Appendix for
Learning with Bad Training Data via Iterative Trimmed Loss Minimization

Yanyao Shen [1] **Sujay Sanghavi** [1]

1. Property of the TL Estimator

Proof of Lemma 3. We use standard techniques for consistency proof, similar to (Čížek, 2008). First, let f be the loss of a single sample, F_n be the loss of sum of n smallest losses over the total sample size. $f_{\lfloor \alpha n \rfloor}$ is the $\lfloor \alpha n \rfloor$-th smallest loss. We can re-write F_n into the following two terms:

$$\begin{aligned} F_n(\theta) =& \frac{1}{n}\sum_{i=1}^{n} f(s_i;\theta)\cdot \mathbb{I}\left\{f(s_i;\theta) \le f_{\lfloor \alpha n\rfloor}(\theta)\right\} \\ =& \frac{1}{n}\sum_{i=1}^{n} f(s_i;\theta)\cdot \left(\mathbb{I}\left\{f(s_i;\theta) \le f_{\lfloor \alpha n\rfloor}(\theta)\right\} - \mathbb{I}\left\{f(s_i;\theta)\le D_\theta^{-1}(\alpha)\right\}\right) & (1) \\ &+ \frac{1}{n}\sum_{i=1}^{n} f(s_i;\theta)\cdot \mathbb{I}\left\{f(s_i;\theta)\le D_\theta^{-1}(\alpha)\right\} & (2) \end{aligned}$$

where D_θ is the distribution function of $f_\theta(s)$, and D_θ^{-1} is its inverse function, which calculates the quantile value. On the other hand, define F to be the expected average trimmed loss, i.e.,

$$F(\theta) =\mathbb{E}\left[f(s_i;\theta)\cdot \mathbb{I}\left\{f(s_i;\theta)\le D_\theta^{-1}(\alpha)\right\}\right] \tag{3}$$

Then, the difference between $F_n(\theta)$ and $F(\theta)$ can be separated into two terms: the first term is the difference between (2) and (3), which asymptotically goes to zero due to the law of large numbers; on the other hand, the term (1) goes to zero because of the convergence of order statistics to the quantile. See (Čížek, 2008) for showing the consistency of both terms under the regularity conditions.

By definition, TL $\hat{\theta}^{(\mathrm{TL})}$ satisfies $\Pr\left[F_n(\hat{\theta}^{(\mathrm{TL})}) < F_n(\theta^\star)\right] = 1$. For any $\epsilon > 0$,

$$\begin{aligned} 1 =& \Pr\left[F_n(\hat{\theta}^{(\mathrm{TL})}) < F_n(\theta^\star)\right] \\ =& \Pr\left[F_n(\hat{\theta}^{(\mathrm{TL})}) < F_n(\theta^\star), \hat{\theta}^{(\mathrm{TL})} \in \mathcal{U}(\theta^\star,\epsilon)\right] + \Pr\left[F_n(\hat{\theta}^{(\mathrm{TL})}) < F_n(\theta^\star), \hat{\theta}^{(\mathrm{TL})} \in \mathcal{B}\backslash\mathcal{U}(\theta^\star,\epsilon)\right] \\ \le& \Pr\left[\hat{\theta}^{(\mathrm{TL})} \in \mathcal{U}(\theta^\star,\epsilon)\right] + \Pr\left[\inf_{\theta\in\mathcal{B}\backslash\mathcal{U}(\theta^\star,\epsilon)} F_n(\theta) < F_n(\theta^\star)\right] & (4) \end{aligned}$$

where in the last inequality, we use the fact that the probability measure on a set is no less than the probability measure on its subset. Notice that our goal is to show $\hat{\theta}^{(\mathrm{TL})}$ is in $\mathcal{U}(\theta^\star,\epsilon)$ with probability 1. This is true as long as the second term is

[1] ECE Department, University of Texas at Austin, TX, USA. Correspondence to: Yanyao Shen <shenyanyao@utexas.edu>, Sujay Sanghavi <sanghavi@mail.utexas.edu>.

Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

zero. The second term in the above can be controlled by

$$\begin{aligned}
&\Pr\left[\inf_{\theta\in\mathcal{B}\setminus\mathcal{U}(\theta^\star,\epsilon)} F_n(\theta) < F_n(\theta^\star)\right] \\
=&\Pr\left[\inf_{\theta\in\mathcal{B}\setminus\mathcal{U}(\theta^\star,\epsilon)} [F_n(\theta) - F(\theta) + F(\theta)] < F_n(\theta^\star)\right] \\
\le&\Pr\left[\inf_{\theta\in\mathcal{B}\setminus\mathcal{U}(\theta^\star,\epsilon)} [F_n(\theta) - F(\theta)] < F_n(\theta^\star) - \inf_{\theta\in\mathcal{B}\setminus\mathcal{U}(\theta^\star,\epsilon)} F(\theta)\right] && (5)\\
\le&\Pr\left[\sup_{\theta\in\mathcal{B}} |F_n(\theta) - F(\theta)| > \inf_{\theta\in\mathcal{B}\setminus\mathcal{U}(\theta^\star,\epsilon)} F(\theta) - F_n(\theta^\star)\right] && (6)\\
\le&\Pr\left[2\sup_{\theta\in\mathcal{B}} |F_n(\theta) - F(\theta)| > \inf_{\theta\in\mathcal{B}\setminus\mathcal{U}(\theta^\star,\epsilon)} F(\theta) - F(\theta^\star)\right] && (7)
\end{aligned}$$

where (5) is due to triangle inequality, in (6), we flip the sign on both sides and upper bound the difference by the abstract value. (7) again uses triangle inequality, in order to separate the population loss on θ and the sample loss on $\theta^\star$. As we have discussed at the beginning, under regularity conditions, $F_n(\theta) - F(\theta)$ goes to zero asymptotically. More specifically, for any $\epsilon > 0$, $\Pr\left[\sup_{\theta\in\mathcal{B}} |F_n(\theta) - F(\theta)| > \frac{\delta(\epsilon)}{2}\right] \to 0$ as $n \to +\infty$. On the other hand, $\Pr\left[\inf_{\theta\in\mathcal{B}\setminus\mathcal{U}(\theta^\star,\epsilon)} F(\theta) - F(\theta^\star) < \delta(\epsilon)\right] = 0$, which is given by the idenfication condition. Combining with (4), and triangle inequality, we have $|S_n(\hat{\theta}^{(\mathrm{TL})}) - S_n(\theta^\star)| \to 0$ with probability 1, as $n \to \infty$. □

2. Clarification of the ITLM Algorithm

For different settings, we use the same procedure as described in Algorithm 1 and 2 (in the main paper), but may select different hyper-parameters. We summarize the alternatives we use for all the settings as follows:

(a) In the linear setting, choosing a large M with small step size η corresponds to finding the closed form solution, which is the setting we analyze;

(b) For generalized linear setting, we analyze for $M = 1$ and $N = |S|$, which corresponds to a single full gradient update per round;

(c) For all experiments with DNNs, we run re-initialization for every round of update to make it harder to stuck at bad local minimum;

(d) For training generative model using GANs, we use the loss of discriminator's output in step 4 in Algorithm 1, and use the joint loss of both the generator and the discriminator in Algorithm 2;

(e) For CIFAR-10 classification tasks with (i) bad labels, (ii) backdoor samples, we choose smaller M for the first 4 rounds, which corresponds to early stopping. As motivated in Section 1 (in the main paper), early stopping may help us better filter out bad samples since the later rounds may overfit on them.

Comparison with several very recent works: Several recent works also use similar ideas that rely on smaller loss samples (Han et al., 2018b; Yu et al., 2019; Han et al., 2018a). On the theoretical side, we provide theoretical analysis/insights for why these types of methods would work. On the practical side, different from their algorithms, ITLM re-initialize the model for every round, which we believe helps the model to avoid bad local minimum. Also, ITLM uses early stopping, which helps to deal with extremely noisy setting (e.g., 80% random label noise).

3. Proofs for ILTM Algorithm

Proof of Lemma 5. Let θ_t be the learned parameter at round t, and θ_{t+1} be the learned parameter in the next round, following Algorithm 1. More specifically, a subset S_t of size αn with the smallest losses $(y_i - \theta_t^\top \cdot \phi(x_i))^2$ is selected. θ_{t+1} is the minimizer on the selected set of sample losses. Denote W_t as the diagonal matrix whose diagonal entry $W_{t,ii}$ equals 1 when the i-th sample is in set S_t, otherwise 0. Then, assume that we take infinite steps and reach the optimal solution (we will discuss how to extend this to arbitrary M_t with small step size later), we have :

$$\theta_{t+1} = \left(\Phi(X)^\top W_t \Phi(X)\right)^{-1} \Phi(X)^\top W_t y$$

where $\Phi(X)$ is an $n \times d$ matrix, whose i-th row is $\phi(x_i)^\top$, and we have used the fact that $W_t^2 = W_t$. Remind that for the feature matrix $\Phi(X)$, we have defined

$$\begin{aligned}
\psi^-(k) &= \min_{W : W \in \mathcal{W}_k} \sigma_{\min}\left(\Phi(X)^\top W \Phi(X)\right), \\
\psi^+(k) &= \max_{W : W \in \mathcal{W}_k} \sigma_{\max}\left(\Phi(X)^\top W \Phi(X)\right),
\end{aligned}$$

which will be used in the later analysis. For $\Phi(X)$ whose every row follows i.i.d. sub-Gaussian random vector, by using concentration of the spectral norm of Gaussian matrices, and uniform bound, $\Phi(X)$ is a regular feature matrix, see, e.g., Theorem 17 in (Bhatia et al., 2015), and other literatures (Davenport et al., 2009).

On the other hand, denote $W^\star$ as the ground truth diagonal matrix for the samples, i.e., $W^\star_{ii} = 1$ if the i-th sample is a clean sample, otherwise $W^\star_{ii} = 0$. Accordingly, define $S^\star$ as the ground truth set of clean samples. For clearness of the presentation, we may drop the subscript t when there is no ambiguation. For bad samples, the output is written in the form of $y_i = r_i + e_i$, where e_i represents the observation noise, and r_i depends on the specific setting we consider (we will discuss more in later Theorems). Under this general representation, we can re-write the term θ_{t+1} as

$$\begin{aligned}
\theta_{t+1} =& \left(\Phi(X)^\top W \Phi(X)\right)^{-1} \Phi(X)^\top W \left(W^\star \Phi(X)\theta^\star + (I - W^\star) r + e\right) \\
=& \theta^\star + \left(\Phi(X)^\top W \Phi(X)\right)^{-1} \left(\Phi(X)^\top W W^\star \Phi(X)\theta^\star + \Phi(X)^\top W r - \Phi(X)^\top W W^\star r - \Phi(X)^\top W \Phi(X)\theta^\star + \Phi(X)^\top W e\right) \\
=& \theta^\star + \left(\Phi(X)^\top W \Phi(X)\right)^{-1} \Phi(X)^\top \left(W W^\star - W\right)\left(\Phi(X)\theta^\star - r - e\right) + \left(\Phi(X)^\top W \Phi(X)\right)^{-1} \Phi(X)^\top W W^\star e
\end{aligned}$$

by basic linear algebra. Therefore, the ℓ_2 distance between the learned parameter and ground truth parameter can be bounded by:

$$\begin{aligned}
& \|\theta_{t+1} - \theta^\star\|_2 \\
=& \left\| \left(\Phi(X)^\top W \Phi(X)\right)^{-1} \Phi(X)^\top \left(W W^\star - W\right)\left(\Phi(X)\theta^\star - r - e\right) + \left(\Phi(X)^\top W \Phi(X)\right)^{-1} \Phi(X)^\top W W^\star e \right\|_2 \\
\leq& \underbrace{\left\| \left(\Phi(X)^\top W \Phi(X)\right)^{-1} \right\|_2}_{\mathcal{T}_1} \cdot \left(\underbrace{\left\| \Phi(X)^\top \left(W W^\star - W\right)\left(\Phi(X)\theta^\star - r - e\right) \right\|_2}_{\mathcal{T}_2} + \underbrace{\left\| \Phi(X)^\top W W^\star e \right\|_2}_{\mathcal{T}_3} \right)
\end{aligned}$$

where basic spectral norm inequalities and triangle inequalities. For the term $\mathcal{T}_1$, notice that W selects αn rows of $\Phi(X)$, i.e., $\mathtt{Tr}(W) = \alpha n$. Therefore, $\mathcal{T}_1 \leq \frac{1}{\psi^-(\alpha n)}$.

Next, the term $\mathcal{T}_2$ can be bounded as:

$$\begin{aligned}
\mathcal{T}_2^2 =& \left\| \Phi(X)^\top \left(W - W W^\star\right)\left(\Phi(X)\theta^\star - r - e\right) \right\|_2^2 \\
=& \left(\Phi(X)\theta^\star - r - e\right)^\top \left[\left(W - W W^\star\right) \Phi(X)\Phi(X)^\top \left(W - W W^\star\right)\right] \left(\Phi(X)\theta^\star - r - e\right) \\
\leq& 2\left(\Phi(X)\theta^\star - \Phi(X)\theta_t\right)^\top \left[\left(W - W W^\star\right) \Phi(X)\Phi(X)^\top \left(W - W W^\star\right)\right] \left(\Phi(X)\theta^\star - \Phi(X)\theta_t\right) \\
& + 2\left(\Phi(X)\theta_t - r - e\right)^\top \left[\left(W - W W^\star\right) \Phi(X)\Phi(X)^\top \left(W - W W^\star\right)\right] \left(\Phi(X)\theta_t - r - e\right) \\
\leq& 2\sigma_{\max}\left(\Phi(X)^\top (W - W W^\star) \Phi(X)\right)^2 \|\theta^\star - \theta_t\|_2^2 \qquad (8) \\
& + 2\underbrace{\left(\Phi(X)\theta_t - r - e\right)^\top \left[\left(W - W W^\star\right) \Phi(X)\Phi(X)^\top \left(W - W W^\star\right)\right] \left(\Phi(X)\theta_t - r - e\right)}_{\varphi(S_t, S^\star, \|\theta^\star - \theta_t\|_2)^2} \qquad (9)
\end{aligned}$$

The last term (9) is defined as $\varphi_t := \varphi(S_t, S^\star, \|\theta^\star - \theta_t\|_2) = \left\| \sum_{i \in S \backslash S^\star} (\phi(x_i)^\top \theta_t - r_i - e_i)\phi(x_i) \right\|_2$. For the term (8), let $|S_t \backslash S^\star|$ be the number of bad samples in S_t. Then, the eigenvalue is bounded by $\psi^+(|S_t \backslash S^\star|)$.

The term $\mathcal{T}_3$ can be bounded as:

$$\mathcal{T}_3^2 = \left\|\Phi(X)^\top W W^\star e\right\|_2^2 \le e^\top \Phi(X)\Phi(X)^\top e = \sum_{i=1}^{d} \left(\sum_{j=1}^{n} e_j \phi(x_j)_i\right)^2 \le c\sum_{i=1}^{n} \|\phi(x_i)\|_2^2 \log n\sigma^2$$

where the last inequality holds with high probability, and all the randomness comes from the measurement noise e. The last inequality is based on the sub-exponential concentration property.

Then, as a summary, combining the results for all three terms, we have:

$$\|\theta^\star - \theta_{t+1}\|_2 \le \frac{\sqrt{2}\psi^+(|S\backslash S^\star|)}{\psi^-(\alpha n)}\|\theta^\star - \theta_t\|_2 + \frac{\sqrt{2}\varphi(S_t, S^\star, \|\theta^\star - \theta_t\|_2)}{\psi^-(\alpha n)} + \frac{c\sqrt{\sum_{i=1}^{n}\|\phi(x_i)\|_2^2 \log n}}{\psi^-(\alpha n)}\sigma$$

□

Discussion on finite M_t As we mentioned before, for the simplicity of the result, we consider θ_{t+1} as a full update on the subset of samples. However, based on this current framework, we can also analyze for finite M_t, with small step size η. The key idea is that in the linear setting, we can connect the updated parameter at each epoch with a closed form solution to a penalized minimization problem. More specifically, accordng to (Suggala et al., 2018), define

$$\dot{\theta}(t) := \frac{d}{dt}\theta(t) = -\nabla f(\theta(t)), \theta(0) = \theta_0$$

and

$$\underline{\theta}(\nu) = \arg\min_{\theta} f(\theta) + \frac{1}{2\nu}\|\theta - \theta_0\|_2^2$$

where $f(\theta) = \frac{1}{2|S|}\sum_{i\in S}(y_i - \phi(x_i)^\top\theta)^2$. Then, $\theta(t)$ and $\underline{\theta}(\nu)$ have the following relationship:

$$\|\theta(t) - \underline{\theta}(\nu(t))\|_2 \le \frac{\|\nabla f(\theta_0)\|_2}{m}\left(e^{-mt} + \frac{c}{1 - c - e^{cMt}}\right)$$

where $\nu(t) = \frac{1}{cm}\left(e^{cMt} - 1\right)$, for $m = \sigma_{\min}(\frac{1}{|S|}\Phi(X)^\top W\Phi(X))$, $M = \sigma_{\max}(\frac{1}{|S|}\Phi(X)^\top W\Phi(X))$, $c = \frac{2m}{M+m}$. Since $\underline{\theta}(\nu)$ has a closed form solution in this linear setting, by connecting θ^{t+1} with $\underline{\theta}$, we are able to bound θ^{t+1} using similar proof technique as above.

Proof of Lemma 6. Define $F : \mathbb{R}^n \to \mathbb{R}^n$ as an entry-wise $f(\cdot)$-operation.

$$\begin{aligned}
\theta_{t+1} =& \theta_t - \frac{\eta}{\alpha n}\sum_{i\in S_t}\left(f\left(\phi(x_i)^\top\theta_t\right) - y_i\right)\cdot f'\left(\phi(x_i)^\top\theta_t\right)\cdot\phi(x_i) \\
=& \theta_t - \frac{\eta}{\alpha n}\Phi(X)^\top \mathtt{Diag}\left(F'\left(\Phi(X)\theta_t\right)\right)W_t\left(F\left(\Phi(X)\theta_t\right) - y\right) \\
=& \theta_t - \frac{\eta}{\alpha n}\Phi(X)^\top \mathtt{Diag}\left(F'\left(\Phi(X)\theta_t\right)\right)W_t\left(F\left(\Phi(X)\theta_t\right) - W^\star F\left(\Phi(X)\theta^\star\right) - \left(I - W^\star\right)\left(r + e\right) - W^\star e\right) \\
=& \theta_t - \frac{\eta}{\alpha n}\Phi(X)^\top \mathtt{Diag}\left(F'\left(\Phi(X)\theta_t\right)\right)W_t\left(F\left(\Phi(X)\theta_t\right) - W^\star F\left(\Phi(X)\theta^\star\right) - (I - W^\star)F\left(\Phi(X)\theta^\star\right)\right) \\
& - \frac{\eta}{\alpha n}\Phi(X)^\top \mathtt{Diag}\left(F'\left(\Phi(X)\theta_t\right)\right)W_t\left((I - W^\star)F\left(\Phi(X)\theta^\star\right) - \left(I - W^\star\right)\left(r + e\right) - W^\star e\right) \\
=& \theta_t - \frac{\eta}{\alpha n}\Phi(X)^\top \mathtt{Diag}\left(F'\left(\Phi(X)\theta_t\right)\right)W_t\left(F\left(\Phi(X)\theta_t\right) - F\left(\Phi(X)\theta^\star\right)\right) \\
& - \frac{\eta}{\alpha n}\Phi(X)^\top \mathtt{Diag}\left(F'\left(\Phi(X)\theta_t\right)\right)\left(W_t - W_t W^\star\right)\left(F\left(\Phi(X)\theta^\star\right) - r - e\right) \\
& + \frac{\eta}{\alpha n}\Phi(X)\mathtt{Diag}\left(F'\left(\Phi(X)\theta_t\right)\right)W_t W^\star e
\end{aligned}$$

We simplify the notation using $H_t \triangleq \mathtt{Diag}\left(F'\left(\Phi(X)\theta_t\right)\right)$. Also, by mean value theorem, for any a, b, there exists some $c \in [a, b]$, such that $\frac{f(b)-f(a)}{b-a} = f'(c)$. Therefore, for the term $F(\Phi(X)\theta_t) - F(\Phi(X)\theta^\star)$, there exists a diagonal matrix C_t, such that $F(\Phi(X)\theta_t) - F(\Phi(X)\theta^\star) = C_t\Phi(X)\left(\theta_t - \theta^\star\right)$. Therefore, we have

$$\begin{aligned}\|\theta_{t+1} - \theta^\star\|_2 \le & \left(\underbrace{1 - \frac{\eta}{\alpha n}\Phi(X)^\top H_t W_t C_t \Phi(X)}_{\mathcal{U}_1}\right)\|\theta_t - \theta^\star\|_2 + \frac{\eta}{\alpha n}\underbrace{\left\|\Phi(X)^\top H_t (W_t - W_t W^\star)\left(F(\Phi(X)\theta^\star) - r - e\right)\right\|_2}_{\mathcal{U}_2} \\ & + \frac{\eta}{\alpha n}\underbrace{\|\Phi(X) H_t W_t W^\star e\|_2}_{\mathcal{U}_3}.\end{aligned}$$

Here,

$$\mathcal{U}_1 \le 1 - \eta a^2 \frac{\psi^-(\alpha n)}{\alpha n}, \mathcal{U}_3 \le b\xi_t\sigma$$

For $\mathcal{U}_2$, define $\tilde{\phi}_t$ similar to ϕ_t:

$$\tilde{\varphi}_t = \left\| \sum_{i \in S_t \backslash S^\star} \left(w(\phi(x_i)^\top\theta^\star) - r_i - e_i\right) w'(\phi(x_i)^\top\theta^\star)\phi(x_i) \right\|.$$

As a result, we have:

$$\|\theta_{t+1} - \theta^\star\|_2 \le \left(1 - \frac{\eta}{\alpha n} a^2 \psi^-(\alpha n)\right)\|\theta_t - \theta^\star\|_2 + \eta\frac{\tilde{\varphi}_t + \xi_t b\sigma}{\alpha n}$$

□

Proof of Theroem 7. Now we consider recovery in the context of aribitrary corrupted output, and random noise setting.

Notice that since samples in $S_t \backslash S^\star$ are selected because of smaller losses, and $\alpha < \alpha^\star$, there exists a permutation matrix P, such that the following inequality holds element-wise:

$$(W - WW^\star)\left|\Phi(X)\theta_t - r - e\right| \le (W - WW^\star)P\left|\Phi(X)(\theta_t - \theta^\star) - e\right|.$$

Accordingly, given a valid permutation matrix P, ϕ_t is further bounded by

$$\begin{aligned}&\phi(S_t, S^\star, \|\theta^\star - \theta_t\|_2)^2 \\ \le& \left(\Phi(X)(\theta_t - \theta^\star) - e\right)^\top NP^\top(W - WW^\star)\Phi(X)\Phi(X)^\top(W - WW^\star)PN\left(\Phi(X)(\theta_t - \theta^\star) - e\right) \\ \le& 2(\theta_t - \theta^\star)^\top\Phi(X)^\top NP^\top(W - WW^\star)\Phi(X)\Phi(X)^\top(W - WW^\star)PN\Phi(X)(\theta_t - \theta^\star) \qquad (10) \\ &+ 2e^\top NP^\top(W - WW^\star)\Phi(X)\Phi(X)^\top(W - WW^\star)PNe \qquad (11) \\ \le& 2\psi^+(|S_t\backslash S^\star|)^2\left\|\theta_t - \theta^\star\right\|_2^2 + 2c\psi^+(|S_t\backslash S^\star|)n\sigma^2, \qquad (12)\end{aligned}$$

where the last inequality (12) holds with high probability. Here, N is some diagonal matrix whose entries are either 1 or -1. More specifically, (10) can be bounded by $2\tilde{\sigma}^2\|\theta_t - \theta^\star\|_2^2$, where $\tilde{\sigma}$ is the top singular value of the matrix

$$\Phi(X)^\top(W - WW^\star)PN\Phi(X).$$

Equivalently, it can be written as

$$\tilde{\sigma} = \max_{u,v:\|u\|_2=\|v\|_2=1} u^\top\Phi(X)^\top(W - WW^\star)PN\Phi(X)v.$$

If we denote $\Phi(X)v$ and $\Phi(X)u$ as $\tilde{v}$, $\tilde{u}$ respectively, then

$$\tilde{\sigma} \le \sum_{i=1}^{|S\backslash S^\star|}|\tilde{u}_{r_i}\tilde{v}_{t_i}| \le \max\left\{\sum_{i=1}^{|S\backslash S^\star|}\tilde{u}_{r_i}^2, \sum_{i=1}^{|S\backslash S^\star|}\tilde{v}_{t_i}^2\right\},$$

for some sequences $\{r_i\}$ and $\{t_i\}$. This shows that the top singular value is indeed bounded by

$$\max\left\{\sigma_{\max}\left(\Phi(X)^\top (W - WW^\star)\Phi(X)\right), \sigma_{\max}\left(\Phi(X)^\top NP^\top (W - WW^\star)PN\Phi(X)\right)\right\},$$

which is bounded by $\psi^+(|S_t \backslash S^\star|)$, since both $W - WW^\star$ and $NP^\top(W - WW^\star)PN$ have $\mathtt{Tr}(W - WW^\star)$ non-zero entries in the diagonal.

The term (11) is bounded because of the feature regularity property. Notice that $(W - WW^\star)\Phi(X)\Phi(X)^\top(W - WW^\star)$ has the same non-zero eigenvalues as $\Phi(X)^\top(W - WW^\star)\Phi(X)$.

Therefore, with high probability,

$$\begin{aligned}\phi_t \leq & \sqrt{2\psi^+(|S_t \backslash S^\star|)^2 \|\theta^\star - \theta_t\|_2^2 + 2c\psi^+(|S_t \backslash S^\star|)n\sigma^2} \\ \leq & \sqrt{2}\psi^+(|S_t \backslash S^\star|)\|\theta^\star - \theta_t\|_2 + \sqrt{2c\psi^+(|S_t \backslash S^\star|)n}\sigma.\end{aligned}$$

Combining previous results, with high probability, we have

$$\|\theta^\star - \theta_{t+1}\|_2 \leq \underbrace{\frac{2\sqrt{2}\psi^+(|S_t \backslash S^\star|)}{\psi^-(\alpha n)}}_{\kappa_t} \|\theta^\star - \theta_t\|_2 + \frac{\sqrt{2c\psi^+(|S_t \backslash S^\star|)n}}{\psi^-(\alpha n)}\sigma + \frac{c\sqrt{\sum_{i=1}^n \|\phi(x_i)\|_2^2 \log n}}{\psi^-(\alpha n)}\sigma. \tag{13}$$

The above result holds for both the setting of random output and arbitrary corruption setting. For arbitrary output setting, since $\psi^+(|S_t \backslash S^\star|)$ can be upper bounded by $\mathcal{O}(n)$, we have:

$$\|\theta^\star - \theta_{t+1}\|_2 \leq \frac{1}{2}\|\theta^\star - \theta_t\|_2 + c\sigma + \frac{c\xi_t}{n}\sigma.$$

In the random output setting, however, in fact we can calculate how the quantity $|S_t \backslash S^\star|$ changes, and have a better characterization of the convergence. Based on Theorem A.1, we have:

$$\kappa_t \leq c\left\{\sqrt{\|\theta_t - \theta^\star\|_2^2 + \sigma^2} \vee \frac{\log n}{n}\right\},$$

for any fixed θ_t. One can use a standard ϵ-net argument to show that the above indeed holds for any θ_t. Therefore, for the case of random output corruption,

$$\|\theta^\star - \theta_{t+1}\|_2 \leq \kappa_t\|\theta^\star - \theta_t\|_2 + c\sqrt{\kappa_t}\sigma + \frac{c\xi_t}{n}\sigma,$$

for $\kappa_t \leq c\{\sqrt{\|\theta_t - \theta^\star\|_2^2 + \sigma^2} \vee \frac{\log n}{n}\}$. □

Proof of Theorem 8. In the context of mixed model setting, we are interested in when the algorithm will find the component that it is closest to. The proof outline is similar to Theorem 7. However, for the case of mixture output, two parts in (13) need re-consideration: the first part is to show that there is an $\Omega(n)$ lower bound for $\psi^-(\alpha n)$ for arbitrary constant α. Notice that in Theorem 17 of (Bhatia et al., 2015), α can not be too small, e.g., 0.1. The main idea of their proof was to use a uniform bound over all possible Ws, which depends on n. However, we take another route and using ϵ-net argument on the parameter space. Notice that we can choose an ϵ-net in $\mathbb{R}^d$, which includes $(1 + \frac{2}{\epsilon})^d$ points (Vershynin, 2016). For any fixed θ, notice that the square of $\min_{W \in \mathcal{W}_{\alpha n}} \sigma_{\min}(\Phi(X)^\top W \Phi(X))$ corresponds to the sum of the minimum αn squares, which is greater than $c_1 n$ with high probability (Boucheron et al., 2012). On the other hand, for arbitrary $\tilde{\theta}$, the additional error is at most $\epsilon\psi^+(\alpha n) = \mathcal{O}(\epsilon n)$. By using the uniform bound over all fixed θ, and choosing $n \geq Cd \log d$ for some large constant c, we can see that $\psi^-(\alpha n)$ is lower bounded by $\Omega(n)$ with high probability. For getting the second term in (13), we use the same idea as in the proof of Theorem **??**. For any fixed θ_t, the residuals for all the samples can be considered as generated from m components, and can be reduced to a two-component setting. Therefore, the numerator in κ_t is again controlled by Theorem A.1. Combining these results, we have $\kappa_t \leq c\left\{\frac{\sqrt{\|\theta_t - \theta^\star_{(j)}\|_2^2 + \sigma^2}}{\min_{k \in [m] \backslash \{j\}} \sqrt{\|\theta_t - \theta^\star_{(k)}\|_2^2 + \sigma^2}} \vee \frac{\log n}{n}\right\}$.

As a consequence,

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

where we require $n = \Omega(d \log d)$. Notice that for small α, in order to make κ_t less than one, the noise should not be too large. Otherwise, even if θ_t is very close to $\theta^\star$, because of the noise and the high density of bad samples, $|S_t \backslash S^\star|$ would still be quite large, and the update will not converge.

□

Theorem 9. *Following the setting in Lemma 6, for the given $\alpha \leq \alpha^\star$, $\Phi(X)$ being a regular feature matrix, and $\alpha^\star > c_{\mathtt{th}}$, sample size $n = \Omega(d \log d)$, w.h.p., we have:*

$$\|\theta^\star - \theta_{t+1}\|_2 \leq \left(1 - c_1 \eta (a^2 - \kappa_t b^2\right) \|\theta_t - \theta^\star\|_2 + c_2 b \sqrt{\kappa_t}\sigma + \frac{\eta b \xi_t}{n}\sigma,$$

where for r being arbitrary output, $\kappa_t \leq \frac{1}{2}$. For r being random sub-Gaussian output, $\kappa_t \leq c\{\frac{b}{a}\sqrt{\|\theta_t - \theta^\star\|_2^2 + \sigma^2} \vee \frac{\log n}{n}\}$.

Theorem 10. *Following the setting in Lemma 6, for the mixed regression setting in (2), suppose for some $j \in [m]$, $\alpha \leq \alpha^\star_{(j)}$. Then, for $n = \Omega(d \log d)$, w.h.p., the next iterate θ_{t+1} of the algorithm satisfies*

$$\|\theta_{t+1} - \theta^\star_{(j)}\|_2 \leq \left(1 - c_1 \eta (a^2 - \kappa_t b^2\right) \|\theta_t - \theta^\star_{(j)}\|_2 + c_1 b \sqrt{\kappa_t}\sigma + \frac{c_2 \eta b \xi_t}{n}\sigma,$$

where $\kappa_t \leq c \left\{ \frac{b\sqrt{\|\theta_t - \theta^\star_{(j)}\|_2^2 + \sigma^2}}{a \min_{k \in [m]\backslash\{j\}} \sqrt{\|\theta_t - \theta^\star_{(k)}\|_2^2 + \sigma^2}} \vee \frac{\log n}{n} \right\}$.

The proof idea for the above two Theorems are similr to what we have shown in the proof of Theorem 7 and Theorem 8.

Theorem A.1. *Suppose we have two Gaussian distributions $\mathcal{D}_1 = \mathcal{N}(0, \Delta^2), \mathcal{D}_2 = \mathcal{N}(0, 1)$. We have $\alpha^\star n$ i.i.d. samples from $\mathcal{D}_1$ and $(1 - \alpha^\star)n$ i.i.d. samples from $\mathcal{D}_2$. Denote the set of the top αn samples with smallest abstract values as $S_{\alpha n}$, where $\alpha \leq \alpha^\star$. Then, with high probability, for $\Delta \leq 1$, at most $\left(c \max\left\{\Delta\left(1 - \alpha^\star\right) n \sqrt{\log n}, \log n\right\}\right)$ samples in $S_{\alpha n}$ are from $\mathcal{D}_2$.*

Proof. **Step I.** We know that for random normal i.i.d. Gaussian variables $x_i, i \in [n]$,

$$P\left[\max_{i \in [n]} |x_i| \geq \sqrt{2 \log 2n} + t\right] \leq 2e^{-\frac{t^2}{2}}.$$

Therefore, for $\alpha^\star n$ samples from $\mathcal{D}_1$, with high probability, the maximum abstract value is in the order of $O(\sqrt{\log n}\Delta)$.

Step II. On the other hand, for a random $u_2 \sim \mathcal{D}_2$, we know that for small positive values $\delta = c\sqrt{\log n}\Delta$, $P\left[|u_2| \leq \delta\right] \leq \sqrt{\frac{2}{\pi}}\delta$ gives a tight upper bound. Let $\mathcal{M}_{\delta,i}$ be the event *sample u_i from $\mathcal{D}_2$ has abstract value less than δ*, and a Bernoulli random variable $m_{i,\delta}$ that is the indicator of event $\mathcal{M}_{\delta,i}$ holds or not. Then,

$$\mathbb{E}\left[\sum_{i=1}^{(1-\alpha^\star)n} m_{i,\delta}\right] \leq \sqrt{\frac{2}{\pi}}\delta\left(1 - \alpha^\star\right) n.$$

For independent random variable x_is, $i \in [\tilde{n}]$ that lie in interval $[0, 1]$, with $X = \sum_i x_i$ and $\mu = \mathbb{E}[X]$, Chernoff's inequality tells us

$$P\left[X \geq (1 + \gamma)\mu\right] \leq e^{-\frac{\mu\gamma^2}{3}}, \;\; \forall \gamma \in [0, 1]$$
$$P\left[X \geq (1 + \gamma)\mu\right] \leq e^{-\frac{\mu\gamma}{3}}, \;\; \forall \gamma > 1$$

As a consequence, for $m_{i,\delta}$s, we have

$$P\left[\sum_{i=1}^{(1-\alpha^\star)n} m_{i,\delta} \geq (1+\gamma)\sqrt{\frac{2}{\pi}}\delta\left(1-\alpha^\star\right)n\right] \leq e^{-\frac{\gamma^2\sqrt{\frac{2}{\pi}}\delta(1-\alpha^\star)n}{3}},\ \forall\gamma\in[0,1]$$

$$P\left[\sum_{i=1}^{(1-\alpha^\star)n} m_{i,\delta} \geq (1+\gamma)\sqrt{\frac{2}{\pi}}\delta\left(1-\alpha^\star\right)n\right] \leq e^{-\frac{\gamma\sqrt{\frac{2}{\pi}}\delta(1-\alpha^\star)n}{3}},\ \forall\gamma > 1.$$

For the first case, where $\gamma\in[0,1]$, we can set $\gamma = c\sqrt{\frac{\log n}{\delta(1-\alpha^\star)n}}$ to get high probability guarantee. The constraint on γ requires $\delta n > c\log n$ for some fixed c. On the contrary, when this is violated, i.e., when δ is much smaller, then, by the Chernoff bound for the case $\gamma > 1$, we can set $\gamma = \frac{c\log n}{\delta(1-\alpha^\star)n}$.

Combining Step I and Step II. To summarize, for some fixed constant c, with high probability:

- For $\Delta > \frac{c\sqrt{\log n}}{n}$, at most $2c\delta\left(1-\alpha^\star\right)n = c\sqrt{\log n}\Delta\left(1-\alpha^\star\right)n$ samples in $S_{\alpha n}$ are from $\mathcal{D}_2$.
- For $\Delta \leq \frac{c\sqrt{\log n}}{n}$, at most $(1+\gamma)\,\delta\left(1-\alpha^\star\right)n = c\log n$ samples in $S_{\alpha n}$ are from $\mathcal{D}_2$.

□

4. Additional Synthetic Experiments

In this section, we present the full results for the synthetic experiments, which aligns with our theoretic results in Section 5 (in the main paper). We focus on discussing behaviors for the linear case first, and then provide results on the non-linear setting.

Synthetic experiments for random output setting We generate the data according to (1), with $w(x) = x$, where we choose $\theta^\star$ to be a random unit vector with dimension $d = 100$, every feature vector $\phi(x_i)$ is generated i.i.d. as a d-dimension normal spherical Gaussian. Random output r_i is generated i.i.d. following $\mathcal{N}(0,1)$, which makes the distribution of both the bad and good outputs the same. We generate in total $n = 1000$ samples, where $\alpha^\star$-fraction of them are clean samples and the rest are bad samples (with random output). The noise vector e is generated i.i.d. Gaussian with variance σ^2.

Synthetic experiments for mixed regression setting We generate the data following (2) with $w(x) = x$, for the settng of two components. The rest of the settings are similar to the random output setting, except for the bad samples, we select another $\theta_{(1)}$ with unit norm, orthogonal to $\theta^\star$.

In Figure 1 and Figure 2, we study:

- (**Inconsistency**) The recovery performance as sample size increases, in both small-noise and large-noise settings;
- (**Recovery**) The recovery performance under different good sample ratios;
- (**Mis-specification**) The effect of mis-specified α;
- (**Convergence**) The convergence speed under different noise levels, for both large and small M_t settings.

All y-axis measures the ℓ_2 distance, i.e., $\|\theta_t - \theta^\star\|_2$. Each data point in the plots is based on 100 runs of the same experiment to cancel out the random factors.

Inconsistency Figure 1-(a) & (b) and Figure 2-(a) & (b) show the result for asymptotic behavior. ITLM -1 corresponds to our algorithm with large M_t, which corresponds to our analysis using the closed form solution at each update round. ITLM -2 corresponds to our algorithm with $M_t = 1$. The performance in both settings are quite similar: in the (b) plots with noise level $\sigma = 1$, as sample size increases, the oracle performance is getting better, while the performance of ITLM does not keep improving, which shows the inconsistency of the algorithm. However, in the (a) plots with small noise ($\sigma = 0.1$), the difference between oracle and ITLM is not significant, for sample size less than 25k. However, as sample size keeps getter larger, we will observe the behavior of inconsistency for ITLM . The observation matches with our results in Theorem 7 & 8, where our per-round convergence property will guarantee the recovered parameter is within a noise ball to the ground truth parameter.

Recovery Figure 1-(c) and Figure 2-(c) show the recovery performance when good sample ratio varies. ITLM -1 and ITLM -2 perform similarly. As good sample ratio gets larger, the algorithm is capable of recovering close to the ground truth with high probability. Here, noise level $\sigma = 0.2$, α is set as $\alpha^\star - 5\%$ by default.

Mis-specification In Figure 1-(d) and Figure 2-(d), we study the recovery behavior for different mis-specified αs. We see that the recovery performance is not very sensitive to the selection of α, especially when the dataset has more clean samples.

Convergence In Figure 1-(e) & (f), and Figure 2-(e) & (f), we see the convergence is more than linear before the learned parameter gets into the noise-level close to the ground truth, for both settings. This convergence behavior, for both the random output and mixture output settings, matches with our results in Theorem 7 and Theorem 8.

Non-linear activation functions In Figure 3, we present convergence result for a non-linear setting: we choose $w()$ to be a piece-wise linear function, i.e., $w(x) = x$ if $x < 0$, and $w(x) = 1.2x$ if $x \geq 0$. We keep all other settings exactly the same as in previous synthetic experiments. We see that the ITLM has similar convergence behavior as in the linear setting.

5. Additional Experiments and Implementation Details

All experiments are implemented using MXNet and gluon. Here, we add more experimental details and supporting experimental results.

5.1. Details for the image classification task with random/systematic label errors

Training details: We use batch size 1000 with learning rate 0.3 for subsampled MNIST dataset, and batch size 256 with learning rate 0.1 for CIFAR-10 dataset, with naive sgd as the optimizer. We use 80 epochs for naive training, and decrease step size at the 50 epoch by 5. The results for MNIST dataset is reported as the median of 5 random runs. In all the experiments, there is **no clean sample** in both the training set and the validation set. The reported accuracy is based tested on the true validation set, but the algorithm saves the best model based on the accuracy on the **bad** validation set, which has the same corruption pattern as the training set.

5.2. Additional experiments for image generation

Training details: We use the popular DC-GAN architecture, and the loss for training is re-written in (14), which is also used for the update step in ITLM .

Table 1. MNIST GAN: comparison with other choices

dataset	MNIST						
$\alpha^\star = \frac{\text{\# clean}}{\text{\# total}}$	**Baseline**	**ITLM**	**Centroid**	**1-Step**	$\Delta\tau = 10\%$	$\Delta\tau = 15\%$	$\Delta\tau = 20\%$
70%	70	97.00	61.46	77.77	83.33	78.06	83.59
80%	80	100.00	77.46	76.84	98.80	99.56	97.77
90%	90	100.00	89.57	91.90	98.85	99.01	98.04

Experimental settings: In this part, we present additional experimental results, in order to verify the performance of ITLM under different parameter settings, and compare with other algorithms. More specifically, we present the results using the following methods/algorithms:

- **Baseline**: naive trainig using all the samples;

- **ITLM** : our proposed iterative learning algorithm with 5 iterations, using a mis-specified τ which is 5% less than the true value;
- **Centroid**: using the centroid of the input data to filter out outliers. For classification task, we calculate the centroids for the samples with the same label/class and filter each class separately;
- **1-Step**: **ITLM** algorithm with a single iteration;
- $\Delta\tau = \tau^\star - \tau \in \{10\%, 15\%, 20\%\}$: **ITLM** under different mis-specified τ value,

under MNIST generation with Fashion-MNIST images.

For the generation task (Table 1), we present the ratio of true MNIST samples selected by each method. For the baseline method, since the DC-GAN is trained using all samples, the reported value is exactly the $\tau^\star$.

Results: Table 1 shows the performance of generation quality under different noise levels. We observe that centroid method does not work, which may due to the fact that all MNIST and Fashion-MNIST images are hard to be distinguished as two clusters in the pixel space. Notice that there are in fact 20 clusters (10 from MNIST, and 10 from Fashion-MNIST), and we are interested in 10 of them. ITLM works well since it automatically learns a clustering rule when generating on the noisy dataset. For example, for $\tau^\star = 80\%$, even with a mis-specified $\tau = 60\%$, ITLM is capable of ignoring almost all bad samples. Again, we also observe significant improvement of ITLM over its 1-step counterpart.

We also have results showing that ITLM works well for generation when the corrupted samples are pure Gaussian noise. However, we do not think it is a practical assumption, and the result is not presented here.

In Figure 4, we present a result under large bad sample ratio: 60% clean MNIST images with 40% bad Fashion-MNIST images. The algorithm, after the 5-th iteration, tries to filter out the all digit-type images.

References

Bhatia, K., Jain, P., and Kar, P. Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems*, pp. 721–729, 2015.

Boucheron, S., Thomas, M., et al. Concentration inequalities for order statistics. *Electronic Communications in Probability*, 17, 2012.

Čížek, P. General trimmed estimation: robust approach to nonlinear and limited dependent variable models. *Econometric Theory*, 24(6):1500–1529, 2008.

Davenport, M. A., Laska, J. N., Boufounos, P. T., and Baraniuk, R. G. A simple proof that random matrices are democratic. *arXiv preprint arXiv:0911.0736*, 2009.

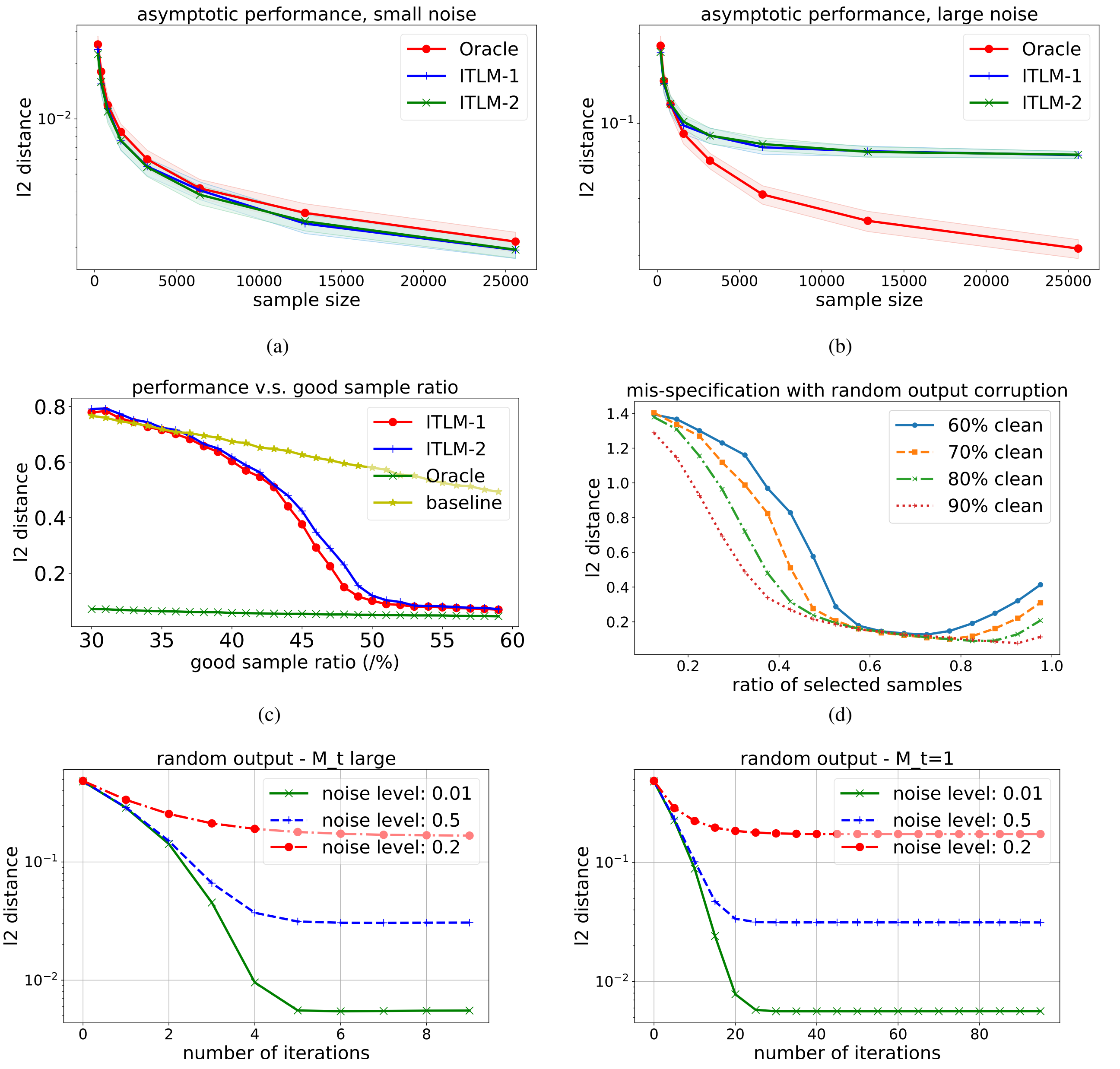
Han, B., Niu, G., Yao, J., Yu, X., Xu, M., Tsang, I., and Sugiyama, M. Pumpout: A meta approach for robustly training deep neural networks with noisy labels. *arXiv preprint arXiv:1809.11008*, 2018a.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., and Sugiyama, M. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, pp. 8527–8537, 2018b.

Suggala, A., Prasad, A., and Ravikumar, P. K. Connecting optimization and regularization paths. In *Advances in Neural Information Processing Systems*, pp. 10631–10641, 2018.

Vershynin, R. High dimensional probability. *An Introduction with Applications*, 2016.

Yu, X., Han, B., Yao, J., Niu, G., Tsang, I. W., and Sugiyama, M. How does disagreement benefit co-teaching? *arXiv preprint arXiv:1901.04215*, 2019.



(a)

(b)

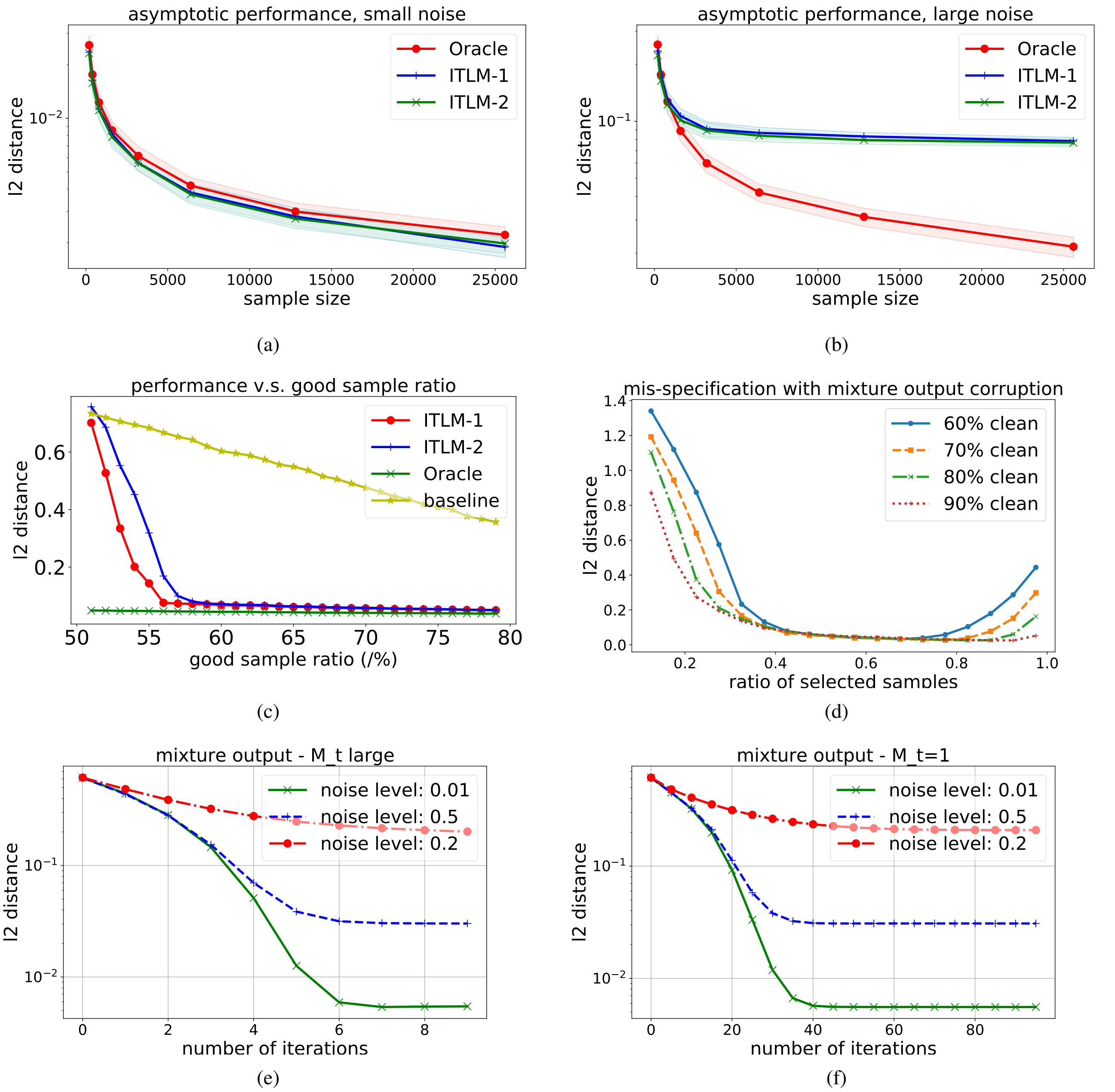
(c)

(d)

(e)

(f)

Figure 1. Synthetic experiments with random output: **(a):** asymptotic performance under small measurement noise; **(b):** asymptotic performance under large measurement noise; **(c):** performance under different good sample ratio; **(d):** the effect of mis-specification; **(e):** convergence rate of ITLM with large M_t (noise from 0.01 to 0.2 ; **(f):** convergence rate of ITLM with small M_t (noise from 0.01 to 0.2).



(a)

(b)

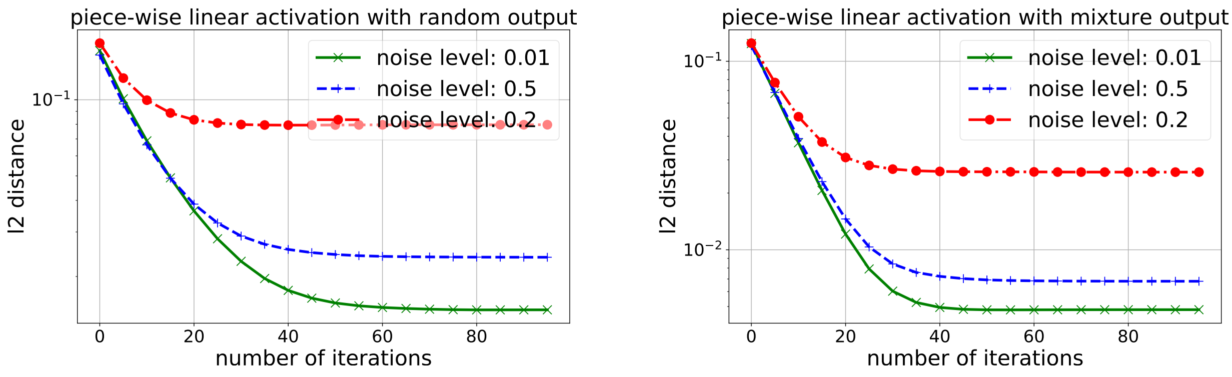
(c)

(d)

(e)

(f)

Figure 2. Synthetic experiments with mixture output: **(a):** asymptotic performance under small measurement noise; **(b):** asymptotic performance under large measurement noise; **(c):** performance under different good sample ratio; **(d):** the effect of mis-specification; **(e):** convergence rate of ITLM with large M_t (noise from 0.01 to 0.2 ; **(f):** convergence rate of ITLM with small M_t (noise from 0.01 to 0.2).



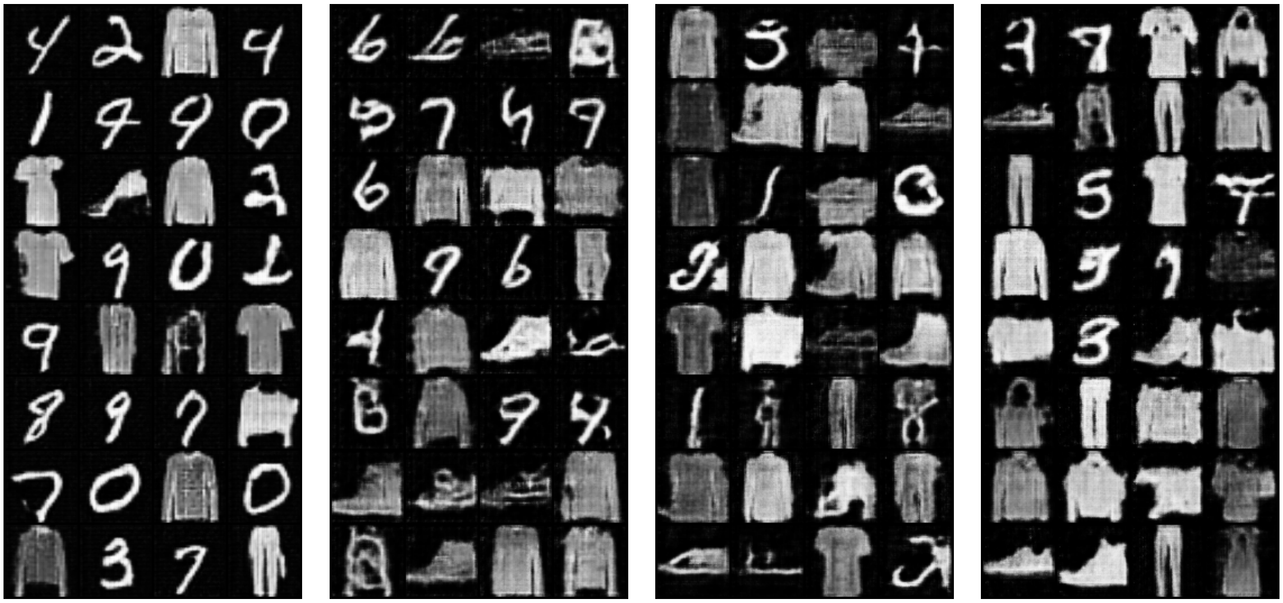
(a)

(b)

Figure 3. Synthetic experiments with non-linear activation function: **(a):** $\|\theta^t - \theta^\star\|_2$ v.s. t for random output setting; **(b):** $\|\theta^t - \theta^\star\|_2$ v.s. t for mixture output setting.

$$L_S^{\texttt{GAN}}(\theta^D,\theta^G) := \frac{1}{|S|}\sum_{i\in S}\log D_{\theta^D}(s_i) + \mathbb{E}_{z\sim p_{\mathcal{Z}}(z)}\left[\log(1-D_{\theta^D}(G_{\theta^G}(z)))\right] \tag{14}$$

$$S_t \leftarrow \arg\min_{S:|S|=\alpha n}\sum_{i\in S} D_{\theta_t^D}(s_i) \tag{15}$$



(a) baseline

(b) 1st iteration

(c) 3rd iteration

(d) 5th iteration

Figure 4. **Illustrative failure case:** This figure shows that when the fraction of bad samples is too large, ILFB cannot clean them out. The setting is exactly the same as in Figure 3 (in the main paper), but now with 60% MNIST clean images + 40% Fashion-MNIST bad images. We can see that now the 5^{th} iteration still retains the fake fashion images.