achieving a small clean loss. The oracle produces a vector $H(w)$ satisfying $H(w) \cdot (w - w^*) \geq c > 0$. We then reduce the problem to online convex optimization (OCO).

Oracle for Separating Hyperplane Unfortunately, unlike the case of convex functions — or as it was used in Diakonikolas et al. (2021b) to perform ReLU regression — we cannot use $\nabla \mathcal{L}_g(w) = \mathbf{E}_{x,y}[\mathrm{sign}(g(w \cdot x) - y)\mathbf{1}(w \cdot x \geq 0)\, x]$ as an oracle for generating a separating hyperplane, since it cannot distinguish $w = 0$ from w^* even when $w^* \neq 0$. This is illustrated in Figure 1 for $g = \mathrm{ReLU}$.

We instead take inspiration from the gradient of a *linear regression* problem. Suppose we are given samples (x_i, y_i) such that $y_i = w^* \cdot x_i + \xi'$, where ξ' is symmetric oblivious noise such that $\Pr[\xi' = 0] \geq \alpha$, and the goal is to recover $\widehat{w}$ which is the ℓ_1 minimizer, i.e., $\widehat{w} := \mathrm{argmin}_{w \in \mathbb{R}^d} \mathcal{L}'(w) := \frac{1}{m}\sum_{i=1}^m |y_i - w \cdot x_i|$. This is a convex program, and a subgradient of $\mathcal{L}'(w)$ is given by $\nabla \mathcal{L}'(w) = \mathbf{E}_{x,y}[\mathrm{sign}(w \cdot x - y)\, x] = \mathbf{E}_x\left[\mathbf{E}_{\xi'}\left[\mathrm{sign}((w \cdot x - w^* \cdot x) - \xi')\right] x\right]$.

We now examine the expectation over ξ'. Since the median of ξ' is 0 and it takes the value 0 with probability at least α, the probability that $\mathrm{sign}((w \cdot x - w^* \cdot x) - \xi') = \mathrm{sign}(w \cdot x - w^* \cdot x)$ is at least $\frac{1+\alpha}{2}$. This implies that $\mathbf{E}_{\xi'}\left[\mathrm{sign}((w \cdot x - w^* \cdot x) - \xi')\right] \geq \alpha\, \mathrm{sign}(w \cdot x - w^* \cdot x)$, since $(w \cdot x - w^* \cdot x) - \xi'$ is more often biased towards $(w \cdot x - w^* \cdot x)$ than it is towards $-(w \cdot x - w^* \cdot x)$ and $\Pr[\xi = 0] \geq \alpha$. Therefore, $\nabla \mathcal{L}'(w) \cdot (w - w^*) \geq \alpha\, \mathbf{E}_x[|w \cdot x - w^* \cdot x|]$.

While we do not have access to $w^* \cdot x + \xi'$ we do have access to $g(w^* \cdot x) + \xi$. At this point we make two observations: (1) Since g is monotonically non-decreasing, it follows that $\mathrm{sign}(g(w \cdot x) - g(w^* \cdot x)) = \mathrm{sign}(w \cdot x - w^* \cdot x)$ whenever $g(w \cdot x) \neq g(w^* \cdot x)$. (2) Since g is 1-Lipschitz, it follows that $|w \cdot x - w^* \cdot x| \geq |g(w \cdot x) - g(w^* \cdot x)|$. An argument analogous to the one above then shows us that $H(w) := \mathbf{E}_{x,y}\left[\mathrm{sign}(g(w \cdot x) - y)\, x\right]$ satisfies $H(w) \cdot (w - w^*) \geq \alpha\, \mathbf{E}_x[|g(w \cdot x) - g(w^* \cdot x)|]$, hence allowing us to separate w's which achieve small clean loss from

those which achieve larger clean loss. In Lemma 7, we demonstrate this in the presence of additive Gaussian noise and without the assumption of symmetry on ξ.

Reduction to Online Convex Optimization In Lemma 10, we show that if we have a good estimate of the quantile at which ξ is 0, we can use our separating hyperplane oracle as the gradient oracle for online gradient descent to optimize the clean loss $\mathcal{L}_g^*(w) := \mathbf{E}_x [|g(w \cdot x) - g(w^* \cdot x)|]$. Since this function is nonconvex, our reduction leaves us with a set of candidates which are iterates of our online gradient descent procedure. Our minimizer is one of these candidates. We then prune out candidates which do not explain the data.

Pruning Bad Candidates Finally, Lemma 22 shows that we can efficiently prune implausible candidates if the list of candidates contains a vector close to w^*. For simplicity of exposition, assume for now that $w^* \in \mathcal{W}$. Our pruning procedure relies on the following two observations: (1) There is no way to find disjoint subsets of the space of x's such that $(y - g(w^* \cdot x))$ takes the value 0 at different quantiles when conditioned on these subsets. (2) Suppose that, for some A, we identify E^+ and E^- such that $x \in E^+$ implies $g(w \cdot x) - g(w^* \cdot x) - A > \tau$ and $x \in E^-$ implies $g(w \cdot x) - g(w^* \cdot x) - A < -\tau$. Then the quantiles at which $(y - g(w \cdot x)) = (g(w^* \cdot x) - g(w \cdot x) + \xi + \epsilon)$ take the value 0 in these two sets differ by at least α.

We can use these observations to determine if a given candidate is equal to w^* or not, by looking at the quantity $g(w \cdot x) - g(w^* \cdot x) - A$. Specifically, we try to find two subsets E^+ and E^- such that $g(w \cdot x) - g(w^* \cdot x)$ is large and positive when $x \in E^+$, and is large and negative when $x \in E^-$. We reject w by comparing the quantiles of $(y - g(w \cdot x))$ when conditioned on x belonging to E^+ and E^-. While we do not know what w^* is beforehand, we know that $w^* \in \mathcal{W}$, and so we iterate over elements in $\mathcal{W}$ to check for the existence of a partition which will allow us to reject w. If $w = w^*$ such a partition is not possible, and w will not be rejected. On the other hand, each candidate remaining in the list will be close to a translation of $g(w^* \cdot x)$, and one of the candidates will be w^*.

1.3. Prior Work

Given the extensive literature on robust regression, here we focus on the most relevant work.

GLM regression Various formalizations of GLM regression have been studied extensively over the past decades; see., e,g., Nelder and Wedderburn (1972); Kalai and Sastry (2009); Kakade et al. (2011); Klivans and Meka (2017). Recently, there has been increased focus on GLM regression for activation functions that are popular in deep learning, including ReLUs. This problem has previously been considered both in the context of far weaker noise models, such as the realizable setting (Soltanolkotabi, 2017; Kalan et al., 2019; Yehudai and Ohad, 2020), as well as in the context of far more challenging noise models (Goel et al., 2019; Diakonikolas et al., 2020b; Goel et al., 2020; Diakonikolas et al., 2021a, 2020a, 2022a,d; Wang et al., 2023).

Even in the realizable setting, it turns out that the squared loss has exponentially many local minima for the logistic activation function (Auer et al., 1995). On the positive side, Diakonikolas et al. (2020a) gave an efficient learner in the presence of adversarial label noise with constant factor approximation guarantees under logconcave distributions. This algorithmic result was generalized to broader families of activations under much milder distributional assumptions in Diakonikolas et al. (2022d); Wang et al. (2023). On the other hand, without distributional assumptions, even approximate learning is hard (Hardt and Moitra, 2013; Manurangsi and Reichman, 2018; Diakonikolas et al., 2022a). In a related direction, there have been attempts to study the problem in the distribution-free

setting under semi-random label noise. Specifically, Diakonikolas et al. (2021b) studied this problem in the presence of bounded (Massart) noise, where the adversary can arbitrarily corrupt a randomly selected subset of at most half the samples. Prior to that, Karmakar et al. (2020) studied it in the realizable setting under a noise model similar to (but more restrictive than) the Massart noise model, while Chen et al. (2020a) studied GLMs under Massart noise for classification.

In our work, we study this problem for general GLMs in the oblivious setting, with the goal of being able to tolerate $1 - o(1)$ fraction of the samples being corrupted. In this setting, we recover the candidate solution to arbitrarily small precision in ℓ_1 norm (with respect to the objective). Since $\|x\|_2 \leq 1$ and $\|w\|_2 \leq R$, it is easy to also provide the corresponding guarantees in the ℓ_2 norm, which was the convention in older works (Kakade et al., 2011; Goel et al., 2019).

The Oblivious Noise Model The oblivious noise model could be viewed as an attempt at characterizing the most general noise model that allows almost all points to be arbitrarily corrupted, while still allowing for recovery of the target function with vanishing error. This model has been considered for natural statistical problems, including PCA, sparse recovery (Pesme and Flammarion, 2020; d'Orsi et al., 2021a), as well as linear regression in the online setting (Dalalyan and Thompson, 2019) and the problem of estimating a signal x^* with additive oblivious noise (d'Orsi et al., 2022).

The setting closest to the one considered in this paper is that of linear regression. Until very recently, the problem had been studied primarily in the context of Gaussian design matrices, i.e., when x's are drawn from $\mathcal{N}(0, \Sigma)$. One of the main goals in this line of work is to design an algorithm that can tolerate the largest possible fraction of the labels being corrupted. Initial works on linear regression either were not consistent as the error did not go to 0 with increasing samples (Wright and Ma, 2010; Nasrabadi et al., 2011) or failed to achieve the right convergence rates or breakdown point (Tsakonas et al., 2014; Bhatia et al., 2015). Suggala et al. (2019) provided the first consistent estimator achieving an error of $O(d/\alpha^2 m)$ for any $\alpha > 1/\log\log(m)$. Later, d'Orsi et al. (2021b) improved this rate to $\alpha > 1/d^c$ for constant c, while also generalizing the class of design matrices.

Most of these prior results focused on either the oblivious noise being symmetric (or median 0), or the underlying distribution being (sub)-Gaussian. In some of these settings (such as that of linear regression) it is possible to reduce the general problem to this restrictive setting, as is done in Norman et al. (2022). However, for GLM regression, we cannot exploit the symmetry that is either induced by the distribution or the class of linear functions. In terms of lower bounds, Chen and d'Orsi (2022) identify a "well-spreadness" condition (the column space of the measurements being far from sparse vectors) as a property that is necessary for recovery even when the oblivious noise is symmetric. Notably, these lower bounds are relevant when the goal is to perform parameter recovery or achieve a rate better than σ^2/m. In our paper, we instead give the first result for a far more general problem and with the objective of minimizing the clean loss, but not necessarily parameter recovery. Our lower bound follows from the fact that we cannot distinguish between translations of the target function from the data without making any assumptions on the oblivious noise.

2. Preliminaries

Basic Notation We use $\mathbb{R}$ to denote the set of real numbers. For $n \in \mathbb{Z}_+$ we denote $[n] := \{1, \ldots, n\}$. We assume $\mathrm{sign}(0) = 0$. We denote by $\mathbf{1}(E)$ the indicator function of the event E. We use $\mathrm{poly}(\cdot)$ to indicate a quantity that is polynomial in its arguments. Similarly, $\mathrm{polylog}(\cdot)$ denotes a quantity that is polynomial in the logarithm of its arguments. For two functions $f, g : \mathbb{R} \to \mathbb{R}$, we

say $f \lesssim g$ if there exist constants $C_1, C_2 > 0$ such that for all $x \geq C_1$, $f(x) \leq C_2 g(x)$. For two numbers $a, b \in \mathbb{R}$, $\min(a, b)$ returns the smaller of the two. We say that a function f is L-Lipschitz if $f(x) - f(y) \leq L\|x - y\|_2$.

Linear Algebra Notation We typically use small case letters for deterministic vectors and scalars. For a vector v, we let $\|v\|_2$ denote its ℓ_2-norm. We denote the inner product of two vectors u, v by $u \cdot v$. We denote the d-dimensional radius-R ball centered at the origin by $B_d(R)$.

Probability Notation For a random variable X, we use $\mathbf{E}[X]$ for its expectation and $\Pr[X \in E]$ for the probability of the random variable belonging to the set E. We use $\mathcal{N}(\mu, \sigma^2)$ to denote the Gaussian distribution with mean μ and variance σ^2. When D is a distribution, we use $X \sim D$ to denote that the random variable X is distributed according to D. When S is a set, we let $\mathbf{E}_{X \sim S}[\cdot]$ denote the expectation under the uniform distribution over S. When clear from context, we denote the empirical expectation and probability by $\widehat{\mathbf{E}}$ and $\widehat{\Pr}$.

2.1. Facts

The proofs of the following facts can be found in Appendix B.

Fact 5 *Let ξ be oblivious noise such that $\Pr[\xi = 0] \geq \alpha$. Then the quantity*

$$F_{\sigma,\xi}(t) := \mathop{\mathbf{E}}_{\epsilon,\xi}[\mathrm{sign}(t + \epsilon + \xi)] - \mathop{\mathbf{E}}_{\epsilon,\xi}[\mathrm{sign}(\epsilon + \xi)]$$

satisfies the following: (1) $F_{\sigma,\xi}$ is strictly increasing, (2) $\mathrm{sign}(F_{\sigma,\xi}(t)) = \mathrm{sign}(t)$, and (3) For any $\gamma \leq 2$, whenever $|t| \geq \gamma\sigma$, $|F_{\sigma,\xi}(t)| > (\gamma\alpha/4)$ and whenever $|t| \leq \gamma\sigma$, $|F_{\sigma,\xi}(t)| \leq (\alpha t/4\sigma)$.

Fact 6 *Let X be a random variable on $\mathbb{R}$. Fix $\tau > 0$ and $\eta > 0$. Define the events E_A^+ and E_A^- such that $\Pr[E_A^+] = \Pr[X > A + \tau]$ and $\Pr[E_A^-] = \Pr[X < A - \tau]$. Then if the following first condition is not true, the second condition is: (1) $\exists A \in \mathbb{R}$ such that $\Pr[E_A^+] \geq \eta$ and $\Pr[E_A^-] \geq \eta$. (2) $\exists A^* \in \mathbb{R}$ such that $\Pr[E_{A^*}^+] \leq \eta$ and $\Pr[E_{A^*}^-] \leq \eta$.*

3. Oblivious Regression via Online Convex Optimization

3.1. A Direction of Improvement

We assume prior knowledge of a constant c that approximates $\mathbf{E}_{\xi,\epsilon}[\mathrm{sign}(\xi + \epsilon)]$. In the following key lemma, we demonstrate an oracle for a hyperplane that separates all vectors that are Δ-separated from w^*. For the following results in Section 3 and later in the paper, we use γ to denote $\min(\Delta/4\sigma, 1/2)$.

Lemma 7 (Separating Hyperplane) *Let $\mathcal{D} = \text{GLM-Ob}(g, \sigma, w^*)$ as defined in Definition 1 and define $\gamma = \min(\Delta/4\sigma, 1/2)$. Suppose $c \in \mathbb{R}$ such that $|\mathbf{E}_{\xi,\epsilon}[\mathrm{sign}(\xi + \epsilon)] - c| \leq \gamma\alpha\Delta/32R$. Then, for $w \in B_d(R)$, $H_c(w) := \mathbf{E}_{x,y}\left[(\mathrm{sign}(y - g(w \cdot x)) - c)\ x\right]$ satisfies*

$$H_c(w) \cdot (w^* - w) \geq (\gamma\alpha/4) \mathop{\mathbf{E}}_{x}[|(g(w^* \cdot x) - g(w \cdot x))|] - (\gamma^2\alpha\sigma/4) - (\gamma\alpha\Delta/16).$$

Specifically, if $\mathbf{E}_x[|(g(w^* \cdot x) - g(w \cdot x))|] > \Delta$, we have that $H_c(w) \cdot (w^* - w) \geq (\alpha\Delta^2)/(32\sigma)$ if $\Delta \leq 2\sigma$, and $H_c(w) \cdot (w^* - w) \geq \alpha\Delta/8$ if $\Delta > 2\sigma$.

Proof Let $F_{\sigma,\xi}(t) := \mathbf{E}_{\epsilon,\xi}[\text{sign}(t+\epsilon+\xi) - \text{sign}(\epsilon+\xi)]$. Then we can write

$$\begin{aligned}
H_c(w)\cdot(w^*-w) &= \mathop{\mathbf{E}}_{x,\epsilon,\xi}[(\text{sign}(g(w^*\cdot x) - g(w\cdot x) + \epsilon + \xi) - c)\ (x\cdot(w^*-w))] \\
&= \mathop{\mathbf{E}}_{x}\left[\left(F_{\sigma,\xi}(g(w^*\cdot x) - g(w\cdot x)) + (\mathop{\mathbf{E}}_{\xi,\epsilon}[\text{sign}(\xi+\epsilon)] - c)\right)\ (x\cdot(w^*-w))\right] \\
&= \mathop{\mathbf{E}}_{x}\left[F_{\sigma,\xi}(g(w^*\cdot x) - g(w\cdot x))(x\cdot(w^*-w))\right] \\
&\quad + (\mathop{\mathbf{E}}_{\xi,\epsilon}[\text{sign}(\xi+\epsilon)] - c)\ \mathop{\mathbf{E}}_{x}\left[x\cdot(w^*-w)\right].
\end{aligned}$$

By Fact 5 and the fact that g is monotone, it follows that $\text{sign}(F_{\sigma,\xi}(g(w^*\cdot x) - g(w\cdot x))) = \text{sign}(g(w^*\cdot x) - g(w\cdot x)) = \text{sign}(x\cdot(w^*-w))$ whenever $g(w^*\cdot x) \neq g(w\cdot x)$. Combining this with the fact that $g(\cdot)$ is 1-Lipschitz, we get

$$\begin{aligned}
&\mathop{\mathbf{E}}_{x}\left[F_{\sigma,\xi}(g(w^*\cdot x) - g(w\cdot x))(x\cdot(w^*-w))\right] \\
&\geq \mathop{\mathbf{E}}_{x}\left[F_{\sigma,\xi}(g(w^*\cdot x) - g(w\cdot x))(g(w^*\cdot x) - g(w\cdot x))\right]
\end{aligned}$$

Continuing the calculation above, we see

$$\begin{aligned}
H_c(w)\cdot(w^*-w) \geq \mathop{\mathbf{E}}_{x}\left[F_{\sigma,\xi}(g(w^*\cdot x) - g(w\cdot x))(g(w^*\cdot x) - g(w\cdot x))\right] \\
- 2R\,|\mathop{\mathbf{E}}_{\xi,\epsilon}[\text{sign}(\xi+\epsilon)] - c|\,,
\end{aligned}$$

where the bound on the second quantity follows from the fact that $\|w\|_2, \|w^*\|_2 \leq R$ and $\|x\|_2 \leq 1$. Fact 5 implies that $|F_{\sigma,\xi}(t)| \geq \gamma\alpha/4$ if $|t| \geq \gamma\sigma$, whenever $\gamma \leq 2$. We now consider the event $E_\gamma := \{x \mid |g(w^*\cdot x) - g(w\cdot x)| \geq \gamma\sigma\}$, which describes the region where there is significant difference between the hypothesis w and the target w^*. Then we can write

$$\begin{aligned}
H_c(w)\cdot(w^*-w) &\geq (\gamma\alpha/4)\ \mathop{\mathbf{E}}_{x}\left[|(g(w^*\cdot x) - g(w\cdot x))|\mathbf{1}(x\in E_\gamma)\right] - 2R\,|\mathop{\mathbf{E}}_{\xi,\epsilon}[\text{sign}(\xi+\epsilon)] - c| \\
&\geq (\gamma\alpha/4)(\mathop{\mathbf{E}}_{x}\left[|(g(w^*\cdot x) - g(w\cdot x))|\right] - \gamma\sigma) - 2R\,|\mathop{\mathbf{E}}_{\xi,\epsilon}[\text{sign}(\xi+\epsilon)] - c| \\
&\geq (\gamma\alpha/4)\mathop{\mathbf{E}}_{x}\left[|(g(w^*\cdot x) - g(w\cdot x))|\right] - (\gamma^2\alpha\sigma/4) - 2R\,|\mathop{\mathbf{E}}_{\xi,\epsilon}[\text{sign}(\xi+\epsilon)] - c|\,.
\end{aligned}$$

In the case that $\mathbf{E}_x\left[|(g(w^*\cdot x) - g(w\cdot x))|\right] > \Delta$, we would like to set the parameter γ such that $(\gamma^2\alpha\sigma/4) + 2R\,|\mathbf{E}_{\xi,\epsilon}[\text{sign}(\xi+\epsilon)] - c| \leq \gamma\alpha\Delta/8$, ensuring that the right hand side above is strictly positive. By assumption, we know that c satisfies $|\mathbf{E}_{\xi,\epsilon}[\text{sign}(\xi+\epsilon)] - c| \leq (\gamma\alpha\Delta/32R)$, so it suffices for γ to satisfy $(\gamma^2\alpha\sigma/2) \leq \gamma\alpha\Delta/8$, i.e., $\gamma \leq \Delta/4\sigma$, in addition to $\gamma \leq 2$. Here, we set $\gamma = \min(\Delta/4\sigma, 1/2)$. Putting these together, we see that when $\Delta \leq 2\sigma$, it holds

$$\begin{aligned}
H_c(w)\cdot(w^*-w) &\geq (\gamma\alpha/4)\mathop{\mathbf{E}}_{x}\left[|(g(w^*\cdot x) - g(w\cdot x))|\right] - (\gamma^2\alpha\sigma/4) - 2R\,|\mathop{\mathbf{E}}_{\xi,\epsilon}[\text{sign}(\xi+\epsilon)] - c| \\
&\geq (\gamma\alpha/4)\mathop{\mathbf{E}}_{x}\left[|(g(w^*\cdot x) - g(w\cdot x))|\right] - (\gamma^2\alpha\sigma/4) - (\gamma\alpha\Delta/16) \\
&\geq (\alpha\Delta)/(16\sigma)\mathop{\mathbf{E}}_{x}\left[|(g(w^*\cdot x) - g(w\cdot x))|\right] - (\alpha\Delta^2)/(64\sigma) - (\alpha\Delta^2/64\sigma).
\end{aligned}$$

In the case that we look at a vector w that is Δ-separated from w^*, the lower bound we get is $(\alpha\Delta^2)/(32\sigma)$ when $\Delta \leq 2\sigma$, while the lower bound is $\alpha\Delta/8$ when $\Delta > 2\sigma$. ■

The following corollary allows us to extend Lemma 7 to the empirical setting. The proof of the corollary can be found in Appendix A.

Corollary 8 (Empirical Separating Hyperplane) *Let $(x_i, y_i)_{i=1}^m \sim$ GLM-Ob$(g, \sigma, w^*)^m$, where $m \gtrsim R^2 \ln(1/\delta)/(\gamma\alpha\Delta)^2$. Assume c satisfies the assumption in Lemma 7. Define $\widehat{H}_c(w) := (1/m) \sum_{i=1}^m [(\mathrm{sign}(g(w \cdot x_i) - y_i) - c)\ x_i]$. Then, for any w, it holds*

$$\widehat{H}_c(w) \cdot (w - w^*) \geq (\gamma\alpha/4) \mathop{\mathbf{E}}_x [|(g(w^* \cdot x) - g(w \cdot x))|] - \gamma^2\alpha\sigma/4 - 3\ (\gamma\alpha\Delta/32)$$

with probability at least $1 - \delta$.

While not directly useful in the proof we present here, as pointed out by a reviewer, we note that our direction of improvement $\widehat{H}_c(w)$ as defined in Corollary 8 can be interpreted to be the gradient of the convex surrogate loss $(1/m) \sum_{i=1}^m \int_0^{w \cdot x_i} (\mathrm{sign}(g(z) - y_i) + c)\, dz$. This has an analogy to the "matching loss" $(1/m) \sum_{i=1}^m \int_0^{w \cdot x_i} (g(z) - y_i)\, dz$ as considered for the case of ℓ_2 GLM regression introduced in the work of Auer (1997) and used extensively in subsequent works.

3.2. Reduction to Online Convex Optimization

If c is a good approximation of $\mathbf{E}_{\xi,\epsilon}[\mathrm{sign}(\xi + \epsilon)]$, we can reduce the problem to online convex optimization to now get a set of candidates, one of which is close to the true solution.

OCO Setting The typical online convex optimization scenario can be modelled as the following game: at time $t - 1$ the player must pick a candidate point w_t belonging to a certain constrained set W. At time t the true convex loss $f_t(\cdot)$ is revealed and the player suffers a loss of $f_t(w_t)$. This continues for a total of T rounds. Algorithms for these settings typically upper bound the regret ($R(\{w_i\}_{i=1}^T)$), which is the performance with respect to the optimal fixed point in hindsight, $R(\{w_i\}_{i=1}^T) := \sum_{i=1}^T f_t(w_t) - \min_{w^* \in W} \left(\sum_{i=1}^T f_t(w^*) \right)$.

We specialize Theorem 3.1 from Hazan (2016) to our setting to get the following lemma.

Lemma 9 (Theorem 3.1 from Hazan (2016)) *Suppose $v_1, \ldots, v_T \in \mathbb{R}^d$ such that for all $t \in [T]$ and $\|v_t\|_2 \leq G$. Then online gradient descent with step sizes $\{\eta_t = \frac{R}{G\sqrt{t}} \mid t \in [T]\}$, for linear cost functions $f_t(w) := v_t \cdot w$, outputs a sequence of predictions $w_1, \ldots, w_T \in B_d(R)$ such that $\sum_{t=1}^T f_t(w_t) - \min_{\|w\|_2 \leq R} \sum_{t=1}^T f_t(w) \leq (3/2)\, GR\sqrt{T}$.*

An application of this lemma then gives us our result for reducing the problem to OCO.

Lemma 10 (Reduction to OCO) *Suppose $(x_1, y_1), \ldots, (x_m, y_m)$ are drawn from GLM-Ob(g, σ, w^*) and c satisfies the assumption in Lemma 7. Let $T \gtrsim (R/\gamma\alpha)^2$ and $m \gtrsim R^2 \ln(T/\delta)/(\gamma\alpha\Delta)^2$. Then there is an algorithm which recovers a set of candidates $w_1, \ldots, w_T$ with probability $1 - \delta$ such that*

$$\min_{w_t} \left\{ \mathop{\mathbf{E}}_x [|g(w_t \cdot x) - g(w^* \cdot x)|] \right\} \leq 3\Delta.$$

Proof At round t, the player proposes weight vector w_t, at which point the function $f_t(\cdot)$ is revealed to be $f_t(w) := v_t \cdot w$ where $v_t := \widehat{H}_c(w_t)$ as defined in Corollary 8. Note that a union bound over the T final candidates will ensure that with $m \gtrsim R^2 \ln(T/\delta)/(\gamma\alpha\Delta)^2$ samples, with probability $1 - \delta$, for every $t \in [1, T]$, $\widehat{H}_c(w_t)$ satisfies the conclusion of Corollary 8.

An application of Lemma 9 to this setting gives

$$\frac{1}{T} \sum_{t=1}^T f_t(w_t) \leq \min_{\|w\| \leq R} \left(\frac{1}{T} \sum_{t=1}^T f_t(w) \right) + \frac{(3/2)GR}{\sqrt{T}} \leq \frac{1}{T} \sum_{t=1}^T f_t(w^*) + \frac{(3/2)GR}{\sqrt{T}}\ .$$

Rearranging this and applying Corollary 8 we get

$$\begin{aligned}\frac{(3/2)GR}{\sqrt{T}} &\geq \frac{1}{T}\left(\sum_{t=1}^{T} f_t(w_t) - \sum_{t=1}^{T} f_t(w^*)\right) = \frac{1}{T}\left(\sum_{t=1}^{T} v_t \cdot (w_t - w^*)\right) \\ &\geq \frac{\gamma\alpha}{4T}\left(\sum_{t=1}^{T} \mathop{\mathbf{E}}_{x}\left[|g(x \cdot w_t) - g(x \cdot w^*)|\right]\right) - \gamma^2\alpha\sigma/4 - 3\left(\gamma\alpha\Delta/32\right) \\ &\geq (\gamma\alpha/4)\min_{w_t}\left\{\mathop{\mathbf{E}}_{x}\left[|g(x \cdot w_t) - g(x \cdot w^*)|\right]\right\} - \gamma^2\alpha\sigma/4 - 3\left(\gamma\alpha\Delta/32\right),\end{aligned}$$

where the final inequality follows from the fact that the minimum is smaller than the average. Rearranging this gives us $\frac{6GR}{\gamma\alpha\sqrt{T}} + \gamma\sigma + (3/8)\Delta \geq \min_{w_t}\{\mathbf{E}_x\left[|g(x \cdot w_t) - g(x \cdot w^*)|\right]\}$. Substituting $\|v_t\|_2 \leq G = 2$ and $\gamma = \min(\Delta/4\sigma, 1/2)$, we get

$$O\left(\frac{R}{\gamma\alpha\sqrt{T}}\right) + 2\Delta \geq \min_{w_t}\left\{\mathop{\mathbf{E}}_{x}\left[|g(x \cdot w_t) - g(x \cdot w^*)|\right]\right\}$$

and so, setting $T \gtrsim (R/\gamma\alpha)^2$ ensures that we achieve an error of 3Δ. ■

Note that if the desired lower bound was a convex function (instead of $\mathbf{E}_x[|g(x \cdot w) - g(x \cdot w^*)|]$), we would not have to take the minimum of all the iterates in the proof. We could instead use Jensen's inequality to take the loss of the average iterates. Unfortunately, because the objective can be non-convex due to the nonlinearity of the activation function g, we can't just use the averaged iterates.

4. Pruning Implausible Candidates

Lemma 10 can generate potential solutions to achieve a low clean loss with respect to $g(w^* \cdot x)$ if c is a good approximation of $\mathbf{E}_{\xi,\epsilon}[\text{sign}(\xi + \epsilon)]$. Unfortunately, it is difficult to verify the accuracy of these candidates on the data since it is impossible to differentiate between translations of $g(w^* \cdot x)$ due to the generality of the setting and since $\mathbf{E}_{\xi,\epsilon}[\text{sign}(\xi + \epsilon)]$ is unknown. Our algorithm generates T candidates for each value of c in a uniform partition of $[-1, 1]$. One the candidates is close to w^*, however, the problem of spurious candidates still remains. In this section, we discuss how to determine which of the candidate solutions is the best fit for the data.

Even though it is difficult to test if a single hypothesis achieves a small clean loss, it is surprisingly possible to find a good hypothesis out of a list of candidates. Algorithm 3 describes a tournament-style testing procedure which produces a set of candidates approximately equal to $g(w^* \cdot x)$, and if efficient identifiability holds for the instance, this list will only contain one candidate. The proof of Lemma 22 is presented in Appendix C.

Lemma 11 (Pruning bad candidates) *Let $\delta > 0$. Suppose $\exists \widehat{w} \in \mathcal{W}$ such that*

$$\mathop{\mathbf{E}}_{x}\left[|g(\widehat{w} \cdot x) - g(w^* \cdot x)|\right] \leq \min\{\Delta, \tau^2/16\}.$$

Then Algorithm 3 draws $m \gtrsim \log(|\mathcal{W}|^2/\delta)/(\alpha\tau(\min\{\tau/\sigma, 1\}))^2 + R^2\log(|\mathcal{W}|^2/\delta)/\Delta^2$ samples, runs in time $\tilde{O}(m|\mathcal{W}|^2)$, and with probability $1 - \delta$ returns a list of candidates containing $\widehat{w}$ such that each candidate satisfies $\Pr[|g(w^* \cdot x) - g(w \cdot x) - A_w| > \tau] \leq 1 - \tau$ for some $A_w \in \mathbb{R}$. If (Δ, τ)-identifiability (Definition 2) holds, the algorithm only returns a single candidate $\widehat{w}$ which achieves a clean loss of 4Δ.

Proof Sketch For the sake of exposition, suppose $\widehat{w} = w^*$ and the empircal estimates equal the true expectation. Define the events $E_s^+(u,v) := \{x \mid g(u\cdot x) - g(v\cdot x) > s\}$ and $E_s^-(u,v) := \{x \mid g(u\cdot x) - g(v\cdot x) < s\}$. An application of Fact 6 to the random variable $g(w^*\cdot x) - g(w\cdot x)$ implies that for any τ if the following first condition is false, then the second condition is true:

1. $\exists A \in \mathbb{R}$ such that $\Pr[E^+_{A+\tau}(w^*, w)] \geq \tau/2$ and $\Pr[E^-_{A-\tau}(w^*, w)] \geq \tau/2$.
2. $\exists A \in \mathbb{R}$ such that $\Pr[E^+_{A+\tau}(w^*, w)] \leq \tau/2$ and $\Pr[E^-_{A-\tau}(w^*, w)] \leq \tau/2$.

If w satisfies Condition 1, then $g(w^*\cdot x) - g(w\cdot x) - A$ takes values $> \tau$ and $\leq \tau$ when $x \in E^+_{A+\tau}(w^*, w)$ and $E^-_{A-\tau}(w^*, w)$ respectively. This means the quantile at which $(y - g(w\cdot x))$ takes the value 0 is different conditioned on x coming from both these sets. Let $R^+ := \{(y - g(w\cdot x)) \mid x \in E^+_{A+\tau}(w^*, w)\}$ and $R^- := \{(y - g(w\cdot x)) \mid x \in E^-_{A-\tau}(w^*, w)\}$ Our algorithm rejects w if there is an A such that $|\widehat{\mathbf{E}}[\mathrm{sign}(r - A) \mid r \in R^+] - \widehat{\mathbf{E}}[\mathrm{sign}(r - A) \mid r \in R^-]|$ is large. This will be the case since elements of R^+ and R^- are drawn from the distribution of $\xi + \epsilon$ shifted by at least τ in opposite directions, and ξ places a mass of α at 0.

Hence, all remaining candidates satisfy Condition 2, which means they are approximate translations of $g(w^*\cdot x)$. Also, since w^* is never rejected, we know that w^* also belongs to this list. If (Δ, τ)-identifiability holds, every element of the final list achieves clean loss Δ. We can test this by checking of every pair of candidates in the list is 2Δ-close, and if they are, returning any element of the list.

Algorithm 1 Prune Implausible Candidates

input: $\tau, \alpha, \sigma, R, \mathcal{W} = \{w_1, \ldots, w_p\}$
Draw $m = C\log(|\mathcal{W}|^2/\delta)/(\alpha\tau(\min\{\tau/2\sigma, 1\}))^2$ samples $\{(x_k, y_k)\}_{k=1}^m$ for some constant C.
for $i \leftarrow 1 ... p$ **do**
  for $j \leftarrow i+1 ... p$ **do**
    Let $E_A^+ := \{x_k | g(w_i\cdot x_k) - g(w_j\cdot x_k) > A\}$ and $E_A^- := \{x_k | g(w_i\cdot x_k) - g(w_j\cdot x_k) < A\}$
    Compute range U^+ of A such that $|E^+_{A+\tau/2}| \geq \alpha m \min\{\tau/2\sigma, 1/4\}$ via binary search on at most m distinct $g(w_i\cdot x_k) - g(w_j\cdot x_k) - \tau/2$ and similarly U^- for $|E^-_{A-\tau/2}|$
    Let $A \leftarrow$ any number in $U^+ \cap U^-$
    if *no such A exists* **then**
      continue to $(j+1)$-th inner loop
    Compute $R^+ = \{r | r = y_i - g(w_i\cdot x)$ for $x \in E^+_{A+\tau/2}\}$ and similarly R^- for $E^-_{A-\tau/2}$
    if $|\widehat{\mathbf{E}}[\mathrm{sign}(r - A) \mid r \in R^+] - \widehat{\mathbf{E}}[\mathrm{sign}(r - A) \mid r \in R^-]| > \alpha \min\{\tau/16\sigma, 1/8\}$ **then**
      reject w_i and continue with $(i+1)$-th outer loop
for $i \leftarrow 1 \ldots p$ **do**
  for $j \leftarrow 1 \ldots p$ **do**
    if $\frac{1}{m}\sum_{t=1}^m |g(w_i\cdot x_t) - g(w_j\cdot x_t)| > 3\Delta$ **then**
      return $\mathcal{W}$
Sample $\widehat{w}$ uniformly from $\mathcal{W}$.
return $\{\widehat{w}\}$.

Algorithm 2 Oblivious GLM Regression

input: $\{(x_i, y_i) \mid i \in [m]\} \sim$ GLM-Ob$(g, \sigma, w^*)^m$, R, σ, τ, α where w^* is unknown and $\|w^*\| \leq R$.
Let P be a uniform parititon of $[-1, 1]$ with granularity $\gamma\alpha\Delta/64R$.
for *c in P* **do**
  Set the parameter Δ in Lemma 10 to be $\min(\Delta/3, \tau^2/48)$.
  Generate a list of T candidates $\mathcal{W}_c$ given by each step of the algorithm in Lemma 10.
Run Algorithm 3 with parameters $\alpha, \sigma, \cup_{c\in P}\mathcal{W}_c$ to get list $\mathcal{L}$.
Return $\mathcal{L}$.

5. Main Results

Finally, we state and prove our two main results. Our first result is a lower bound, demonstrating the necessity of our condition for efficient identifiability. Our second result is our algorithmic guarantee, demonstrating that if efficient identifiability holds, our algorithm returns a hypothesis achieving a small clean loss.

5.1. Necessity of the Identifiability Condition for Unique Recovery

Theorem 12 *Suppose GLM-Ob(g, σ, w^*) is not (Δ, τ)-identifiable, and suppose $u, v \in \mathbb{R}^d$ and $A \in \mathbb{R}$ witness this, i.e. u, v are Δ-separated but satisfy $\Pr_x[|g(u \cdot x) - g(v \cdot x) - A| > \tau] \leq \tau$. Then any algorithm that distinguishes between u and v with probability at least $1 - \delta$ requires $m = \Omega(\min(\sigma, 1)\ln(1/\delta)/\tau)$ samples.*

Proof Given A and τ, consider the event E defined by $|g(u \cdot x_i) - g(v \cdot x_i) - A| > \tau$. This occurs with probability $\leq \tau$. A single sample observed in event E can be enough to tell the difference between u and v, and so, to distinguish between u and v with a probability of at least $1 - \delta$, one must observe $\Omega(\ln(1/\delta)/\tau)$ samples from E.

If no samples from E are observed, then all (x_i, y_i) satisfy $|g(u \cdot x_i) - g(v \cdot x_i) - A| \leq \tau$. In this case, an oblivious noise adversary can construct oblivious noises ξ_u, ξ_v for instances of u, v such that the corrupted labels $g(u \cdot x_i) + \xi_u$ and $g(v \cdot x_i) + \xi_v$ only differ by at most τ. This means that y_i can either be generated from $g(u \cdot x_i) + \xi_u + \epsilon$ or $g(v \cdot x_i) + \xi_v + \epsilon$, which are close to each other in total variation distance. By Fact 20, any algorithm to distinguish u and v using inliers requires at least $\Omega(\sigma\ln(1/\delta)/\tau)$ samples. The lower bound corresponds to the minimum of the two sample complexities, so any algorithm to distinguish u and v with probability at least $1 - \delta$ needs $\Omega(\min(\sigma, 1)\ln(1/\delta)/\tau)$ samples. ■

5.2. Main Algorithmic Result

Here, we state the formal version of Theorem 4. This follows from putting together Lemma 10 and Lemma 22, applied to Algorithm 2. We restate and prove this in

Theorem 13 (Main Result) *We first define a few variables and their relationships to Δ (the desired final accuracy), α (the probability of being an inlier), R (an upper bound on $\|w^*\|$) and σ (the standard deviation of the additive Gaussian noise).*

Let $\Delta' = \min(\Delta, \tau^2/16)$. $\gamma = \min(\Delta/4\sigma, 1/2)$, $T \gtrsim (R/\gamma\alpha)^2$, $m_1 \gtrsim R^2\ln(T/\delta)/(\gamma\alpha\Delta)^2$ and $W \gtrsim T(\gamma\alpha\Delta/64R)$.

There is an algorithm, which, given Δ, α, R and σ runs in time $O(dTm_1)$, draws $m_1 \gtrsim \alpha^{-2}\log(R/\Delta\alpha\delta)\left(R^2\sigma^2/\Delta^4\right)$ samples from GLM-Ob(g,σ,w^*) and returns a $T(\gamma\alpha\Delta/64R)$-sized list of candidates, one of which achieves excess loss at most Δ.

Moreover, if the instance is (Δ,τ)-identifiable then, there is an algorithm which takes the parameters $\Delta, \alpha, \sigma, R$ and $\tau' \le \tau$, draws

$$m \gtrsim \alpha^{-2}\log(W/\delta)\left(R^2\sigma^2/(\Delta'^4 + 1/(\tau'\min(\tau'/\sigma,1))^2\right)$$

samples from GLM-Ob(g,σ,w^*), runs in time $O(dmW^2)$ and returns a single candidate.

References

P. Auer. Learning nested differences in the presence of malicious noise. *Theoretical Computer Science*, 185(1):159–175, 1997.

P. Auer, M. Herbster, and M. K. K Warmuth. Exponentially many local minima for single neurons. In D. Touretzky, M.C. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1995. URL https://proceedings.neurips.cc/paper/1995/file/3806734b256c27e41ec2c6bffa26d9e7-Paper.pdf.

K. Bhatia, P. Jain, and P. Kar. Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, pages 721–729, 2015.

H. Chen and T. d'Orsi. On the well-spread property and its relation to linear regression. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 3905–3935. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/chen22d.html.

S. Chen, F. Koehler, A. Moitra, and M. Yau. Classification under misspecification: Halfspaces, generalized linear models, and evolvability. *Advances in Neural Information Processing Systems*, 33:8391–8403, 2020a.

S. Chen, F. Koehler, A. Moitra, and M. Yau. Online and distribution-free robustness: Regression and contextual bandits with huber contamination. *arXiv preprint arXiv:2010.04157*, 2020b.

A. Dalalyan and P. Thompson. Outlier-robust estimation of a sparse linear model using ℓ_1-penalized huber's m-estimator. *Advances in neural information processing systems*, 32, 2019.

I. Diakonikolas and D. Kane. Near-optimal statistical query hardness of learning halfspaces with massart noise. In *Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 4258–4282. PMLR, 2022.

I. Diakonikolas, S. Goel, S. Karmalkar, A. R. Klivans, and M. Soltanolkotabi. Approximation schemes for relu regression. In *Conference on Learning Theory, COLT 2020*, volume 125 of *Proceedings of Machine Learning Research*, pages 1452–1485. PMLR, 2020a.

I. Diakonikolas, D. Kane, and N.Zarifis. Near-optimal SQ lower bounds for agnostically learning halfspaces and relus under gaussian marginals. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020b.

I. Diakonikolas, D. M. Kane, T. Pittas, and N. Zarifis. The optimality of polynomial regression for agnostic learning under gaussian marginals in the sq model. *Proceedings of Machine Learning Research vol*, 134:1–33, 2021a.

I. Diakonikolas, J. H. Park, and C. Tzamos. Relu regression with massart noise. *Advances in Neural Information Processing Systems*, 34:25891–25903, 2021b.

I. Diakonikolas, D. Kane, P. Manurangsi, and L. Ren. Hardness of learning a single neuron with adversarial label noise. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2022*, volume 151 of *Proceedings of Machine Learning Research*, pages 8199–8213. PMLR, 2022a. URL https://proceedings.mlr.press/v151/diakonikolas22a.html.

I. Diakonikolas, D. Kane, L. Ren, and Y. Sun. SQ lower bounds for learning single neurons with massart noise. In *NeurIPS*, 2022b. URL http://papers.nips.cc/paper_files/paper/2022/hash/97b983c974551153d20ddfabb62a5203-Abstract-Conference.html.

I. Diakonikolas, D. M. Kane, P. Manurangsi, and L. Ren. Cryptographic hardness of learning halfspaces with massart noise. *CoRR*, abs/2207.14266, 2022c. doi: 10.48550/arXiv.2207.14266. URL https://doi.org/10.48550/arXiv.2207.14266. Conference version in NeurIPS'22.

I. Diakonikolas, V. Kontonis, C. Tzamos, and N. Zarifis. Learning a single neuron with adversarial label noise via gradient descent. In *Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 4313–4361. PMLR, 2022d. URL https://proceedings.mlr.press/v178/diakonikolas22c.html.

I. Diakonikolas, D. M. Kane, and L. Ren. Near-optimal cryptographic hardness of agnostically learning halfspaces and relu regression under gaussian marginals. *CoRR*, abs/2302.06512, 2023. doi: 10.48550/arXiv.2302.06512. URL https://doi.org/10.48550/arXiv.2302.06512. Conference version in ICML'23.

T. d'Orsi, C. H. Liu, R. Nasser, G. Novikov, D. Steurer, and S. Tiegel. Consistent estimation for pca and sparse regression with oblivious outliers. *Advances in Neural Information Processing Systems*, 34:25427–25438, 2021a.

T. d'Orsi, G. Novikov, and D. Steurer. Consistent regression when oblivious outliers overwhelm. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 2297–2306. PMLR, 2021b. URL http://proceedings.mlr.press/v139/d-orsi21a.html.

T. d'Orsi, R. Nasser, G. Novikov, and D. Steurer. Higher degree sum-of-squares relaxations robust against oblivious outliers. *CoRR*, abs/2211.07327, 2022. doi: 10.48550/arXiv.2211.07327. URL https://doi.org/10.48550/arXiv.2211.07327.

S. Goel, S. Karmalkar, and A. R. Klivans. Time/accuracy tradeoffs for learning a relu with respect to gaussian marginals. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, pages 8582–8591, 2019.

S. Goel, A. Gollakota, and A. R. Klivans. Statistical-query lower bounds via functional gradients. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020.

M. Hardt and A. Moitra. Algorithms and hardness for robust subspace recovery. In *Proc. 26th Annual Conference on Learning Theory (COLT)*, pages 354–375, 2013.

E. Hazan. Introduction to online convex optimization. *Found. Trends Optim.*, 2(3–4):157–325, aug 2016. ISSN 2167-3888. doi: 10.1561/2400000013. URL https://doi.org/10.1561/2400000013.

S. M. Kakade, V. Kanade, O. Shamir, and A. Kalai. Efficient learning of generalized linear and single index models with isotonic regression. *Advances in Neural Information Processing Systems*, 24, 2011.

A. T. Kalai and R. Sastry. The isotron algorithm: High-dimensional isotonic regression. In *COLT*, 2009.

S. M. M. Kalan, M. Soltanolkotabi, and S. Avestimehr. Fitting relus via sgd and quantized sgd. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 2469–2473. IEEE, 2019.

S. Karmakar, A. Mukherjee, and R. Muthukumar. A study of neural training with iterative non-gradient methods. *arXiv e-prints*, pages arXiv–2005, 2020.

A. Klivans and R. Meka. Learning graphical models using multiplicative weights. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 343–354. IEEE, 2017.

P. Manurangsi and D. Reichman. The computational complexity of training relu (s). *arXiv preprint arXiv:1810.04207*, 2018.

N. Nasrabadi, T. Tran, and N. Nguyen. Robust lasso with missing and grossly corrupted observations. *Advances in Neural Information Processing Systems*, 24, 2011.

R. Nasser and S. Tiegel. Optimal SQ lower bounds for learning halfspaces with massart noise. In *Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 1047–1074. PMLR, 2022. URL https://proceedings.mlr.press/v178/nasser22a.html.

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

T. Norman, N. Weinberger, and K. Y. Levy. Robust linear regression for general feature distribution. *arXiv preprint arXiv:2202.02080*, 2022.

S. Pesme and N. Flammarion. Online robust regression via sgd on the l1 loss. *Advances in Neural Information Processing Systems*, 33:2540–2552, 2020.

M. Soltanolkotabi. Learning relus via gradient descent. In *Advances in Neural Information Processing Systems*, pages 2007–2017, 2017.

A. S. Suggala, K. Bhatia, P. Ravikumar, and P. Jain. Adaptive hard thresholding for near-optimal consistent robust regression. In *Conference on Learning Theory*, pages 2892–2897. PMLR, 2019.

E. Tsakonas, J. Jaldén, N. D. Sidiropoulos, and B. Ottersten. Convergence of the huber regression m-estimate in the presence of dense outliers. *IEEE Signal Processing Letters*, 21(10):1211–1214, 2014. doi: 10.1109/LSP.2014.2329811.

P. Wang, N. Zarifis, I. Diakonikolas, and J. Diakonikolas. Robustly learning a single neuron via sharpness. *CoRR*, abs/2306.07892, 2023. doi: 10.48550/arXiv.2306.07892.

J. Wright and Y. Ma. Dense error correction via ℓ_1-minimization. *IEEE Transactions on Information Theory*, 56(7):3540–3560, 2010.

G. Yehudai and S. Ohad. Learning a single neuron with gradient methods. In *Conference on Learning Theory*, pages 3756–3786. PMLR, 2020.

Appendix A. Concentration and Anti-Concentration

Lemma 14 (Hoeffding) *Let $X_1, \dots X_n$ be independent random variables such that $X_i \in [a_i, b_i]$. Then $S_n := \frac{1}{n}\sum_{i=1}^n X_i$, then for all $t > 0$*

$$\Pr[|S_n - \mathbf{E}[S_n]| \geq t] \leq \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

Lemma 15 (Empirical Separating Hyperplane) *Let $(x_i, y_i)_{i=1}^m \sim$ GLM-Ob$(g, \sigma, w^*)^m$ where $m \gtrsim R^2 \ln(1/\delta)/(\gamma\alpha\Delta)^2$. Assume c satisfies the assumption in Lemma 7. Define $\widehat{H}_c(w) := (1/m)\sum_{i=1}^m [(\mathrm{sign}(g(w \cdot x_i) - y_i) - c)\ x_i]$. Then for any w,*

$$\widehat{H}_c(w) \cdot (w - w^*) \geq (\gamma\alpha/4)\, \mathop{\mathbf{E}}_x [|(g(w^* \cdot x) - g(w \cdot x))|] - \gamma^2\alpha\sigma/4 - 3\,(\gamma\alpha\Delta/32)$$

with probability at least $1 - \delta$.

Proof From Lemma 7, we know that $H_c(w) \cdot (w - w^*) \geq (\gamma\alpha/4)\, \mathbf{E}_x [|(g(w^* \cdot x) - g(w \cdot x))|] - \gamma^2\alpha\sigma/2$ where $H_c(w) := \mathbf{E}_{x,y} [(\mathrm{sign}(y - g(w \cdot x)) - c)\ x]$. Consider the random variable given by $H_c(w) \cdot v - \widehat{H}_c(w) \cdot v$ for any fixed vector v. Upon examination, we can see that the quantity $(\mathrm{sign}(g(x \cdot x_i) - y_i) - c)\, x_i \cdot v$ has bounded absolute value at most $2\|v\| \leq 4R$ because $|c| \leq 1$. Then the concentration follows from a simple application of Hoeffding's inequality (Lemma 14).

$$\Pr\left[H_c(w) \cdot v - \widehat{H}_c(w) \cdot v \geq t\right] \leq \exp\left(-\frac{mt^2}{8R^2}\right)$$

Then, setting $v = w - w^*$, $t = \gamma\alpha\Delta/32$ and $m = C\left(R^2 \ln(1/\delta)/(\gamma\alpha\Delta)^2\right)$ for some large enough constant C, we have that

$$\widehat{H}_c(w) \cdot (w - w^*) \geq (\gamma\alpha/4)\, \mathop{\mathbf{E}}_x [|(g(w^* \cdot x) - g(w \cdot x))|] - \gamma^2\alpha\sigma/4 - 3\,(\gamma\alpha\Delta/32)\,.$$

■

Appendix B. Proofs of Basic Facts

Fact 16 *Given estimates $\widehat{a}, \widehat{b}$ of quantities a, b satisfying $0 \leq a \leq 1$, $L \leq b \leq 1$ and $|\widehat{a} - a| \leq e$ and $|\widehat{b} - b| \leq e$ where $e \leq L/2$, the quotient $\widehat{a}/\widehat{b}$ satisfies $|(\widehat{a}/\widehat{b}) - (a/b)| \leq 8e/L^2$.*

Proof We see that $(a/b) - (\widehat{a}/\widehat{b}) \leq (a/b) - (a - e/(b + e)) \leq (ea + eb)/(b(b - e)) \leq 8e/L^2$. The other direction follows by a similar argument. ■

Fact 17 *Let $\epsilon \sim \mathcal{N}(0, \sigma^2)$, then $h_\sigma(t) : \mathbb{R} \to \mathbb{R}$ defined as $h_\sigma(t) := \mathbf{E}_\epsilon[\mathrm{sign}(t + \epsilon)]$ satisfies:*

1. *$h_\sigma(-t) = -h_\sigma(t)$.*
2. *$h_\sigma(t)$ is strictly increasing.*
3. *$|h_\sigma(t)| \leq 1$.*

4. *For every $\tau < 2$, For all $t \notin [-\tau\sigma, \tau\sigma]$, $|h_\sigma(t)| \geq (\tau/4)$, and whenever $|t| \leq \tau\sigma$, $|h_\sigma(t)| \geq (1/4)(t/\sigma)$.*

Proof Suppose $\sigma \neq 0$, if $\sigma = 0$ these properties follow from properties of the sign function.

The first three follow easily from the fact that

$$h_\sigma(t) = \Pr_\epsilon[t + \epsilon > 0] - \Pr_\epsilon[t + \epsilon \leq 0] = \text{sign}(t) \Pr_\epsilon[-|t| \leq \epsilon \leq |t|] = \text{sign}(t)\,(1 - 2\Pr_\epsilon[\epsilon > |t|]).$$

To see the final property, observe that

$$h_\sigma(t) = \text{sign}(t) \Pr_\epsilon[-|t| \leq \epsilon \leq |t|] = \text{sign}(t) \Pr_{x \sim \mathcal{N}(0,1)}[-t/\sigma \leq x \leq t/\sigma].$$

By Gaussian anticoncentration, whenever $t/\sigma < 2$, $\Pr_{x\sim\mathcal{N}(0,1)}[-t/\sigma \leq x \leq t/\sigma] \geq (t/4\sigma)$ proving the second part of this claim. Since h is strictly increasing, we see that whenever $|t| > \tau\sigma$ and $\tau < 2$ $|h_\sigma(t)| \geq \tau/4$, proving the first part of the claim. ■

Fact 18 *Let ξ be oblivious noise such that $\Pr[\xi = 0] \geq \alpha$, then $F_{\sigma,\xi}(t) := \mathbf{E}_{\epsilon,\xi}[\text{sign}(t + \epsilon + \xi)] - \mathbf{E}_{\epsilon,\xi}[\text{sign}(\epsilon + \xi)]$ satisfies the following:*

1. *$F_{\sigma,\xi}$ is strictly increasing.*
2. *$\text{sign}(F_{\sigma,\xi}(t)) = \text{sign}(t)$.*
3. *For any $\tau \leq 2$, Whenever $|t| \geq \sigma\tau$, $|F_{\sigma,\xi}(t)| > (\tau\alpha/4)$ and whenever $|t| \leq \sigma\tau$, $|F_{\sigma,\xi}(t)| \leq (\alpha/4)(t/\sigma)$*

Proof The first property follows from the fact that if $t_1 - t_2 > 0$ then $F_{\sigma,\xi}(t_1) - F_{\sigma,\xi}(t_2) = \mathbf{E}_\xi\left[h_\sigma(t_1 + \xi) - h_\sigma(t_2 + \xi)\right] > 0$. Hence $F_{\sigma,\xi}(t)$ is strictly increasing.

The second property follows by definition and the first property, $F_{\sigma,\xi}(0) = 0$ and since $F_{\sigma,\xi}$ is strictly increasing, $\text{sign}(F_{\sigma,\xi}(t)) = \text{sign}(t)$.

Note that h_σ is strictly increasing. Let $a > c\sigma$.

$$\begin{aligned} F_{\sigma,\xi}(a) &= \mathop{\mathbf{E}}_{\xi}\left[h_\sigma(a + \xi) - h_\sigma(\xi)\right] \\ &> \alpha \mathop{\mathbf{E}}_{\xi|\xi=0}\left[h_\sigma(a) - h_\sigma(0)\right] \\ &= \alpha h_\sigma(a)\ . \end{aligned}$$

The first inequality above follows from the fact that h_σ is montone and $a > 0$. The final property above now follows from Property 4 of Fact 17. A similar argument holds when $a < -c\sigma$. ■

Remark 19 Note that a similar result holds for other distributions as long as the measurement noise has some density around the origin. More precisely, if $\Pr_\epsilon[|\epsilon| \leq \sigma] \geq C$ then $\Pr_{\epsilon,\xi}[|\epsilon + \xi| \leq \sigma] \geq \alpha C$. This implies that $F_{\sigma,\xi}(t)$ as defined above satisfies $|F_{\sigma,\xi}(t)| \geq C\alpha$ whenever $|t| \geq \sigma$, for ϵ satisfying the constraint $\Pr_\epsilon[|\epsilon| \leq \sigma] \geq C$. Hence, the Gaussianity of our observation noise is not crucial, but the noise needs to have some density around the origin. Indeed, the oblivious noise is free to incorporate any other distribution as well.

Fact 20 *Let* $\mathbf{p}$ *and* $\mathbf{q}$ *be univariate probability distributions on* $\mathbb{R}$ *and denote total variation distance as* d_{TV}*. Then any algorithm requires* $\Omega(\ln(1/\delta)/d_{TV}(\mathbf{p},\mathbf{q}))$ *samples to successfully distinguish between* $\mathbf{p},\mathbf{q}$ *with probability* $1-\delta$*.*

Fact 21 *Let* X *be a random variable on* $\mathbb{R}$*. Fix* $\tau>0$ *and* $\eta>0$*. Define the events* E_A^+ *and* E_A^- *such that* $\Pr[E_A^+]=\Pr[X>A+\tau]$ *and* $\Pr[E_A^-]=\Pr[X<A-\tau]$*. Then if the first condition below is not true, the second is.*

1. $\exists A\in\mathbb{R}$ *such that* $\Pr[E_A^+]\geq\eta$ *and* $\Pr[E_A^-]\geq\eta$*.*
2. $\exists A^*\in\mathbb{R}$ *such that* $\Pr[E_{A^*}^+]\leq\eta$ *and* $\Pr[E_{A^*}^-]\leq\eta$*.*

Proof Assume the first statement is false. Then the negation implies that $\forall A\in\mathbb{R}$, either $\Pr[E_A^+]\leq\eta$ or $\Pr[E_A^-]\leq\eta$. We want to show that the "or" statement translates to an "and" statement for a particular A^*.

Note that $\Pr[E_A^-]$ as a function of A can be seen as the CDF of X without the equality portion where $X=A-\tau$. This means that $\Pr[E_A^-]$ is left-continuous with respect to A. Therefore, we can define A^* such that $\forall A\in(-\infty,A^*]$, $\Pr[E_A^-]\leq\eta$ and for any $A>A^*$, $\Pr[E_A^-]>\eta$. Then, by our initial assumption, it must be the case that $\forall A\in(A^*,\infty)$, $\Pr[E_A^+]\leq\eta$. In contrast to $\Pr[E_A^-]$ being left-continuous, we can infer that $\Pr[E_A^+]$ as a function of A is right-continuous with respect to A. Therefore by right-continuity, $\Pr[E_{A^*}^+]\leq\eta$. This proves the existence of such A^* of the second condition and concludes the proof. ■

Appendix C. Pruning Implausible Solutions

Algorithm 3 Prune Implausible Candidates

input: $\alpha,\sigma,R,\mathcal{W}=\{w_1,\ldots,w_p\},\tau$
Draw $m=C\log(|\mathcal{W}|^2/\delta)/(\alpha\tau(\min\{\tau/2\sigma,1\}))^2$ samples $\{(x_k,y_k)\}_{k=1}^m$ for some constant C.
for $i\leftarrow 1...p$ **do**
  for $j\leftarrow i+1...p$ **do**
    Let $E_A^+:=\{x_k|g(w_i\cdot x_k)-g(w_j\cdot x_k)>A\}$ and $E_A^-:=\{x_k|g(w_i\cdot x_k)-g(w_j\cdot x_k)<A\}$
    Compute the range of A such that $|E_{A+\tau/2}^+|\geq\alpha m\min\{\tau/2\sigma,1/4\}$ via binary search on at most m distinct $g(w_i\cdot x_k)-g(w_j\cdot x_k)-\tau/2$ and similarly for $|E_{A-\tau/2}^-|$
    Let $A\leftarrow$ any number in the intersection of two ranges
    if *no such* A *exists* **then**
      continue to $(j+1)$-th inner loop
    Compute $R^+=\{r|r=y_i-g(w_i\cdot x)$ for $x\in E_{A+\tau/2}^+\}$ and similarly R^- for $E_{A-\tau/2}^-$
    if $|\widehat{\mathbf{E}}[\mathrm{sign}(r-A)\mid r\in R^+]-\widehat{\mathbf{E}}[\mathrm{sign}(r-A)\mid r\in R^-]|>\alpha\min\{\tau/16\sigma,1/8\}$ **then**
      reject w_i and continue with $(i+1)$-th outer loop
for $i\leftarrow 1\ldots p$ **do**
  for $j\leftarrow 1\ldots p$ **do**
    if $\frac{1}{m}\sum_{t=1}^m|g(w_i\cdot x_t)-g(w_j\cdot x_t)|>3\Delta$ **then**
      return $\mathcal{W}$
Sample $\widehat{w}$ uniformly from $\mathcal{W}$.
return $\{\widehat{w}\}$.

Lemma 22 (Pruning bad candidates) *Let $\delta > 0$. Suppose $\exists \widehat{w} \in \mathcal{W}$ such that*

$$\mathop{\mathbf{E}}_{x}\left[|g(\widehat{w}\cdot x) - g(w^*\cdot x)|\right] \le \min\{\Delta, \tau^2/16\}.$$

Then Algorithm 3 draws $m \gtrsim \log(|\mathcal{W}|^2/\delta)/(\alpha\tau(\min\{\tau/\sigma, 1\}))^2 + R^2\log(|\mathcal{W}|^2/\delta)/\Delta^2$ samples, runs in time $\tilde{O}(dm|\mathcal{W}|^2)$, and with probability $1-\delta$ returns a list of candidates containing $\widehat{w}$ such that each candidate satisfies $\Pr[|g(w^*\cdot x) - g(w\cdot x) - A_w| > \tau] \le 1-\tau$ for some $A_w \in \mathbb{R}$. If (Δ,τ)-identifiability (Definition 2) holds, the algorithm only returns a single candidate $\widehat{w}$ which achieves a clean loss of 4Δ.

Proof Define the events $E_s^+(\widehat{w}, w) := \{x \mid g(\widehat{w}\cdot x) - g(w\cdot x) > s\}$ and $E_s^-(\widehat{w}, w) := \{x \mid g(\widehat{w}\cdot x) - g(w\cdot x) < s\}$. An application of Fact 6 to the random variable $g(\widehat{w}\cdot x) - g(w\cdot x)$ implies that for any τ_0, η_0 if the first condition below is not true, the second is.

1. $\exists A \in \mathbb{R}$ such that $\Pr[E^+_{A+\tau_0}(\widehat{w}, w)] \ge \eta_0$ and $\Pr[E^-_{A-\tau_0}(\widehat{w}, w)] \ge \eta_0$.
2. $\exists A \in \mathbb{R}$ such that $\Pr[E^+_{A+\tau_0}(\widehat{w}, w)] \le \eta_0$ and $\Pr[E^-_{A-\tau_0}(\widehat{w}, w)] \le \eta_0$.

In the first part of our proof, we show that the algorithm rejects w if Condition 1 holds. We will need the following lemma about R^+ and R^- as defined in our algorithm.

Claim 23 *Suppose $C := \{x \mid |g(\widehat{w}\cdot x) - g(w^*\cdot x)| \le \tau_0/2\}$. Then for any choice of τ_0 and η_0, if Condition 1 holds, there is a choice of $\delta' = 2\Delta/\tau_0$ satisfying,*

1. $\Pr[C \mid E^+_{A+\tau_0}(\widehat{w}, w)] \ge 1 - \delta'/\eta_0$.
2. $\max\{\Pr[\overline{C} \mid E^+_{A+\tau_0}(\widehat{w}, w)], \Pr[\overline{C} \mid E^-_{A-\tau_0}(\widehat{w}, w)]\} \le \delta'/\eta_0$.
3. $x \in E^+_{A+\tau_0}(\widehat{w}, w) \cap C$ *implies* $\operatorname{sign}(y(x) - g(w\cdot x) - A) \ge \operatorname{sign}(\epsilon + \xi + \tau_0/2)$ *and* $x \in E^-_{A-\tau_0}(\widehat{w}, w) \cap C$ *implies* $\operatorname{sign}(y(x) - g(w\cdot x) - A) \le \operatorname{sign}(\epsilon + \xi - \tau_0/2)$.

Proof Since Condition 1 holds, $\Pr[E^+_{A+\tau_0}(\widehat{w}, w)] \ge \eta_0$ and $\Pr[E^-_{A-\tau_0}(\widehat{w}, w)] \ge \eta_0$.

We now lower bound the probability of C. By assumption, $\mathbf{E}_x\left[|g(\widehat{w}\cdot x) - g(w^*\cdot x)|\right] \le \Delta$. An application of Markov's inequality implies $\Pr[|g(\widehat{w}\cdot x) - g(w^*\cdot x)| \ge \tau_0/2] \le 2\Delta/\tau_0$. Choosing $2\Delta/\tau_0 = \delta'$ implies $\Pr[C] \ge 1 - \delta'$.

The first property now follows from the fact that $\Pr[E^+_{A+\tau_0}(\widehat{w}, w) \cap C] \ge \Pr[E^+_{A+\tau_0}(\widehat{w}, w)] - \delta'$. Finally, Bayes rule and the fact that $\Pr[E^+_{A+\tau_0}(\widehat{w}, w)] \ge \eta_0$, implies $\Pr[C \mid E^+_{A+\tau_0}(\widehat{w}, w)] \ge 1 - \delta'/\Pr[E^+_{A+\tau_0}(\widehat{w}, w)] \ge 1 - \delta'/\eta_0$.

The second property follows from the Bayes rule, $\Pr[E^+_{A+\tau_0}(\widehat{w}, w)] \ge \eta_0$ and $\Pr[E^-_{A-\tau_0}(\widehat{w}, w)] \ge \eta_0$, and the fact that $\Pr[\overline{C}] \le \delta'$.

For $x \in E^+_{A+\tau_0}(\widehat{w}, w) \cap C$, the third property follows from the fact that $\operatorname{sign}(\cdot)$ is monotonically increasing and the fact that if $g(\widehat{w}\cdot x) - g(w\cdot x) - A > \tau_0$ and $|g(\widehat{w}\cdot x) - g(w^*\cdot x)| \le \tau_0/2$, then $g(w^*\cdot x) - g(w\cdot x) - A > \tau_0/2$. A similar argument for the case when $x \in E^-_{A-\tau_0}(\widehat{w}, w) \cap C$ proves our result.

■

Let $R^+ := \{y_i - g(w\cdot x_i) \mid x_i \in E^+_{A+\tau_0}(\widehat{w}, w)\}$ and $R^- := \{y_i - g(w\cdot x_i) \mid x_i \in E^-_{A-\tau_0}(\widehat{w}, w)\}$ for a specific choice of τ_0. Our algorithm rejects w if there is an A such that

$|\widehat{\mathbf{E}}[\text{sign}(r-A) \mid r \in R^+] - \widehat{\mathbf{E}}[\text{sign}(r-A) \mid r \in R^-]| \geq \alpha \min\{\tau_0/8\sigma, 1/8\}$. An application of the properties from Claim 23 shows us that if Condition 1 holds for w, then this is indeed the case for the true distribution.

$$
\begin{aligned}
&|\mathbf{E}[\text{sign}(g(w^* \cdot x) - g(w \cdot x) + \xi + \epsilon - A) \mid x \in E^+_{A+\tau_0}(\widehat{w}, w)] \\
&\qquad - \mathbf{E}[\text{sign}(g(w^* \cdot x) - g(w \cdot x) + \xi + \epsilon - A) \mid x \in E^-_{A-\tau_0}(\widehat{w}, w)]| \\
&= |\Pr[C \mid E^+_{A+\tau_0}(\widehat{w}, w)] \ \mathbf{E}[\text{sign}(g(w^* \cdot x) - g(w \cdot x) + \xi + \epsilon - A) \mid x \in E^+_{A+\tau_0}(\widehat{w}, w) \cap C] \\
&\qquad + \Pr[\overline{C} \mid E^+_{A+\tau_0}(\widehat{w}, w)] \ \mathbf{E}[\text{sign}(g(w^* \cdot x) - g(w \cdot x) + \xi + \epsilon - A) \mid x \in E^+_{A+\tau_0}(\widehat{w}, w) \cap \overline{C}] \\
&\qquad - \Pr[C \mid E^-_{A-\tau_0}(\widehat{w}, w)] \ \mathbf{E}[\text{sign}(g(w^* \cdot x) - g(w \cdot x) + \xi + \epsilon - A) \mid x \in E^-_{A-\tau_0}(\widehat{w}, w) \cap C] \\
&\qquad - \Pr[\overline{C} \mid E^-_{A-\tau_0}(\widehat{w}, w)] \ \mathbf{E}[\text{sign}(g(w^* \cdot x) - g(w \cdot x) + \xi + \epsilon - A) \mid x \in E^-_{A-\tau_0}(\widehat{w}, w) \cap \overline{C}]| \\
&\geq |(1 - \delta'/\eta_0) \ \mathbf{E}[\text{sign}(g(w^* \cdot x) - g(w \cdot x) + \xi + \epsilon - A) \mid x \in E^+_{A+\tau_0}(\widehat{w}, w) \cap C] \\
&\qquad - \ \mathbf{E}[\text{sign}(g(w^* \cdot x) - g(w \cdot x) + \xi + \epsilon - A) \mid x \in E^-_{A-\tau_0}(\widehat{w}, w) \cap C]| \\
&\qquad - 2\delta'/\eta_0 \\
&\geq |\mathbf{E}[\text{sign}(g(w^* \cdot x) - g(w \cdot x) + \xi + \epsilon - A) \mid x \in E^+_{A+\tau_0}(\widehat{w}, w) \cap C] \\
&\qquad - \mathbf{E}[\text{sign}(g(w^* \cdot x) - g(w \cdot x) + \xi + \epsilon - A) \mid x \in E^-_{A-\tau_0}(\widehat{w}, w) \cap C]| - 3\delta'/\eta_0 \\
&\geq |\mathbf{E}[\text{sign}(\xi + \epsilon + \tau_0/2) \mid x \in E^+_{A+\tau_0}(\widehat{w}, w) \cap C] \\
&\qquad - \mathbf{E}[\text{sign}(\xi + \epsilon - \tau_0/2) \mid x \in E^-_{A-\tau_0}(\widehat{w}, w) \cap C]| - 3\delta'/\eta_0 \\
&= |\mathbf{E}[\text{sign}(\xi + \epsilon + \tau_0/2)] - \mathbf{E}[\text{sign}(\xi + \epsilon - \tau_0/2)]| - 3\delta'/\eta_0 \\
&= 2\Pr[|\xi + \epsilon| \leq \tau_0/2] - 3\delta'/\eta_0 \\
&\geq 2\alpha \Pr[|\epsilon| \leq \tau_0/2] - 3\delta'/\eta_0 \\
&\geq \alpha \min\{\tau_0/4\sigma, 1/4\}.
\end{aligned}
$$

The final inequality follows by setting $\delta' < \alpha\eta_0 \min\{\tau/12\sigma, 1/12\})$, and the fact that whenever $\tau_0 < 2$, $\Pr[|\epsilon| \leq \tau_0/2] \geq \min\{\tau_0/2\sigma, 1/2\}$. We will estimate $\mathbf{E}[\text{sign}(g(w^* \cdot x) - g(w \cdot x) + \xi + \epsilon - A) \mid x \in E^+_{A+\tau_0}(\widehat{w}, w)] - \mathbf{E}[\text{sign}(g(w^* \cdot x) - g(w \cdot x) + \xi + \epsilon - A) \mid r \in E^-_{A-\tau_0}(\widehat{w}, w)]$ upto an error of $\alpha \min\{\tau_0/8\sigma, 1/8\}$. For a fixed $\widehat{w}, w^*$ and w, this follows by estimating $\Pr[E^+_{A+\tau_0}(\widehat{w}, w)]$ and $\mathbf{E}[\text{sign}(g(w^* \cdot x) - g(w \cdot x) + \xi + \epsilon - A)\mathbf{1}(x \in E^+_{A+\tau_0}(\widehat{w}, w))]$ (and the corresponding E^- terms) each to an accuracy of $\eta_0^2 \alpha \min\{\tau_0/64\sigma, 1/64\}$. Since both of these are expectations of random variables bounded by one, Hoeffding's Lemma (Lemma 14) implies that $(64/\alpha^2\eta_0^2(\min\{\tau_0/64\sigma, 1/64\})^2)\log(1/\delta)$ samples suffice to achieve this approximation with a probability of $1 - \delta$. Let $\widehat{\Pr}$ and $\widehat{\mathbf{E}}$ denote the empirical expectation and probability respectively, then an application of Fact 16 to $\Pr[E^+_{A+\tau_0}(\widehat{w}, w)]$, $\mathbf{E}[\text{sign}(g(w^* \cdot x) - g(w \cdot x) + \xi + \epsilon - A)\mathbf{1}(x \in E^+_{A+\tau_0}(\widehat{w}, w))]$ and their respective empirical estimates implies $|\widehat{\mathbf{E}}[\text{sign}(r-A) \mid r \in R^+] - \widehat{\mathbf{E}}[\text{sign}(r-A) \mid r \in R^+]| \leq 8\eta_0^2 \alpha \min\{\tau_0/64\sigma, 1/64\}/\eta_0^2 \leq \alpha \min\{\tau_0/8\sigma, 1/8\}$.

A union bound over all possible $\mathcal{W}$ candidates for w and $\widehat{w}$ tells us that a sample complexity of $(64/\alpha^2\eta_0^2(\min\{\tau_0/64\sigma, 1/64\})^2)\log(|\mathcal{W}|^2/\delta)$ suffices.

Suppose w is not rejected, then we know that Condition 2 holds. Another application of Markov's inequality similar to before gives us $\Pr[|g(\widehat{w} \cdot x) - g(w^* \cdot x)| \geq \tau_0/2] \leq 2\Delta/\tau_0 = \delta'$. Any x satisfying $|g(\widehat{w} \cdot x) - g(w^* \cdot x)| \leq \tau_0$ and $g(w^* \cdot x) - g(w \cdot x) - A > 2\tau_0$ must also satisfy $g(\widehat{w} \cdot x) - g(w \cdot x) - A > \tau_0$. This implies that $\Pr[E^+_{A+2\tau_0}(w^*, w)] \leq \eta_0 + \delta'$. A similar argument

shows that $\Pr[E^-_{A-2\tau_0}(w^*, w)] \le \eta_0 + \delta'$. By choosing $2\tau_0 = \tau$ and $\eta_0 + \delta' = \tau/2$, every hypothesis we return satisfies $\Pr[E^-_{A-\tau}(w^*, w)] \le \tau/2$ and $\Pr[E^+_{A+\tau}(w^*, w)] \le \tau/2$. The constraints on the variables are satisfied when $\eta_0 = \delta' = \tau/4$ and $\tau_0 = \tau/2$, which amounts to $\Delta < \tau^2/16$.

If (Δ, τ)-identifiability holds, every element w in the set of candidates that remains satisfies $\mathbf{E}_x[|g(w^* \cdot x) - g(w \cdot x)|] \le \Delta$. To check that this is the case, the algorithm tests if every pair of candidates u, v in $\mathcal{W}$ is at most 3Δ-close, i.e. $\widehat{\mathbf{E}}_x[|g(u \cdot x) - g(v \cdot x)|] \le 3\Delta$. If this is the case, we return any candidate in the set. Otherwise we get a polynomial sized list $\mathcal{L}$ with $\widehat{w} \in \mathcal{L}$.

■

Appendix D. Proof of Main Theorem

Here we state and prove our main theorem, which is a more detailed version of Theorem 4.

Theorem 24 (Main Result) *We first define a few variables and their relationships to Δ (the desired final accuracy), α (the probability of being an inlier), R (an upper bound on $\|w^*\|$) and σ (the standard deviation of the additive Gaussian noise).*

Let $\Delta' = \min(\Delta, \tau^2/16)$. $\gamma = \min(\Delta/4\sigma, 1/2)$, $T \gtrsim (R/\gamma\alpha)^2$, $m_1 \gtrsim R^2 \ln(T/\delta)/(\gamma\alpha\Delta)^2$ and $W \gtrsim T(\gamma\alpha\Delta/64R)$.

There is an algorithm, which, given Δ, α, R and σ runs in time $O(dTm_1)$, draws $m_1 \gtrsim \alpha^{-2}\log(R/\Delta\alpha\delta)\left(R^2\sigma^2/\Delta^4\right)$ samples from GLM-Ob(g, σ, w^*) and returns a $T(\gamma\alpha\Delta/64R)$-sized list of candidates, one of which achieves excess loss at most Δ.

Moreover, if the instance is (Δ, τ)-identifiable then, there is an algorithm which takes the parameters $\Delta, \alpha, \sigma, R$ and $\tau' \le \tau$, draws

$$m \gtrsim \alpha^{-2}\log(W/\delta)\left(R^2\sigma^2/(\Delta'^4 + 1/(\tau'\min(\tau'/\sigma, 1))^2\right)$$

samples from GLM-Ob(g, σ, w^*), runs in time $O(dmW^2)$ and returns a single candidate.

Proof Recall that P is a uniform partition of $[-1, 1]$ with granularity $p = \gamma\alpha\Delta/64R$. For each $c \in P$ we run the algorithm from Lemma 10 for $T \gtrsim (R/\gamma\alpha)^2$ steps where $\gamma = \sigma/4\Delta$. From the lemma, we know that when $|c - \mathbf{E}_{\xi,\epsilon}[\mathrm{sign}(\xi + \epsilon)]| \le p$ one of the candidates generated by the online gradient descent algorithm satisfies $\mathbf{E}[|g(w^* \cdot x) - g(\widehat{w} \cdot x)|] \le \Delta$.

For the second part of the theorem, we set $\Delta' = \min(\Delta, \tau^2/16)$ and run the OCO algorithm above to get a larger list of candidates, one of which achieves excess loss Δ'. Finally, we collect all $T/p \lesssim (1/\Delta')\,(R/\gamma\alpha)^3 = |\mathcal{W}|$ candidates and run our pruning algorithm Algorithm 3 on them. Then Lemma 22 returns a list satisfying our final guarentee.

Putting together the sample complexities of the lemmas, we see that for this second part,

$$m \gtrsim \log(|\mathcal{W}|^2/\delta)/(\alpha\tau(\min(\tau/2\sigma, 1)))^2 + R^2\ln(T/\delta)/(\gamma\alpha\Delta')^2\ .$$

■