A APPENDIX

Proof of Lemma 2. Since w_{t+1} and w_t represents the model parameters in $(t+1)$th iteration and tth iteration respectively, following the first stage of our framework. More specifically, a subset S_t of size αn with the smallest losses is selected. w_{t+1} is the minimizer on the selected set. Denote W_t as the diagonal matrix whose diagonal entry $W_{t,ii}$ equals 1 when the ith instance is in set S_t, otherwise 0. Then, assume that we take infinite steps and reach the optimal solution, we have:

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

where $\Phi(X)$ is an $n \times d$ matrix, whose ith row is $\delta(x_i)^\top$, and we have used the fact that ${W_t}^2 = W_t$. Remained that for the feature matrix $\Phi(X)$, we have defined in Equation 7. 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.

On the other hand, denote W^* as the ground truth diagonal matrix for the training instances, *i.e.*, $W^*_{ii} = 1$ if the ith instance is a clean instance, otherwiseW^*_{ii}. Accordingly, define S^* as the ground truth set of clean instances. For clearness of the presentation, we may drop the subscript t when there is no ambiguation. For mislabeled instances, the output is written in the form of $y_i = r_i + n_i$ where n_i represents the observation noise, and y_i depends on the specific setting we consider. Under this general representation, we can re-write the term w_{t+1} as

$$\begin{aligned}
w_{t+1} &= (\Phi(X)^\top W \Phi(X))^{-1}\Phi(X)^\top W(W^*\Phi(X)w_c^* + (I - W^*)r + e) \\
&= w_c^* + (\Phi(X)^\top W\Phi(X))^{-1}(\Phi(X)^\top WW^*\Phi(X)w_c^* \\
&+ \Phi(X)^\top Wr - \Phi(X)^\top WW^*r - \Phi(X)^\top W\Phi(X)w_c^* + \Phi(X)We) \\
&= w_c^* + (\Phi(X)^\top W\Phi(X))^{-1}\Phi(X)^\top (WW^* - W)(\Phi(X)w_c^* - r - e) \\
&+ (\Phi(X)^\top W\Phi(X))^{-1}\Phi(X)^\top WW^*e
\end{aligned}$$

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

$$\begin{aligned}
\|w_{t+1} - w_c^*\| = &\|(\Phi(X)^\top W\Phi(X))^{-1}\Phi(X)^\top (WW^* - W) \\
&(\Phi(X)w_c^* - r - e) + (\Phi(X)^\top W\Phi(X))^{-1}\Phi(X)^\top WW^*e\|_2 \\
&\le \underbrace{\|(\Phi(X)^\top W\Phi(X))^{-1}\|_2}_{v_1} \cdot \\
&\Bigg(\underbrace{\|\Phi(X)^\top (WW^* - W)(\Phi(X)w_c^* - r - e)\|_2}_{v_2} \\
&+ \underbrace{\|\Phi(X)^\top WW^*e\|_2}_{v_3}\Bigg)
\end{aligned}$$

where basic spectral norm inequalities and triangle inequalities. For the term v_1, notice that W selects αn rows of $\Phi(X)$, *i.e.*, $Tr(W) = \alpha n$. Therefore, $v_1 \le \frac{1}{\Psi^-(\alpha n)}$.

Next, the term v_2 can be bounded as:

$$\begin{aligned}
v_2^2 &= \|(\Phi(X)^\top (W - WW^*)(\Phi(X)w_c^* - r - e)\|_2^2 \\
&= (\Phi(X)w_c^* - r - e)^\top \\
&[(W - WW^*)\Phi(X)\Phi(X)^\top (W - WW^*)](\Phi(X)w_c^* - r - e) \\
&\le 2(\Phi(X)w_c^* - \Phi(X)w_t)^\top \\
&[(W - WW^*)\Phi(X)\Phi(X)^\top (W - WW^*)](\Phi(X)w_c^* - \Phi(X)w_t) \\
&+ 2(\Phi(X)w_t - r - e)^\top \\
&[(W - WW^*)\Phi(X)\Phi(X)^\top (W - WW^*)](\Phi(X)w_t - r - e) \\
&\le 2\sigma_{max}(\Phi(X)^\top (W - WW^*)\Phi(X))^2\|w_c^* - w_t\|_2^2 \\
&+ 2(\Phi(X)w_t - r - e)^\top \\
&[(W - WW^*)\Phi(X)\Phi(X)^\top (W - WW^*)](\Phi(X)w_t - r - e)
\end{aligned}$$

For the first term $2\sigma_{max}(\Phi(X)^\top(W - WW^*)\Phi(X))^2\|w_c^* - w_t\|_2^2$, let $|S_t \setminus S^*|$ be the number of mislabeled samples in S_t. Then, the eigenvalue is bounded by $\Psi^+(|S_t \setminus S^*|)$. And the last term $2(\Phi(X)w_t - r - e)^\top[(W - WW^*)\Phi(X)\Phi(X)^\top(W - WW^*)](\Phi(X)w_t - r - e)$ is defined as $\kappa_t := \kappa(S_t, S^*, \|w_c^* - w_t\|_2) = \|\sum_{i\in S\setminus S^*}(\delta(x_i)^\top w_t r_i - e_i)\delta(x_i)\|_2$. The term v_3 can be bounded as:

$$\begin{aligned} v_3^2 &= \|\Phi(X)^\top WW^* e\|_2 \\ &\leq e^\top \Phi(X)\Phi(X)^\top e \\ &= \sum_{i=1}^{d}(\sum_{j=1}^{n} e_j \delta(x_j)_i)^2 \\ &\leq c\sum_{i=1}^{n}\|\delta(x_i)\|_2^2 \log n\sigma^2 \end{aligned}$$

where the last inequality holds with high probability by the subexponential concentration property, and all randomness comes from the measurement noise e.Then, as a summary, combining the results for all three terms, we have:

$$\begin{aligned} \|w_{t+1} - w_c^*\|_2 \leq & \frac{\sqrt{2}\Psi^+(|S_t \setminus S^*|)}{\Psi^-(\alpha n)}\|w_t - w_c^*\|_2 \\ & + \frac{\sqrt{2}\kappa(S_t, S^*, \|w_c^* - w_t\|_2)}{\Psi^-(\alpha n)} \\ & + c\frac{\sqrt{\sum_{i=1}^{n}\|(\delta(x_i)\|_2^2 \log n\sigma}}{\Psi^-(\alpha n)} \end{aligned}$$