On Robust Learning of Ising Models

Erik M. Lindgren
University of Texas at Austin
erikml@utexas.edu

Vatsal Shah
University of Texas at Austin
vatsalshah1106@utexas.edu

Yanyao Shen
University of Texas at Austin
shenyanyao@utexas.edu

Alexandros G. Dimakis
University of Texas at Austin
dimakis@austin.utexas.edu

Adam Klivans
University of Texas at Austin
klivans@cs.utexas.edu

Abstract

Ising Models are one of the most popular class of probability distributions with applications in wide ranging fields such as physics, engineering and finance. In this paper, we attempt to learn the underlying graphical model robustly in presence of adversarial corruptions. In this work, we establish new lower and upper bounds for robustly learning Ising models.

1 Introduction

Ising models are an important class of probability distributions that model simple dependencies between binary random variables. They have been used to model network behavior in various domains, such as social networks, biology, and game theory [Daskalakis et al., 2011, 2017, Ellison, 1993, Montanari and Saberi, 2010]. Recent work due to Klivans and Meka [2017] develops an algorithm with essentially optimal run-time and sample complexity for the problem of *structure learning* for Ising models. That is, given samples from the unknown Ising model, the algorithm recovers all of its edge weights with small error.

The main thrust of this paper is to understand whether structure learning for Ising models can be made robust; i.e., can we efficiently recover the Ising model if an adversary is corrupting some draws from the underlying distribution? In the strongest setting, the adversary is typically allowed to observe the entire dataset and replace a constant fraction η of samples with arbitrary values.

In this work, we establish new lower and upper bounds for robustly learning Ising models based on the sparsity λ of the model and the smallest absolute edge weight α. For the lower bounds, we construct two Ising models over different graphs. We show that if an adversary is allowed to corrupt even $\eta = \alpha \exp(-O(\lambda))$ fraction of samples then no algorithm can differentiate the two distributions.

We complement our lower bound by establishing a robustness guarantee of the Sparsitron algorithm of Klivans and Meka [2017]. We show that the Sparsitron algorithm is robust to an adversary who can arbitrarily corrupt a fraction $\eta = \alpha^2 \exp(-O(\lambda))$ of samples from the Ising model. The number of samples required is the same as in the uncorrupted case, specifically, on the order of exponential in the sparsity of the model and logarithmic in the dimension of the model.

1.1 Related Work

Ising Models Bresler [2015] was the first to establish tractable algorithms for learning sparse Ising models with sample complexity that, for a fixed sparsity, depends only on the logarithm of the number of variables. The dependence on the sparsity was improved from doubly exponential to exponential by Vuffray et al. [2016], Lokhov et al. [2018] using convex programming. Klivans and Meka [2017], improved the running time to be essentially optimal using multiplicative weights. They were able to generalize this approach to learn non-binary and higher order graphical models.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

Robust Estimation There are a number of classic techniques for robust estimation of low-dimensional distributions [Huber and Ronchetti, 2011, Hampel et al., 2011]. Diakonikolas et al. [2016] and Lai et al. [2016] were the first to propose tractable algorithms for robust *high-dimensional* estimation. Diakonikolas et al. [2018] considers robust learning of Bayesian networks with known graphical structure. In this work, they mention robust learning of Ising models as an interesting open problem. Additionally, Kapoor et al. [2018] considers robust bandit learning under the same adversarial model as what we consider for the experts problem.

2 Problem Setup

An Ising model is defined over a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $|\mathcal{V}| = n$. Vertices $v_i \in \mathcal{V}$ correspond to binary random variables $x_i \in \{-1, +1\}$ (alternatively known as spins). Edges $(v_i, v_j) \in \mathcal{E}$, also called as couplings, are denoted using non-zero real parameters A_{ij}. There are also an external field parameters θ_i for each vertex v_i that bias the variable towards a particular value. An Ising model distribution $\mathcal{D}$ is the distribution such that the probability of a spin configuration $\underline{x} = \{x_1, x_2, \ldots, x_n\}$ is given by:

$$\mathrm{P}(\underline{x}) = \frac{1}{Z} \exp\left(\sum_{(v_i,v_j)\in\mathcal{E}} A_{ij} x_i x_j + \sum_{v_i \in \mathcal{V}} \theta_i \right),$$

where Z is a normalization constant. The set of neighbors of node $v_i \in \mathcal{V}$ is described by $\partial v_i = \{v_j : (v_i, v_j) \in \mathcal{E}\}$. Accordingly, we define the width of the Ising model $\lambda = \max_{v_i} \left(\sum_{v_j \in \partial v_i} |A_{ij}| + |\theta_i| \right)$. To construct an estimator of the edge set that is able to reconstruct the original structure with high probability, we will make a few assumptions on the coupling intensity of the model.

A1: The smallest absolute edge weight is greater than α: $\min_{(v_i,v_j)\in\mathcal{E}} |A_{ij}| \geq \alpha$.

A2: The width of the Ising model is bounded by λ: $\max_{v_i} \left(\sum_{v_j \in \partial v_i} |A_{ij}| + |\theta_i| \right) \leq \lambda$.

A common variant of the width guarantee is that the graph G has maximum degree d, and the maximum absolute edge weight $\max_{ij} |A_{ij}| \leq \beta$. Note that in this case, we have the width is bounded by βd.

Now, that we have setup the notation for the Ising model, let us define the two primary contamination models we will consider.

Definition (Huber's η-contamination model)**.** Let $\mathcal{D}$ be a distribution on $\{-1, 1\}^n$. In Huber's η-contamination model, we receive i.i.d. samples from the distribution $(1-\eta)\mathcal{D} + \eta\mathcal{E}$, where $\mathcal{E}$ is an arbitrary distribution.

Definition (η-corrupted samples)**.** Let $\mathcal{D}$ be a distribution on $\{-1, 1\}^n$. We say that a collection of samples U is η-corrupted if they were created by the following process: Generate $m = |D|$ samples by drawing them i.i.d. from $\mathcal{D}$, then an adversary chooses an η-fraction of the samples and replaces them with arbitrary values on $\{-1, 1\}^n$.

In the corrupted model, the adversary can introduce dependencies between the observed data. Ignoring a technical issue[1] for simplicity, the corruption model is stronger than the contamination model. All of our achievability results hold in the corruption model and all of our impossibility results hold in the contamination model.

3 Inachievability Results

We will use the following well-known lemma in both of our subsequent proofs for inachievability (see, for example, Fact 2.3 of [Diakonikolas et al., 2016]). It essentially states that if two distribution are close in total variation distance, then they cannot be distinguished given contaminated samples. It can be proven using Farkas' lemma.

Lemma 1. *Given two distribution $\mathcal{D}_1$ and $\mathcal{D}_2$ such that the total variation distance $d_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_2) \leq \eta$, there exists distributions $\mathcal{E}_1$ and $\mathcal{E}_2$ such that $(1-\eta)\mathcal{D}_1 + \eta\mathcal{E}_1 = (1-\eta)\mathcal{D}_2 + \eta\mathcal{E}_2$.*

[1] Note that the number of corrupted samples in the second model is not random but with high probability for large enough n, the second model is stronger [Diakonikolas et al., 2016]

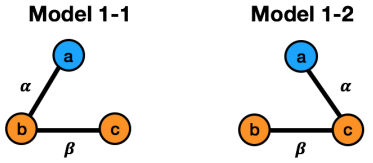

Figure 1: Graphs for the inachievability result in Theorem 2

We first establish that no algorithm can robustly learn an Ising model with bounded width λ when the fraction of contaminated samples is even exponentially small in λ.

Theorem 2. *For all λ and $\alpha > 0$, there exists two Ising models $\mathcal{D}_1$ and $\mathcal{D}_2$ such that both have width λ and minimum edge weight α such that given any number of η-contaminated samples with $\eta > \min\{\alpha, 1\}\exp(-2(\lambda - \alpha))$, then no algorithm with any number of samples can distinguish the two distributions.*

This theorem holds even in the weaker model where the adversary cannot corrupt real samples but can only inject an η-fraction of samples.

Proof. We will create two Ising models with different graph structure that can be completely confused when $\eta > \alpha\exp(-2\lambda)$. Both models are on three vertices a, b, c and have an edge bc with weight β. The first model (Model 1-1 in Fig. 1) has an additional edge ab of weight α, while the second model (Model 1-2 in Fig. 1) has an additional edge ac with weight α. Note that both models have width $\lambda = \alpha + \beta$.

The total variation distance d_{TV} between these models can be calculated as:

$$\begin{aligned} d_{\mathrm{TV}}(\tilde{\mathcal{D}}_1, \tilde{\mathcal{D}}_2) =& \frac{2(e^{-\beta+\alpha} - e^{-\beta-\alpha})}{2e^{\beta+\alpha} + 2e^{\beta-\alpha} + 2e^{-\beta-\alpha} + 2e^{-\beta+\alpha}} \\ =& \frac{1}{(e^{2\beta}+1)} \cdot \frac{e^{2\alpha}-1}{e^{2\alpha}+1} = \tanh(\alpha)\sigma(-2\beta) \leq \min\{\alpha, 1\}e^{-2\beta}. \end{aligned}$$

Thus if an adversary can inject $\min\{\alpha, 1\}\exp(-2\beta)$ fraction of samples, they can make samples from one model look like samples from the other model and vice-versa. □

One special case of interest is when the Ising model has a degree bound d and a bound on the maximum absolute edge weight β. For this special case, we establish a similar lower bound.

Theorem 3. *For all d, $\alpha > 0$ and $\beta > \frac{\ln 2}{3}$, there exists two Ising models $\mathcal{D}_1$ and $\mathcal{D}_2$ on different graphs such that both are d-sparse and have edges weights satisfying $\alpha \leq A_{uv} \leq \beta$ such that given any number of η-contaminated samples with $\eta > \min\{\alpha, 1\}e^{-C\beta d}$, no algorithm can distinguish the two distributions.*

The detailed proof of Theorem 3 can be found in Appendix Section 5.

4 Achievable Results

We complement the prior lower bound with the following robustness guarantee for learning Ising models with corrupted samples.

Theorem 4. *Given an Ising model distribution $\mathcal{D}$ of width λ such that the absolute value of all edge weights are greater than α, suppose we receive N η-corrupted samples from $\mathcal{D}$. If $\eta < C_1 \min\{\alpha^2, 1\}e^{-C_2\lambda}$, then, if $N = O\left(\frac{\exp(C_3\lambda)}{\alpha^4}\log(\frac{n}{\delta\alpha})\right)$ samples, we can recover the Ising model structure with probability $1-\delta$, for some fixed constant C_1, C_2, C_3.*

We establish this by showing robustness of the Sparsitron algorithm introduced by Klivans and Meka [2017] where the authors established how to learn Ising models using *sparse generalized linear models* (GLMs). A generalized linear model is defined by a weight vector w and link function

$\sigma : \mathbb{R} \mapsto [0, 1]$, which in this work we assume to be the logistic function[2]. The model predicts the response $\hat{y}$ to a feature x as $\sigma(w \cdot x)$.

For an Ising model with weight matrix A and mean field θ, we have that

$$\mathrm{P}(x_i = +1 \mid x_{\setminus i}) = \sigma(2\theta_i + 2\sum_j A_{ij}x_j),$$

where σ is the logistic function. Thus, if we set the true label $y_i = 1(x_i = +1)$, then the expected label is a GLM. Klivans and Meka [2017] show how to learn an Ising model by solving

$$\min_w \mathrm{E}[(\sigma(w \cdot x) - y_i)^2].$$

Lemma 5 (Klivans and Meka [2017])**.** *Given an Ising model distribution D of width λ, suppose that w^*, θ^* is the weight vector such that $\mathrm{P}(x_i = 1 \mid x_{\setminus i}) = \sigma(w^* \cdot x + \theta^*)$. If we have a weight vector w, θ such that, for some $\alpha < 1$ and fixed constant C,*

$$\mathrm{E}_{X\sim D}[(\sigma(w \cdot x + \theta) - \sigma(w^* \cdot x + \theta^*))^2] \leq \alpha^2 e^{-C\lambda}$$

then we have

$$\|w - w^*\|_\infty \leq \alpha.$$

Their result shows that if we can learn the appropriate Ising model GLMs with small enough error then we can learn the Ising structure. Their algorithm to learn the sparse GLM is *Sparsitron*, which is based on the Hedge algorithm due Freund and Schapire [1997].

We are able to show that Sparsitron is robust to corrupted samples. See Section 6 for the proof. Since Sparsitron is based on multiplicative weights, the key result needed to establish Theorem 4 is a robustness guarantee of the Hedge algorithm.

4.1 Robustness of the Hedge Algorithm

The Hedge algorithm is a powerful tool used in many applications in learning theory and computer science [Freund and Schapire, 1997, Arora et al., 2012].

In the experts problem, there are n experts. At every iteration we want to generate a distribution over experts p^t. We then observe a loss of each expert ℓ_i^t. We want our distributions to minimize the total loss $L = \sum_{t=1}^T p^t \cdot \ell^t$.

The Hedge algorithm learns distributions in the following way:

1. Initialize a weight $w_i^1 = 1$ for each expert i.
2. At iteration t, use the distribution $p^t = \frac{w^t}{\|w^t\|_1}$.
3. After observing the loss ℓ_i^t for each expert i, set $w_i^{t+1} = w_i^t(1 - \varepsilon)^{\ell_i^t}$.

We want to control the loss of the Hedge algorithm in the following adversarial noise model, where for an η-fraction of losses, the true loss ℓ^t is different from the observed loss $\tilde{\ell}^t$.

Definition (η-corrupted experts problem)**.** In every iteration t for every expert i, there is a loss ℓ_i^t such that $0 \leq \ell_i^t \leq 1$. However we observe the loss $\tilde{\ell}_i^t$. In the η-corruption model, the losses ℓ_i^t are all created, however before we start observing data the adversary sees all the losses and creates the observed losses $\tilde{\ell}_i^t$ such that at most an η-fraction of the observed losses differ from the true losses. Our goal is to generate a distribution p^t at every iteration over the experts such that the total loss at the T-th iteration $L = \sum_{t=1}^T p^t \cdot \ell^t$ is minimized. However, we have to do this without observing the true losses, only the observed losses $\tilde{\ell}^t$.

A simple modification of the proof of the noiseless guarantee of Hedge shows that if we run the hedge algorithm on the observed losses, we can still bound the true loss.

Lemma 6. *Let $L_i = \sum_{t=1}^T \ell_i^t$. In the η-corrupted loss model, the multiplicative weights algorithm on the observed losses $\tilde{\ell}^t$ achieves total loss $L = \sum_{t=1}^T p^t \cdot \ell^t$ such that*

$$L \leq \frac{\ln n}{\varepsilon} + (1 + \varepsilon)L_i^T + 3\eta T.$$

The proof is in Section 6 of the Appendix. Note that this is an additive factor of $O(\eta T)$ from the guarantee in the noiseless case.

[2]For GLMs, we can use any 1-Lipschitz link functions.

Acknowledgments

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1110007. This research has been supported by NSF Grants CCF 1422549, 1618689, DMS 1723052, CCF 1763702, ARO YIP W911NF-14-1-0258 and research gifts by Google, Western Digital, and NVIDIA.

References

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012.

Guy Bresler. Efficiently learning ising models on arbitrary graphs. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 771–782. ACM, 2015.

Constantinos Daskalakis, Elchanan Mossel, and Sébastien Roch. Evolutionary trees and the ising model on the bethe lattice: a proof of steel's conjecture. *Probability Theory and Related Fields*, 149(1-2):149–189, 2011.

Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. Concentration of multilinear functions of the ising model with applications to network data. In *Advances in Neural Information Processing Systems*, pages 12–22, 2017.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 655–664. IEEE, 2016.

Ilias Diakonikolas, Daniel Kane, and Alistair Stewart. Robust learning of fixed-structure bayesian networks. In *NIPS*, 2018.

Glenn Ellison. Learning, local interaction, and coordination. *Econometrica: Journal of the Econometric Society*, pages 1047–1071, 1993.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Robust statistics: the approach based on influence functions*, volume 196. John Wiley & Sons, 2011.

Peter J Huber and Elvezio M Ronchetti. *Robust Statistics*, volume 693. John Wiley & Sons, 2011.

Sayash Kapoor, Kumar Kshitij Patel, and Purushottam Kar. Corruption-tolerant bandit learning. *Machine Learning*, pages 1–29, 2018.

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

Kevin A Lai, Anup B Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 665–674. IEEE, 2016.

Andrey Y Lokhov, Marc Vuffray, Sidhant Misra, and Michael Chertkov. Optimal structure and parameter learning of ising models. *Science advances*, 4(3):e1700791, 2018.

Andrea Montanari and Amin Saberi. The spread of innovations in social networks. *Proceedings of the National Academy of Sciences*, 107(47):20196–20201, 2010.

Marc Vuffray, Sidhant Misra, Andrey Lokhov, and Michael Chertkov. Interaction screening: Efficient and sample-optimal learning of ising models. In *Advances in Neural Information Processing Systems*, pages 2595–2603, 2016.

Appendix

5 Proof of Theorem 3

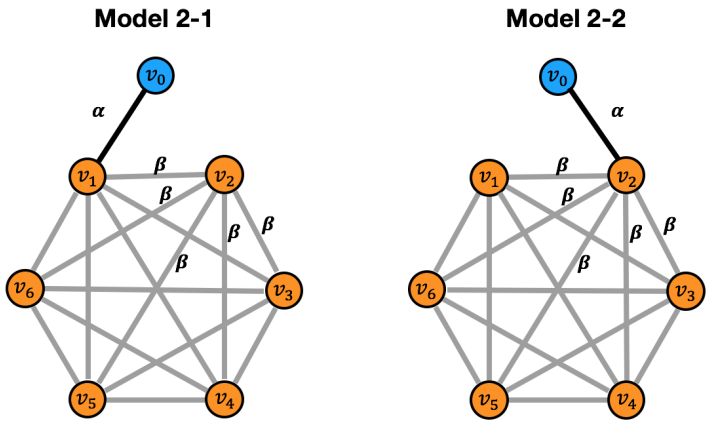

Figure 2: Graphs for Theorem 3

Proof. We will create two Ising model distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ that satisfy the conditions of Theorem 3 and show that they have total variations distance $d_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_2) \leq \alpha e^{-\beta(d-2)}$.

Both models will have a clique of size d on vertices $v_1, v_2, \ldots, v_d$ such that every edge weight in the clique is β. Both models will also have a vertex v_0. Model $2-1$ will have an edge between v_0 and v_1 with edge weight α. Model $2-2$ will have an edge between v_0 and v_2 with edge weight α. In addition to what is described in the figures, we will also assume a graph with $n-d-1$ vertices with identical configurations which are unconnected to the rest of the graph and thus don't affect the energy and consequently the total variation distance calculations.

Define $E(k) = \binom{k}{2} + \binom{d-2-k}{2} - k(d-2-k)$. Note that $E(k)$ is the energy from the edge weights between $\{x_3, x_4, \ldots, x_d\}$ when k of the variables take the value of 1.

We will lower bound the partition function Z by

$$Z \geq (e^{\alpha} + e^{-\alpha}) e^{d} \sum_{k=0}^{d-2} \binom{d-2}{k} e^{\beta E(k)}. \tag{1}$$

To see this, first consider only the configurations such that $x_1 = x_2$. Then group the configurations by the number of variables with the value 1 in $\{x_3, x_4, \ldots, x_d$. For each of the $\binom{d-2}{k}$ configuration of $\{x_3, x_4, \ldots, x_d\}$ with k values equal to 1, we need to also decide the value of $x_1 = x_2$ and the value of x_0. We thus have

$$\begin{aligned} Z \geq & (e^{\alpha} + e^{-\alpha}) \sum_{k=0}^{d-2} \binom{d-2}{k} \left(e^{2\beta(2k-d+2)+\beta E(k)+\beta} + e^{-2\beta(2k-d+2)+\beta E(k)+\beta} \right) \\ \geq & (e^{\alpha} + e^{-\alpha}) e^{\beta} \sum_{k=0}^{d-2} \binom{d-2}{k} e^{\beta E(k)} e^{|2(2k-d+2)|\beta} \end{aligned}$$

We have $\max\{2k-d+2, -(2k-d+2)\} = |2k-d+2|$.

$$
\begin{aligned}
d_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_2) =& \frac{1}{2}\frac{1}{Z} \sum_{x \in \{-1,+1\}^p} |E_1(x) - E_2(x)| \\
=& \frac{1}{2}\frac{4}{Z} \sum_{k=0}^{d-2} \binom{d-2}{k} \left(e^{\alpha} - e^{-\alpha}\right) e^{-\beta} e^{\beta E(k)} \\
\leq& 2\frac{e^{\alpha} - e^{-\alpha}}{e^{\alpha} + e^{-\alpha}} e^{-2\beta} \frac{\sum_{k=0}^{d-2} \binom{d-2}{k} e^{\beta E(k)}}{\sum_{k=0}^{d-2} \binom{d-2}{k} e^{\beta E(k)} e^{|2(2k-d+2)|\beta}} \\
\leq& 2\frac{e^{\alpha} - e^{-\alpha}}{e^{\alpha} + e^{-\alpha}} e^{-2\beta} \frac{\sum_{k=0}^{d-2} \binom{d-2}{k} e^{\beta E(k)}}{e^{\beta E(d-2)} e^{4(d-2)\beta}} \\
\leq& 2\frac{e^{\alpha} - e^{-\alpha}}{e^{\alpha} + e^{-\alpha}} e^{-2\beta} \frac{2^{d-2} e^{\beta E(d-2)}}{e^{\beta E(d-2)} e^{4(d-2)\beta}} \\
\leq& 2\alpha e^{-2\beta} e^{-(d-2)(4\beta - \ln 2)}
\end{aligned}
$$

□

6 Proof of Achieveability

To prove Theorem 4, we just need to show that we can learn the sparse GLM even when the samples are coming from an η-corrupted distribution. The rest of the proof follows from Lemma 5.

We first prove Theorem 6, following the approach from Freund and Schapire [1997] (see also Arora et al. [2012]).

Proof of Theorem 6. We define $\Phi^t = \sum_{i=1}^n w_i^t$. We first upper bound the value of Φ^T.

Suppose that at time t we observe the loss $\tilde{\ell}^t$ and that the observed loss $\tilde{\ell}^t$ is equal to the true loss ℓ^t. We have that

$$
\begin{aligned}
\Phi^{t+1} &= \sum_{i=1}^n w_i^{t+1} = \sum_{i=1}^n w_i^t (1-\gamma)^{\ell_i^t} \leq \sum_{i=1}^n w_i^t (1 - \ell_i^t \gamma) \\
&= \Phi^t - \gamma \Phi^t \sum_{i=1}^n p_i^t \ell_i^t \leq \Phi^t \exp(-\gamma p^t \cdot \ell^t).
\end{aligned}
$$

Now suppose that the observed loss $\tilde{\ell}^t$ is different from the true loss ℓ^t. Note that $\tilde{\ell}_i^t - \ell_i^t \geq -1$. We have that

$$
\begin{aligned}
\Phi^{t+1} &= \sum_{i=1}^n w_i^{t+1} = \sum_{i=1}^n w_i^t (1-\gamma)^{\tilde{\ell}_i^t} = \sum_{i=1}^n w_i^t (1-\gamma)^{\tilde{\ell}_i^t + \ell_i^t - \ell_i^t} \\
&\leq (1-\gamma)^{-1} \sum_{i=1}^n w_i^t (1-\gamma)^{\ell_i^t} \leq (1-\gamma)^{-1} \Phi^t \exp(-\gamma p^t \cdot \ell^t).
\end{aligned}
$$

Since only ηT steps have corrupted losses, we have that

$$
\Phi^T \leq \Phi^1 (1-\gamma)^{-\eta T} \exp(\gamma L), \tag{2}
$$

where $L = \sum_{t=1}^T p^t \cdot \ell^t$ is the total loss.

We now lower bound the value of Φ^T. If the observed loss is equal to the true loss, we have that

$$
\Phi^{t+1} \geq w_i^{t+1} = w_i^t (1-\gamma)^{\ell_i^{t+1}},
$$

otherwise, since $\tilde{\ell}_i^t - \ell_i^t \leq 1$, we have that

$$
\Phi^{t+1} \geq w_i^{t+1} = w_i^t (1-\gamma)^{\tilde{\ell}_i^t - \ell_i^t + \ell_i^t} \geq (1-\gamma) w_i^t (1-\gamma)^{\ell_i^t}.
$$

Since only ηT steps have corrupted losses, we have that

$$\Phi^T \geq w^1(1-\gamma)^{\eta T}(1-\gamma)^{L_i}, \tag{3}$$

where $L_i = \sum_{t=1}^{T} \ell_i^T$ is the loss for expert i.

Since $\Phi^1 = n$ and $w_i^1 = 1$, we can use Equations (2) and (3) to show that

$$(1-\gamma)^{\eta T}(1-\gamma)^{L_i} \leq \Phi^T \leq n(1-\gamma)^{-\eta T}\exp(\gamma L),$$

which implies that

$$L \leq \frac{\ln n}{\gamma} + L_i\frac{-\ln(1-\gamma)}{\gamma} + 2\eta T\frac{-\ln(1-\gamma)}{\gamma}.$$

Using the fact that $-\ln(1-\gamma) \leq \gamma(1+\gamma)$ for $\gamma \leq \frac{1}{2}$, we can conclude that

$$L \leq \frac{\ln n}{\gamma} + L_i(1+\gamma) + 3\eta T.$$

□

We then need to prove that the Sparsitron algorithm is able to efficiently optimize sparse GLMs with an η-corrupted dataset. The Sparsitron algorithm is specified by Algorithm 1. Note that it is a particular instance of the multiplicative weights algorithm.

Algorithm 1 Sparsitron [Klivans and Meka [2017]]

Input: training samples $(x^1, y^1), \ldots, (x^T, y^T)$
Input: test samples $(a^1, b^1), \ldots, (a^T, b^T)$
Input: sparsity parameter λ, weight parameter γ
initialize all weights: $w_i^0 = 1$
for all iterations $t = 1, 2, \ldots, T$:
  $p^t = \frac{w^{t-1}}{\|w^{t-1}\|_1}$
  $\ell^t = \frac{1}{2}(1 + (\sigma(\lambda p^t \cdot x^t) - y^t)x^t)$
  $w_t^t = w_i^{t-1}(1-\gamma)^{\ell_i^t}$
for all iterations $t = 1, 2, \ldots T$:
  $\hat{\varepsilon}(\lambda p^t) = \frac{1}{T}\sum_{j=1}^{T}(u(\lambda p^t \cdot a^j) - b^j)^2$
Return λp^{t^*} for $t^* = \operatorname{argmin}_t \hat{\varepsilon}(\lambda p^t)$

Using Sparsitron, we can establish the following Theorem.

Theorem 7. *Let $\mathcal{D}$ be a distribution on $\{-1,1\}^n \times \{0,1\}$, such that $E_{x,y\sim\mathcal{D}}[y \mid x] = \sigma(w^* \cdot x)$. Assume that $\|w^*\|_1 \leq \lambda$ for a known λ. There exists an algorithm such that, given $T = O(\frac{\lambda^2}{\varepsilon^2}\log\frac{n}{\delta\varepsilon})$ η-corrupted samples from $\mathcal{D}$, learns a weight vector w such that, with probability $1-\delta$, satisfies*

$$\mathrm{E}_{x,y\sim\mathcal{D}}[(\sigma(w\cdot x) - \sigma(w^*\cdot x))^2] \leq \varepsilon + O(\lambda\eta).$$

Using Theorem 7, we can now prove Theorem 4 by setting $\varepsilon, \eta \leq \min\{\alpha^2, 1\}e^{-C\max\{\lambda,1\}}$ for some constant C and applying Lemma 5.

We now prove Theorem 7. The proof is essentially the same as the proof of Theorem 3.1 by Klivans and Meka [2017] in the non-adversarial setting, expect that we use the bound of Theorem 6. For completeness, we duplicate the proof including the error induced by the adversarial setting.

Proof of Theorem 4. We can assume that $w^* \geq 0$ and $\|w^*\|_1 = \lambda$. If not, we can map examples (x, y) to $((x, -x, 0), y)$.

Define the risk of a weight vector v to be $\varepsilon(v) = \mathrm{E}_{x,y\sim D}[(\sigma(v\cdot x) - y)^2]$.

Since the training set and the holdout set are the same size, if the whole dataset is η-corrupted, then each portion of the dataset is at most 2η corrupted.

Let $\tilde{x}^t, \tilde{y}^t$ be the corrupted examples. To be fully rigorous, we need to define a probability space over the examples $(x^t, y^t, \tilde{x}^t, \tilde{y}^t)$. We will assume the adversary takes the following form:

1. The adversary receives the dataset (x^t, y^t).
2. The adversary enumerates all valid η-corrupted datasets.
3. The adversary runs the Sparsitron algorithm for each η-corrupted dataset and calculates the risk of the feature vector learned from each dataset.
4. The adversary sends us the dataset with the highest risk.

This is a deterministic function of the true dataset. If we are robust to this adversary, then we can be robust to any adversary.

Let $Q^t = p^t \cdot \ell^t - \frac{w^*}{\lambda} \cdot \ell^t$. Let

$$Z^t = Q^t - \mathop{\mathrm{E}}_{x^t, y^t, \tilde{x}^t, \tilde{y}^t}[Q^t \mid (x^1, y^1, \tilde{x}^1, \tilde{y}^1), \ldots, (x^{t-1}, y^{t-1}, \tilde{x}^{t-1}, \tilde{y}^{t-1})].$$

Note that $Z^1, \ldots, Z^T$ is a martingale difference sequence with respect to the sequence $(x^1, y^1, \tilde{x}^1, \tilde{y}^1), \ldots, (x^T, y^T, \tilde{x}^T, \tilde{y}^T)$, as Z^t is a function of the values $(x^1, y^1, \tilde{x}^1, \tilde{y}^1), \ldots, (x^T, y^T, \tilde{x}^t, \tilde{y}^t)$. Further, we have that Z^t is bounded between -2 and 2. Thus, by the Azuma-Hoeffding inequality, we have that $\left|\sum_{t=1}^T Z^t\right| \leq O(\sqrt{T \log(1/\delta)})$, with probability $1 - \delta$. We can conclude that, with probability at least $1 - \delta$,

$$\sum_{t=1}^T \mathop{\mathrm{E}}_{x^t, y^t, \tilde{x}^t, \tilde{y}^t}[Q^t \mid (x^1, y^1, \tilde{x}^1, \tilde{y}^1), \ldots, (x^{t-1}, y^{t-1}, \tilde{x}^{t-1}, \tilde{y}^{t-1})] \leq \sum_{t=1}^T Q^t + O(\sqrt{T \log(1/\delta)}). \tag{4}$$

Now we analyze the term

$$\mathop{\mathrm{E}}_{x^t, y^t, \tilde{x}^t, \tilde{y}^t}[Q^t \mid (x^1, y^1, \tilde{x}^1, \tilde{y}^1), \ldots, (x^{t-1}, y^{t-1}, \tilde{x}^{t-1}, \tilde{y}^{t-1})]$$

and relate it to the error of the weight vector λp^t. Recall that

$$\begin{aligned}
&\mathop{\mathrm{E}}_{x^t, y^t, \tilde{x}^t, \tilde{y}^t}[Q^t \mid (x^1, y^1, \tilde{x}^1, \tilde{y}^1), \ldots, (x^{t-1}, y^{t-1}, \tilde{x}^{t-1}, \tilde{y}^{t-1})] \\
&= \mathop{\mathrm{E}}_{x^t, y^t, \tilde{x}^t, \tilde{y}^t}[(p^t - (1/\lambda)w^*) \cdot \ell^t \mid (x^1, y^1, \tilde{x}^1, \tilde{y}^1), \ldots, (x^{t-1}, y^{t-1}, \tilde{x}^{t-1}, \tilde{y}^{t-1})].
\end{aligned}$$

Note that p^t is completely determined by $(\tilde{x}^1, \tilde{y}^1), \ldots, (\tilde{x}^{t-1}, \tilde{y}^{t-1})$ and ℓ^t is completely determined by (x^t, y^t). We thus have

$$\mathop{\mathrm{E}}_{x^t, y^t, \tilde{x}^t, \tilde{y}^t}[Q^t \mid (x^1, y^1, \tilde{x}^1, \tilde{y}^1), \ldots, (x^{t-1}, y^{t-1}, \tilde{x}^{t-1}, \tilde{y}^{t-1})] = \mathop{\mathrm{E}}_{x^t, y^t}[(p^t - (1/\lambda)w^*) \cdot \ell^t].$$

From the proof of Theorem 3.1 of Klivans and Meka [2017], we can conclude that

$$\mathop{\mathrm{E}}_{x^t, y^t, \tilde{x}^t, \tilde{y}^t}[Q^t \mid (x^1, y^1, \tilde{x}^1, \tilde{y}^1), \ldots, (x^{t-1}, y^{t-1}, \tilde{x}^{t-1}, \tilde{y}^{t-1})] = \mathop{\mathrm{E}}_{x^t, y^t}[(p^t - (1/\lambda)w^*) \cdot \ell^t] \geq \frac{1}{2\lambda}\varepsilon(\lambda p^t).$$

Connecting the above with Inequality (4), we have, with probability at least $1 - \delta$,

$$\begin{aligned}
\frac{1}{2\lambda}\sum_{t=1}^T \varepsilon(\lambda p^t) &\leq \sum_{t=1}^T Q^t + O(\sqrt{T \log(1/\delta)}) \\
&= \sum_{t=1}^T p^t \cdot \ell^t - \sum_{t=1}^T \frac{w^*}{\lambda} \cdot \ell^t + O(\sqrt{T \log(1/\delta)}).
\end{aligned} \tag{5}$$

Now, using Theorem 6, with $\gamma = \sqrt{\frac{\ln n}{T}}$, we have that the total loss $L = \sum_{t=1}^T p^t \cdot \ell^T$ satisfies

$$L \leq \min_i L_i + O(\sqrt{T \log n} + \eta T),$$

where $L_i = \sum_{t=1}^T \ell_i^T$ is the loss for expert i. Connecting this with Inequality (5), we have, with probability $1 - \delta$

$$\frac{1}{2\lambda}\sum_{t=1}^T \varepsilon(\lambda p^t) \leq \min_i L_i - \sum_{t=1}^T \frac{w^*}{\lambda} \cdot \ell^t + O(\sqrt{T \log(1/\delta)} + \sqrt{T \log n} + \eta T).$$

Since w^*/λ is a valid distribution, and $\min_i L_i$ is the minimum loss for all distributions, we have that $\min_i L_i - \sum_{t=1}^{T} \frac{w^*}{\lambda} \cdot \ell^t \leq 0$, and thus, with probability $1-\delta$,

$$\frac{1}{2\lambda} \sum_{t=1}^{T} \varepsilon(\lambda p^t) \leq O(\sqrt{T \log(1/\delta)} + \sqrt{T \log n} + \eta T)$$

$$\implies \min_t \varepsilon(\lambda p^t) \leq \frac{1}{T} \sum_{t=1}^{T} \varepsilon(\lambda p^t) \leq \lambda O\left(\sqrt{\frac{\log(1/\delta) + \log n}{T}}\right) + O(\lambda\eta).$$

Setting $T = O(\frac{\lambda^2 \log(n/\delta)}{\varepsilon^2})$, we have, with probability $1-\delta$, that

$$\min_t \varepsilon(\lambda p^t) \leq O(\varepsilon + \lambda\eta).$$

Using our holdout set, we calculate the empirical error for each weight vector λp^t using

$$\hat{\varepsilon}(\lambda p^t) = \frac{1}{T} \sum_{j=1}^{T} (\sigma(\lambda p^t \cdot a^j) - b^j)^2.$$

From Fact 3.2 of Klivans and Meka [2017], since $T \geq O(\frac{\log(T/\delta)}{\varepsilon^2})$, we know that, with probability $1-\delta$, we have that $|\varepsilon(\lambda p^t) - \hat{\varepsilon}(\lambda p^t)| \leq \varepsilon$. However, our examples are η-corrupted. Since each value of $(\sigma(\lambda p^t \cdot a^j) - b^j)^2$ is bounded between -4 and 4, the mean of the corrupted examples differs from the mean of the true examples by an additive factor of $O(\eta)$ at most. Thus, by choosing the weight vector λp^t with the smallest empirical error, we can find a weight vector λp^t such that $\varepsilon(\lambda p^t) \leq O(\varepsilon + \lambda\eta)$. □