Notes on the bad training data paper

Fan Wu

November 7, 2019

1 Motivation

We first provide some motivation behind the paper Shen and Sanghavi [2019] reviewed in these notes.

When training a model using some given training data, if the model is expressive enough to fit all data, it might generalize poorly if a portion of the data is corrupted. This includes scenarios where a portion of the data is labeled incorrectly as well as backdoor attacks, where some adversarial samples are introduced (i.e. both labels and features changed).

Consider the setting where the model is trained in epochs or stages. The key observation motivating the algorithm presented is the following: *especially at the beginning of training, the training error of "bad" samples is higher than that of "clean" samples.* This suggests a natural approach: iteratively alternate between (a) selecting data points with a small loss, and (b) re-training the model on the chosen subset of all samples. This approach is related to minimizing the *trimmed loss*

$$\hat{\theta}^{(TL)} = \arg\min_{\theta} \min_{S:|S|=\lfloor \alpha n \rfloor} \sum_{i \in S} f_\theta(s_i) \tag{1}$$

which involves jointly choosing a subset S of size $\lfloor \alpha n \rfloor$ and an optimal θ, and is intractable in general. The approach proposed in Shen and Sanghavi [2019] can be seen as an iterative way of minimizing the trimmed loss.

Theoretical guarantees for generalized linear models are given, and experimental results for (a) deep image classification with errors in the labels only, (b) generative adversarial networks with bad training images and (c) deep image classification with adversarial images (i.e. both the image and the label) can be found in Shen and Sanghavi [2019].

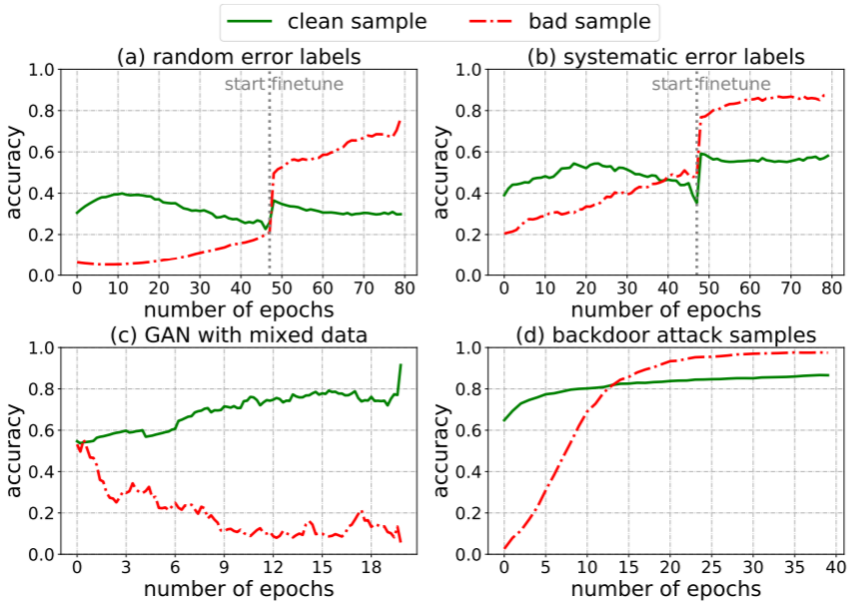

Figure 1: Image taken from Shen and Sanghavi [2019]. Evolution of accuracy for clean and bad training data for (a) classification for CIFAR-10 with 40% random errors in labels, (b) classification for CIFAR-10 with 40% systematic errors in labels, (c) DC_GAN trained on unlabeled mixture of 70% MNIST and 30% Fashion-MNIST images and (d) backdoor attack on classification for CIFAR-10.

2 Setup and notation

2.1 Iterative Trimmed Loss Minimization

Let $s_1, ..., s_n$ be the samples, where $s_i = (x_i, y_i)$, $\theta \in \mathcal{B}$ the model parameter to be learned, $\alpha \in (0, 1)$ the fraction of samples we want to fit and $f_\theta(\cdot)$ the loss function. For a fixed $\theta \in \mathcal{B}$, denote by D_θ and d_θ the distribution and density function of $f_\theta(s)$ respectively. Let $S(\theta) = \mathbb{E}[f_\theta(s)]$ be the population loss, $S_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_\theta(s_i)$ the empirical loss and $F(\theta) = \mathbb{E}[f_\theta(s)\mathbf{I}(f_\theta(s) < D_\theta^{-1}(\alpha)]$ the population trimmed loss. Let $\mathcal{U}(\theta, \epsilon) = \{\tilde{\theta} : |S(\tilde{\theta}) - S(\theta)| < \epsilon\}$ be the set of parameters with population loss close to θ.

We can now formalize the algorithm described in the previous section.

Algorithm 1 Iterative Trimmed Loss Minimization (ITML)

Input: samples $\{s_i\}_{i=1}^n$, number of rounds T, fraction of samples α
(Optional) Initialize: $\theta_0 \leftarrow \arg\min_\theta \sum_{i=1}^n f_\theta(s_i)$
for $t = 0$ **to** $T-1$ **do**
    Choose samples with smallest current loss:

$$S_t \leftarrow \arg \min_{S:|S|=\lfloor \alpha n \rfloor} \sum_{i \in S} f_\theta(s_i)$$

    $\theta_{t+1} = \text{ModelUpdate}(\theta_t, S_t, t)$
end for
Return: θ_T

ModelUpdate refers to the procedure which finds a new θ given a sample set S_t with θ_t as initial value. For completeness, we will describe batch stochastic gradient, which is used as update procedure.

Algorithm 2 BatchSGD_ModelUpdate(θ, S, t)

Input: Initial state θ, set S, round t
Choose: Step size η, number of gradient steps M, batch size N
for $j = 1$ **to** M **do**
    $B_j \leftarrow \text{RANDOM\_SUBSET}(S, N)$
    $\theta^j \leftarrow \theta^{j-1} - \eta\big(\frac{1}{N}\sum_{i \in B_j} \nabla_\theta f_\theta(s_i)\big)$
end for
Return: θ^M

2.2 Generalized linear model

Next, we describe the setting, in which the theoretical results for ITML are proved. Consider a generalized linear model with errors only in the labels:

$$y = \omega(\phi(x)^T\theta^*) + e \qquad \text{(clean sample)} \tag{2}$$

$$y = r + e \qquad \text{(bad sample)} \tag{3}$$

Here x represents the input, y the output, the embedding function ϕ and link function ω are assumed to be known, e is random sub-Gaussian noise with parameter σ^2 and θ^* is the parameter we try to learn. Let α^* be the fraction of clean samples in the data set, $S^* \subset \{1, ..., n\}$ the indices of the clean samples, and consider the squared loss $f_\theta(x, y) = (y - \omega(\phi(x)^T\theta))^2$

Finally, we will assume the feature matrix $\Phi(X)$ to be regular.

Definition 1. *Let $\Phi(X) \in \mathbb{R}^{n\times d}$ be the feature matrix for all samples, where $\phi(x_i)^T$ is the i^{th} row. Let $\mathcal{W}_k = \{W \in \mathbb{R}^{n\times n} : W_{ij} = 0, W_{ii} \in \{0,1\}, \mathrm{Tr}(W) = k\}$, and define*

$$\psi^-(k) = \min_{W\in\mathcal{W}_k} \sigma_{min}(\Phi(X)^T W \Phi(X)) \tag{4}$$

$$\psi^+(k) = \max_{W\in\mathcal{W}_k} \sigma_{max}(\Phi(X)^T W \Phi(X)) \tag{5}$$

where by σ_{min} and σ_{max} we denote the minimum and maximum eigenvalue respectively. We say tht $\Phi(X)$ is a regular feature matrix, if for $k = \alpha n$, $\alpha \in [c, 1]$, $\phi^-(k), \phi^+(k) \in \Theta(n)$ for $n \in \Omega(d \log d)$.

For example, if every row of $\Phi(X)$ is i.i.d. sub-Gaussian, then $\psi^-(k), \psi^+(k) \in \Theta(k)$ and $\Phi(X)$ is regular (see e.g. Bhatia et al. [2015] Theorem 17).

3 Main result

We will state and prove the main results only for the linear case $\omega(x) = x$; the non-linear case can be shown in a very similar way.

We will always make the following, very natural assumptions on the distribution of the samples.

Assumption 2.

Identification condition: *For every $\epsilon > 0$ there exists a $\delta > 0$ such that if $\theta \notin \mathcal{U}(\theta^*, \epsilon)$, then $F(\theta) - F(\theta^*) > \delta$.*

Regularity condition: *D_θ is absolutely continuous for any $\theta \in \mathcal{B}$ and d_θ is bounded uniformly in $\theta \in \mathcal{B}$ and positive in a neighborhood of its α-quantile.*

Under these assumptions, we can show a lemma describing the improvement of ITML in one step in the linear case.

Lemma 3. *Assume $\omega(x) = x$. Under Assumption 1, ITLM with some $\alpha \in (0, 1]$, M large and η small enough, satisfies with high probability*

$$\|\theta_{t+1} - \theta^*\|_2 \leq \frac{\sqrt{2}\psi^+(|S_t \backslash S^*|)}{\psi^-(\alpha n)} \|\theta_t - \theta^*\|_2 + \frac{\sqrt{2}\varphi_t + c\xi_t\sigma}{\psi^-(\alpha n)} \tag{6}$$

where $\varphi_t = \left\| \sum_{i\in S_t\backslash S^*} (\phi(x_i)^T\theta_t - r_i - e_i)\phi(x_i) \right\|_2$ and $\xi_t = \sqrt{\sum_{i=1}^n \|\phi(x_i)\|_2^2 \log n}$

Remark 1. *Lemma 3 bounds the error in the next step based on the current error, how mismatched S_t and S^* are and the regularity parameters. This result immediately implies the consistency of the algorithm if we have no bad samples ($\alpha^* = 1$) and the feature matrix $\Phi(X)$ is regular. In that case, $S^* = \{1, ..., n\}$, and we have $\psi^+(|S_t \backslash S^*|) = \varphi_t = 0$. Further, $\xi_t \in \mathcal{O}(\sqrt{n \log n})$ is sublinear in n, so $\frac{c\xi_t\sigma}{\psi^-(\alpha n)} \to 0$ as $n \to \infty$. This is unsurprising, as it can be*

shown under mild assumptions that the minimizer of the trimmed loss is a consistent estimator (e.g. Shen and Sanghavi [2019] Lemma 3).

Note that the distribution of the labels of bad samples, r, affects this result via the quantity $|S_t \backslash S^*|$.

Under suitable regularity conditions, we can control the parameters in Lemma 3 and characterize the convergence rate in the case of arbitrary or random corruption.

Theorem 4. *Assume $\omega(x) = x$. Assume that $\alpha^* > c_{th}$ (c_{th} depends on the regularity of the feature matrix $\Phi(X)$) and $n \in \Omega(d \log d)$. Under Assumption 1, ITLM with $\alpha < \alpha^*$, M large and η small enough, satisfies with high probability*

$$\|\theta_{t+1} - \theta^*\|_2 \leq \kappa_t \|\theta_t - \theta^*\|_2 + c_1 \sqrt{\kappa_t}\sigma + \frac{c_2 \xi_t}{n}\sigma \tag{7}$$

where $\kappa_t \leq c$ when r is arbitrary, and if r, $\phi(x_i)$ and e are all Gaussian, we have $\kappa_t \leq c \max\{\sqrt{\|\theta_t - \theta^*\|_2^2 + \sigma^2}, \frac{\log n}{n}\}$. The constants c, c_1, c_2, c_{th} all depend on the regularity conditions.

Remark 2. *From this result we see that even as n tends to infinity, the second term does not vanish, that is this theoretical guarantee does not give consistency if $\alpha^* < 1$. Indeed, this behavior is observed empirically.*

Finally, we state without proof the analogous result when the output comes from a mixture model. Assume that $S = [n]$ is split into m subsets $S = \bigcup_{j \in [m]} S_{(j)}$, $|S_{(j)}| = \alpha^*_{(j)} n$. The response variable y_i is given by

$$y_i = \omega\left(\phi(x_i)^T \theta^*_{(j)}\right) + e_i \quad \text{for } i \in S_{(j)} \tag{8}$$

Theorem 5. *Assume $\omega(x) = x$ and $n \in \Omega(d \log d)$. In the mixed regression setting, assume that $\alpha < \alpha^*_{(j)}$ for some component $j \in [m]$. Under Assumption 1, ITLM with α, M large and η small enough, satisfies with high probability*

$$\|\theta_{t+1} - \theta^*_{(j)}\|_2 \leq \kappa_t \|\theta_t - \theta^*_{(j)}\|_2 + c_1 \sqrt{\kappa_t}\sigma + \frac{c_2 \xi_t}{n}\sigma \tag{9}$$

where $\kappa_t \leq c \max\left\{ \frac{\sqrt{\|\theta_t - \theta^*_{(j)}\|_2^2 + \sigma^2}}{\min_{k \neq j} \sqrt{\|\theta_t - \theta^*_{(k)}\|^2 + \sigma^2}}, \frac{\log n}{n} \right\}$.

Remark 3. *In order for $\kappa_0 < 1$, the initialization θ_0 has to satisfy $\|\theta_0 - \theta^*_{(j)}\|_2 \leq C(\alpha) \min_{k \neq j} \|\theta_0 - \theta^*_{(k)}\|_2 - \sqrt{1 - C(\alpha)^2}\sigma$ where $C(\alpha) = \min\{\frac{c_3 \alpha}{1-\alpha}, 1\}$. If α is large enough such that $C(\alpha) = 1$, this condition does not depend on the noise σ. For smaller values of α, the condition becomes stricter, because even if the initial estimate θ_0 is very close to $\theta^*_{(j)}$, if the noise is large, the algorithm could select a substantial number of samples from other mixture components, and it would not converge to $\theta^*_{(j)}$.*

4 Proofs

For the proofs of Lemma 3 and Theorem 4 we will assume the feature matrix to be regular in the sense of Definition 1 with sufficiently good parameters. In particular, we will require. We begin with Lemma 3.

Proof. Let θ_t be the current estimate of the parameter θ^*, the goal is to bound the error of θ_{t+1} obtained by one iteration of Algorithm 1. Let S_t be the subset selected at iteration t, and let W_t be the diagonal matrix with $W_{ii} = 1$ if $i \in S$ and $W_{ii} = 0$ otherwise. Assume that we solve the least square problem with only the samples in S_t exactly to obtain θ_{t+1}, that is we take infinitely many steps in ModelUpdate ($M = \infty$). By taking M large and η small enough, we can get arbitrarily close to this solution. θ_{t+1} is given by

$$\theta_{t+1} = (\Phi(X)^T W_t \Phi(X))^{-1} \Phi(X)^T W_t y \tag{10}$$

where we used $W_t^2 = W_t$. Note that θ_{t+1} depends on θ_t only via the selected subset S_t, that is the diagonal matrix W_t.

Denote by W^* the ground truth diagonal matrix for the samples corresponding to the set S^* of clean samples. Recall that for clean samples we have $y_i = \phi(x_i)^T \theta^* + e_i$, and for bad samples $y_i = r_i + e_i$, so we can write

$$\begin{aligned}
\theta_{t+1} &= (\Phi(X)^T W_t \Phi(X))^{-1} \Phi(X)^T W_t [W^* \Phi(X)\theta^* + (I - W^*)r + e] \\
&= \theta^* + (\Phi(X)^T W_t \Phi(X))^{-1} \Phi(X)^T (W_t W^* - W_t)(\Phi(X)\theta^* - r - e) \\
&\quad + (\Phi(X)^T W_t \Phi(X))^{-1} \Phi(X)^T W_t W^* e
\end{aligned} \tag{11}$$

by simply rearranging terms. Using this, we can bound

$$\begin{aligned}
\|\theta_{t+1} - \theta^*\|_2 &= \|(\Phi(X)^T W_t \Phi(X))^{-1} \Phi(X)^T (W_t W^* - W_t)(\Phi(X)\theta^* - r - e) \\
&\qquad + (\Phi(X)^T W_t \Phi(X))^{-1} \Phi(X)^T W_t W^* e\|_2 \\
&\leq \underbrace{\|(\Phi(X)^T W_t \Phi(X))^{-1}\|_2}_{\mathcal{T}_1} \Big(\underbrace{\|\Phi(X)^T (W_t W^* - W_t)(\Phi(X)\theta^* - r - e)\|_2}_{\mathcal{T}_2} \\
&\qquad + \underbrace{\|(\Phi(X)^T W_t \Phi(X))^{-1} \Phi(X)^T W_t W^* e\|_2}_{\mathcal{T}_3} \Big)
\end{aligned} \tag{12}$$

The three terms can be bounded separately. First, recalling that W_t selects αn rows of $\Phi(X)$ and Definition 1, we have $\mathcal{T}_1 \leq \frac{1}{\psi^-(\alpha n)}$. For the second term, note that $(u+v)^T A (u+v) \leq$

$2u^TAu + 2v^TAv$ for positive semidefinite matrices A, so we can write

$$\begin{aligned}
\mathcal{T}_2^2 &= \|\Phi(X)^T(W_tW^* - W_t)(\Phi(X)\theta^* - r - e)\|_2 \\
&= (\Phi(X)\theta^* - r - e)^T[(W_t - W_tW^*)\Phi(X)\Phi(X)^T(W_t - W_tW^*)](\Phi(X)\theta^* - r - e) \\
&\leq 2(\Phi(X)\theta^* - \Phi(X)\theta_t)^T[(W_t - W_tW^*)\Phi(X)\Phi(X)^T(W_t - W_tW^*)](\Phi(X)\theta^* - \Phi(X)\theta_t) \\
&\quad + 2(\Phi(X)\theta_t - r - e)^T[(W_t - W_tW^*)\Phi(X)\Phi(X)^T(W_t - W_tW^*)](\Phi(X)\theta_t - r - e) \\
&\leq 2\sigma_{max}(\Phi(X)^T(W_t - W_tW^*)\Phi(X))^2\|\theta_t - \theta^*\|_2^2 \\
&\quad + 2(\Phi(X)\theta_t - r - e)^T[(W_t - W_tW^*)\Phi(X)\Phi(X)^T(W_t - W_tW^*)](\Phi(X)\theta_t - r - e) \quad (13)
\end{aligned}$$

where by $\sigma_{max}(\cdot)$ we denote the largest eigenvalue. The second term is exactly the definition of $\varphi_t = \left\| \sum_{i \in S_t \backslash S^*} (\phi(x_i)^T\theta_t - r_i - e_i)\phi(x_i) \right\|_2$. For the first term note that $W_t - W_tW^*$ is a diagonal matrix, where the i^{th} entry is 1 if $i \in S_t \backslash S^*$ and 0 otherwise, so we can bound the eigenvalue by $\psi^+(|S_t \backslash S^*|)$. Thus, we have for the second term

$$\mathcal{T}_2 \leq \sqrt{2}\psi^+(|S_t \backslash S^*|) + \sqrt{2}\varphi_t \quad (14)$$

The third term can be bounded by

$$\mathcal{T}_3^2 = \|\Phi(X)^TW_tW^*e\|_2^2 \leq e^T\Phi(X)\Phi(X)^Te = \sum_{i=1}^{d}\left(\sum_{j=1}^{n} e_j\phi(x_j)_i\right)^2 \leq c^2\sum_{j=1}^{n}\|\phi(x_j)\|_2^2 \log n\sigma^2 \quad (15)$$

where the last probability holds with high probability and can be shown as follows. Recall that e_j is i.i.d. sub-Gaussian with variance proxy σ^2, so, fixing i, $\sum_{j=1}^n e_j\phi(x_j)_i$ is sub-Gaussian with variance proxy $\sigma^2\sum_{j=1}^n \phi(x_j)_i^2$. By standard sub-Gaussian concentration, we have that

$$\mathbb{P}\left[\left(\sum_{j=1}^{n} e_j\phi(x_j)_i\right)^2 > \epsilon^2\right] = \mathbb{P}\left[\left|\sum_{j=1}^{n} e_j\phi(x_j)_i\right| > \epsilon\right] \leq 2e^{-\frac{\epsilon^2}{2\sigma^2\sum_{j=1}^n \phi(x_j)_i^2}} \quad (16)$$

In other words, $\left(\sum_{j=1}^n e_j\phi(x_j)_i\right)^2 < c^2\sum_{j=1}^n \phi(x_j)_i^2 \log n\sigma^2$ with probability $1 - n^{\frac{c}{2}}$. Taking the union bound over $i \in d$ and choosing c large enough gives the desired inequality with high probability.

Putting everything together, we have

$$\|\theta_{t+1} - \theta^*\|_2 \leq \frac{\sqrt{2}\psi^+(|S_t \backslash S^*|)}{\psi^-(\alpha n)}\|\theta_t - \theta^*\|_2 + \frac{\sqrt{2}\varphi_t}{\psi^-(\alpha n)} + \frac{c\sqrt{\sum_{i=1}^n \|\phi(x_i)\|_2^2 \log n}}{\psi^-(\alpha n)}\sigma \quad (17)$$

□

To show Theorem 4, we need to control the parameters appearing in the bound in Lemma 3.

Proof. First note that samples in $S_t \backslash S^*$ are selected because they have a smaller loss than samples not in S_t, i.e. for $i \in S_t \backslash S^*$ we have $|\phi(x_i)^T \theta_t - r_i - e_i| \leq |\phi(x_j)^T(\theta_t - \theta^*) - e_j|$ for any $j \in S^* \backslash S_t$. Because we chose $\alpha < \alpha^*$, we have $|S_t \backslash S^*| \leq |S^* \backslash S_t|$, so there is a permutation matrix P_t such that

$$(W_t - W_t W^*)|\Phi(X)\theta_t - r - e| \leq (W_t - W_t W^*)P_t|\Phi(X)(\theta_t - \theta^*) - e| \tag{18}$$

holds elementwise. Using this permutation matrix, and for some diagonal matrix N_t with entries either 1 or -1, we can bound

$$\begin{aligned}
\varphi_t &= \|\Phi(X)^T(W_t - W_t W^*)(\Phi(X)\theta_t - r - e)\|_2 \\
&\leq \|\Phi(X)^T(W_t - W_t W^*)N_t P_t(\Phi(X)(\theta_t - \theta^*) - e)\|_2 \\
&\leq \|\Phi(X)^T(W_t - W_t W^*)N_t P_t \Phi(X)(\theta_t - \theta^*)\|_2 \\
&\quad + \|\Phi(X)^T(W_t - W_t W^*)N_t P_t e\|_2
\end{aligned} \tag{19}$$

where the first inequality follows because if $|u| < |v|$ elementwise, then $u^T\Phi(X)\Phi(X)^T u < |v|^T\Phi(X)\Phi(X)^T|v|$ since $\Phi(X)\Phi(X)^T$ only has non-negative entries. For the first term we have to bound the maximum singular value of the matrix

$$\Phi(X)^T(W_t - W_t W^*)N_t P_t \Phi(X) \tag{20}$$

Note that N_t can contain negative values. Denoting the maximum singular value by σ_{max}, we can write

$$\begin{aligned}
\sigma_{max} &= \max_{\|u\|_2, \|v\|_2 = 1} u^T\Phi(X)^T(W_t - W_t W^*)N_t P_t \Phi(X)v \\
&\leq \sum_{i=1}^{|S_t \backslash S^*|} |\tilde{u}_{r_i}\tilde{v}_{t_i}| \\
&\leq \max\left\{ \sum_{i=1}^{|S_t \backslash S^*|} \tilde{u}_{r_i}^2, \sum_{i=1}^{|S_t \backslash S^*|} \tilde{v}_{r_i}^2 \right\}
\end{aligned} \tag{21}$$

for some sequences r_i and t_i and we substituted $\tilde{u} = \Phi(X)u$, $\tilde{v} = \Phi(X)v$. So σ_{max} is bounded by

$$\max\left\{\sigma_{max}(\Phi(X)^T(W_t - W_t W^*)\Phi(X)), \sigma_{max}(\Phi(X)^T P_t^T N_t(W_t - W_t W^*)N_t P_t \Phi(X))\right\}$$

which is bounded by $\psi^+(|S_t \backslash S^*|)$, since both $(W_t - W_t W^*)$ and $P_t^T N_t(W_t - W_t W^*)N_t P_t$ are diagonal matrices with entries in $\{0, 1\}$ and trace $|S_t \backslash S^*|$.

Using this and sub-Gaussian concentration, we can bound

$$\|\Phi(X)^T(W_t - W_t W^*)N_t P_t e\|_2 \leq \|\Phi(X)^T(W_t - W_t W^*)N_t P_t\|_2 \|e\|_2 \leq \sqrt{\psi^+(|S_t \backslash S^*|)}c\sqrt{n}\sigma \tag{22}$$

with high probability. All in all, we have shown that

$$\varphi_t \leq \psi^+(|S_t \backslash S^*|)\|\theta_t - \theta^*\|_2 + c\sqrt{\psi^+(|S_t \backslash S^*|)n}\sigma \tag{23}$$

holds with high probability. Plugging this back into the bound of Lemma 3, we have

$$\|\theta_{t+1} - \theta^*\|_2 \leq \underbrace{\frac{2\sqrt{2}\psi^+(|S_t \backslash S^*|)}{\psi^-(\alpha n)}}_{\kappa_t} \|\theta_t - \theta^*\|_2 + \frac{\sqrt{2}c\sqrt{\psi^+(|S_t \backslash S^*|)n}}{\psi^-(\alpha n)}\sigma + \frac{c_2\xi_t}{n}\sigma \tag{24}$$

Recall that by our regularity assumption, $\psi^-(\alpha n) \in \Theta(n)$, so $\frac{\sqrt{2}c\sqrt{\psi^+(|S_t \backslash S^*|)n}}{\psi^-(\alpha n)} = c_1\sqrt{\kappa_t}$.

In the case of arbitrary output, under suitable regularity conditions (which are satisfied e.g. if $\Phi(X)$ has i.i.d. sub-Gaussian rows), and if α^* is large enough, $\kappa_t \leq \frac{1}{2}$ can be guaranteed.

If we assume that r is random Gaussian output, we can use that, under suitable regularity conditions, $\frac{\sqrt{\psi^+(|S_t \backslash S^*|)n}}{\psi^-(\alpha n)} \in \Theta\left(\frac{|S_t \backslash S^*|}{n}\right)$, and the quantity $|S_t \backslash S^*|$ can be better controlled using

Theorem 11 **Shen and Sanghavi [2019]** *Suppose we have $\alpha^* n$ samples from a Gaussian distribution $\mathcal{D}_1 = \mathcal{N}(0, \Delta^2)$ and $(1-\alpha^*)n$ from $\mathcal{D}_2 = \mathcal{N}(0, 1)$ with $\Delta < 1$. Denote by $S_{\alpha n}$ the set of αn samples with smallest absolute value, where $\alpha < \alpha^*$. Then, with high probability, at most $c\max\{\Delta(1-\alpha^*)n, \log n\}$ samples in $S_{\alpha n}$ are from $\mathcal{D}_2$.*

Now consider the losses for clean and bad samples for a fixed θ_t:

$$\text{Clean sample} \qquad \phi(x_i)^T(\theta^* - \theta_t) + e_i \tag{25}$$

$$\text{Bad sample} \qquad r - \phi(x_i)\theta_t + e_i \tag{26}$$

With our assumption, both losses are Gaussian and above result applies, giving

$$\kappa_t \leq c\max\left\{\sqrt{\|\theta_t - \theta^*\|_2^2 + \sigma^2}, \quad \frac{\log n}{n}\right\} \tag{27}$$

An $\epsilon-$net argument shows that this indeed holds for all θ_t. □

Note: The original version of Theorem 4 is stated without the Gaussian assumption in the random output case, however we need both an upper and lower bound for the two losses, so it doesn't seem straightforward to extent this result to the sub-Gaussian case (without further assumptions).

References

K. Bhatia, P. Jain, and P. Kar. Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems*, 2015.

Y. Shen and S. Sanghavi. Learning with bad training data via iterative trimmed loss minimization. In *Proceedings of the* 36^{th} *International Conference on Machine Learning*, 2019.