7 Appendix

7.1 Additional Results for Section 3

The following lemma provides upper bounds on the expected gradient of the worst-possible MKL-SGD solution that lies in a ball around $\boldsymbol{w^*}$. Simultaneously satisfying the following bound with the one in Lemma 3 may lead to an infeasible set of ϵ and N'. And thus we use Lemma 4 in conjunction with 3.

Lemma 6. *Let us assume that MKL-SGD converges to $\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}$. For any $\bar{\boldsymbol{w}}_{\boldsymbol{MKL}} \in \mathcal{B}_r(\boldsymbol{w^*})$ that satisfies assumptions N1, N2, A4 and A5, there exists $N' \geq N$ and $\epsilon' \leq \epsilon$ such that,*

$$\left\| \sum_{i \notin \mathbb{O}} p_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}) \nabla f_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}) \right\| \leq \min\left\{ (1-\epsilon^k) L \|\bar{\boldsymbol{w}}_{\boldsymbol{MKL}} - \boldsymbol{w^*}\|, \epsilon^k G(\boldsymbol{w}) \right\}$$

The proof for lemma 2 can be found in the Appendix Section 7.2.8

7.1.1 Squared loss in the scalar setting with outliers centered at different points

We will assume that without loss of generality all the outliers will lie on the same side of w^*. If that's not the case, the bounds which show in the subsequent part will be even stronger. Without loss of generality, assume $0 < w_{b_1} < w_{b_2} < \cdots < w_{b_{|\mathcal{B}|}}$.

The loss functions and $\widetilde{w}$ are redefined as follows:

$$f_i(w) = \begin{cases} l_i(w - w^*)^2 \ \ \forall\ i \notin \mathbb{O} \\ l_i(w - w_{b_i})^2 \ \ \forall\ i \in \mathbb{O}, \end{cases} \tag{14}$$

$$\widetilde{w} := \left\{ w \middle| \ \ w = \min_{\lambda \in (0,1)} \left[\lambda w^* + (1-\lambda) \min_j w_{b_j} \right], f_{l_m}(w) = f_{l_M}(w) \right\} \tag{15}$$

Once again by simple analysis of different points of intersection, we can describe a closed form expression of $\widetilde{\boldsymbol{w}}$ as follows:

$$\widetilde{w} = \frac{\sqrt{l_m} w^* + \sqrt{l_{\tilde{M}}} w_{b_{\tilde{M}}}}{\sqrt{l_m} + \sqrt{l_{\tilde{M}}}}$$

where $\tilde{M} = \arg\max_{j \in \mathbb{O}} \sqrt{l_j} w_{b_j}$ and $\kappa = \dfrac{l_M}{l_m}$ and $\rho = \dfrac{\min_i w_{b_i}}{\max_i w_{b_i}}$

Condition 2. $\hat{p} < \dfrac{1}{1 + \dfrac{\kappa}{\rho\left(1 + \dfrac{1}{\sqrt{\kappa}}\right) - 1}}$

When $\rho = 1$, condition 2 becomes identical to condition 1.

Lemma 7. *If Condition 2 is satisfied and the loss functions and $\widetilde{w}$ are defined as in equation 26. Now let us start at a point where the highest probabilities are assigned to all the bad samples, even in that case, the stationary point attained by MKL-SGD will be such that the highest probabilities are assigned to the good samples.*

7.2 Proofs and supporting lemmas

7.2.1 Proof of Lemma 1

Proof. Proof. $\widetilde{F}(\boldsymbol{w}) = \sum_i p_{m_i(\boldsymbol{w})}(\boldsymbol{w})$. Let us fix a $\boldsymbol{w}$ such that $p_i = p_i(\boldsymbol{w})$. We know that for any p_i, $\sum_i p_i f_i(\boldsymbol{w})$ is strongly convex in $\boldsymbol{w}$ with parameter $\lambda_{\boldsymbol{w}}$. This implies

$$\nabla \widetilde{F}(\boldsymbol{w})^\top (\boldsymbol{w} - \boldsymbol{w^*}) \geq \lambda_{\boldsymbol{w}} \|\boldsymbol{w} - \boldsymbol{w^*}\|^2$$

□

A naive bound for the above Lemma can be:

$$\nabla\widetilde{F}(\boldsymbol{w})^\top(\boldsymbol{w}-\boldsymbol{w^*}) \geq \min_i p_i \sum_i f_i(\boldsymbol{w}) \geq \underbrace{\lambda \min_i p_i}_{\lambda_{\boldsymbol{w}}} \|\boldsymbol{w}-\boldsymbol{w^*}\|^2$$

□

7.2.2 Proof of Theorem 1

Proof. By the definition of the noiseless framework, $\boldsymbol{w^*}$ is the unique optimum of $F(\boldsymbol{w})$ and lies in the optimal set of each $f_i(.)$. We will prove this theorem by contradiction. Assume there exists some $\hat{\boldsymbol{w}} \neq \boldsymbol{w^*}$ that also satisfies optimum of $\nabla\widetilde{F}(\hat{\boldsymbol{w}}) = \mathbf{0}$. At $\hat{\boldsymbol{w}}$, we have $0 =< \nabla\widetilde{F}(\hat{\boldsymbol{w}}), \hat{\boldsymbol{w}} - \boldsymbol{w^*} >= \lambda\|\hat{\boldsymbol{w}} - \boldsymbol{w^*}\|^2$. This implies $\hat{\boldsymbol{w}} = \boldsymbol{w^*}$. □

Theorem 1 and Assumption 2 guarantee that $\lambda_{\boldsymbol{w}} > 0$. If $f(\boldsymbol{w})$ is strongly convex and $g(\boldsymbol{w})$ is convex, then we know that $f(\boldsymbol{w}) + g(\boldsymbol{w})$ is strongly convex. On similar lines we can show that $\lambda > 0$ by splitting the terms in $F(\boldsymbol{w})$ as $p_{min}F(\boldsymbol{w})$ and $(\tilde{F}(\boldsymbol{w}) - p_{min}F(\boldsymbol{w}))$. The first term has $\lambda > 0$ (Assumption 2) and the second term has $\lambda = 0$ (since it is convex). Note, p_{min} is a positive constant independent of $\boldsymbol{w}$ and so the above lemma is for all $\boldsymbol{w}$.

7.2.3 Proof of the claim in Section 3

We will first describe the problem setting again for ease of analysis before elaborating on the proof.

Problem setting Let us assume good and bad samples are centered at the same point with different Lipschitz constants. The loss functions are given as follows:

$$f_i(w) = \begin{cases} l_i(w-w^*)^2 & \forall\ i \notin \mathbb{O} \\ l_i(w-w_B)^2 & \forall\ i \in \mathbb{O}, \end{cases} \tag{16}$$

where $|\mathbb{O}| = b$ such that $n = g + b$. Let $l_m = \min_{i\notin\mathbb{O}} l_i$ and Let $l_M = \max_{i\in\mathbb{O}} l_i$ and $l_{max} = \min_{i\in[n]} l_i$, $l_{min} = \min_{i\in[n]} l_i$. Let us define $\kappa = \dfrac{l_{max}}{l_{min}} \geq \dfrac{l_M}{l_m}$. Let us define $\widetilde{w}$ as follows:

$$\widetilde{w} := \left\{ w \mid \begin{array}{ll} w = \min\limits_{\alpha} \alpha w^* + (1-\alpha)w_B, & \alpha \in (0,1), \\ f_{l_m}(w) = f_{l_M}(w) & \end{array} \right\} \tag{17}$$

By observation, we know for the scalar case $\widetilde{w} = \dfrac{\sqrt{l_m}w^* + \sqrt{l_M}w_B}{\sqrt{l_m} + \sqrt{l_M}}$. Since we are initializing at w_0, the probability of picking bad samples is $\hat{p}$.

Proof. Assume we start at w_B, such that the outlier functions will have the highest weights. Without loss of generality, the probability of picking sample j in the bad set is $p_{m_j(w_B)}(w_B)$.

Let $\bar{w}$ indicate the stationary point of MKL-SGD assuming fixed $p_i(w_B)$ centered at w_B. The probabilities will not change until we reach $\widetilde{w}$. At $\bar{w}$, we have:

$$\begin{aligned} \sum_{i\notin\mathbb{O}} p_i(w_B)\nabla f_i(\bar{w}) &= -\sum_{j\in\mathbb{O}} p_j(w_B)\nabla f_j(\bar{w}) \\ \sum_{i\notin\mathbb{O}} p_i(w_B) l_i(\bar{w} - w^*) &= -\sum_{j\in\mathbb{O}} p_j(w_B) l_j(\bar{w} - w_B) \\ \bar{w} &= \frac{\sum_{i\notin\mathbb{O}} p_i(w_B) l_i w^* + \sum_{j\in\mathbb{O}} p_j(w_B) l_j w_B}{\sum_{i\notin\mathbb{O}} p_i(w_B) l_i + \sum_{j\in\mathbb{O}} p_j(w_B) l_j} \end{aligned}$$

If $\bar{w}$ is closer to w_B then $\widetilde{w}$, then the local minima $\bar{w}$ exists, else we will escape that local minima, since at $\widetilde{w}$, the probabilities change.

When is $\bar{w} < \widetilde{w}$?

A sufficient condition for that is:

$$\begin{aligned}
\frac{\sum_{j\in\mathbb{O}} p_j(w_B) l_j w_B}{\sum_{i\notin\mathbb{O}} p_i(w_B) l_i + \sum_{j\in\mathbb{O}} p_j(w_B) l_j} &\leq \frac{\sqrt{l_M} w_B}{\sqrt{l_m} + \sqrt{l_M}} \\
\frac{1}{1 + \frac{\sum_{i\notin\mathbb{O}} p_i(w_B) l_i}{\sum_{j\in\mathbb{O}} p_j(w_B) l_j}} &\leq \frac{1}{1 + \frac{\sqrt{l_m}}{\sqrt{l_M}}} \\
\frac{\sqrt{l_m}}{\sqrt{l_M}} &\leq \frac{\sum_{i\notin\mathbb{O}} p_i(w_B) l_i}{\sum_{j\in\mathbb{O}} p_j(w_B) l_j}
\end{aligned}$$

We just need,

$$\begin{aligned}
\frac{\sqrt{l_m}}{\sqrt{l_M}} &\leq \frac{\sum_{i\notin\mathbb{O}} p_i(w_B) l_m}{\sum_{j\in\mathbb{O}} p_j(w_B) l_M} \\
\frac{\sqrt{l_m}}{\sqrt{l_M}} &\leq \frac{(1-\hat{p}) l_m}{\hat{p} l_M} \\
\frac{\hat{p}}{1-\hat{p}} &\leq \sqrt{\frac{l_m}{l_M}} \\
\hat{p} &\leq \frac{1}{1 + \sqrt{\frac{l_M}{l_m}}}
\end{aligned}$$

If $\hat{p} \leq \frac{1}{1+\sqrt{\kappa}}$, then the above inequality is satisfied.

This was the condition for the point of intersection between the curves of good sample with the smallest l and the bad sample with the largest l. What happens between w^* and $\widetilde{w}$?

Next, we evaluate the closest point of intersection to $\widetilde{w}$ between w^* and $\widetilde{w}$ and do that recursively.

$$\begin{aligned}
\frac{\sqrt{l_{m1}}}{\sqrt{l_{M1}}} &\leq \frac{(1-\hat{p}) l_m}{\hat{p} l_M} \\
\hat{p} &\leq \frac{1}{1 + \sqrt{\frac{l_M}{l_m}} \sqrt{\frac{l_M}{l_{M1}}} \sqrt{\frac{l_{m1}}{l_m}}}
\end{aligned}$$

Similarly, if $\hat{p} \leq \frac{1}{1+\sqrt{\kappa}^{1.5}}$, then the above inequality is satisfied. □

7.2.4 Proof of Lemma 2

Let $\bar{\boldsymbol{w}}$ be a stationary point of MKL-SGD . Now, we analyze the loss landscape on the line joining $\boldsymbol{w^*}$ and $\boldsymbol{w}_C$ where $\boldsymbol{w}_C = C\bar{\boldsymbol{w}}$ is any arbitrary point [4] in the landscape at a distance as far as the farthest outlier from $\boldsymbol{w^*}$. Let C be a very large number.

[4] Note that we just need $\boldsymbol{w}_C$ for the purpose of landscape analysis and it is not a parameter of the algorithm

The loss functions and $\widetilde{\boldsymbol{w}}$ are redefined as follows:

$$f_i(\boldsymbol{w}) = \begin{cases} l_i\|\boldsymbol{w}-\boldsymbol{w}^*\|^2 & \forall\ i \in \mathbb{O} \\ l_i\|\boldsymbol{w}-\boldsymbol{w}_{b_i}\|^2 & \forall\ i \notin \mathbb{O}, \end{cases}$$

$$\widetilde{\boldsymbol{w}} := \left\{ \boldsymbol{w} \;\middle|\; \begin{array}{l} \boldsymbol{w} = \min\limits_{\alpha\in(0,1)} \alpha\boldsymbol{w}^* + (1-\alpha)\boldsymbol{w}_C, \\ f_{l_m}(\boldsymbol{w}) = f_{l_M}(\boldsymbol{w}) \end{array} \right\}$$

where $|\mathbb{O}| = b$ such that $n = g + b$. Let $l_m = \min_{i\notin\mathbb{O}} l_i$ and Let $l_M = \max_{i\in\mathbb{O}} l_i$ and $l_{max} = \min_{i\in[n]} l_i$, $l_{min} = \min_{i\in[n]} l_i$. Let us define $\kappa = \dfrac{l_{max}}{l_{min}} \geq \dfrac{l_M}{l_m}$.

Now at $\bar{w}$, we have $\nabla\widetilde{F}(\bar{\boldsymbol{w}}) = 0$. Let us assume that the outliers are chosen in such a way that at $\boldsymbol{w}_C$, all the outliers have the lowest loss. As stated in the previous lemma, the results hold irrespective of that. This implies:

$$\sum_{i\notin\mathbb{O}} p_i(\boldsymbol{w}_C)\nabla f_i(\bar{\boldsymbol{w}}) = -\sum_{j\in\mathbb{O}} p_j(\boldsymbol{w}_C)\nabla f_j(\bar{\boldsymbol{w}})$$

$$\sum_{i\notin\mathbb{O}} p_i(\boldsymbol{w}_C)l_i(\bar{\boldsymbol{w}}-\boldsymbol{w}^*) = -\sum_{j\in\mathbb{O}} p_j(\boldsymbol{w}_C)l_j(\bar{\boldsymbol{w}}-\boldsymbol{w}_{b_j})$$

$$\bar{\boldsymbol{w}} = \frac{\sum_{i\notin\mathbb{O}} p_i(\boldsymbol{w}_C)l_i\boldsymbol{w}^* + \sum_{j\in\mathbb{O}} p_j(\boldsymbol{w}_C)l_j\boldsymbol{w}_{b_j}}{\sum_{i\notin\mathbb{O}} p_i(\boldsymbol{w}_C)l_i + \sum_{j\in\mathbb{O}} p_j(\boldsymbol{w}_C)l_j}$$

$$\text{By triangle inequality, } \|\bar{\boldsymbol{w}}-\boldsymbol{w}^*\| \leq \frac{\sum_{j\in\mathbb{O}} p_j(\boldsymbol{w}_C)l_j\|\boldsymbol{w}_{b_j}-\boldsymbol{w}^*\|}{\sum_{i\notin\mathbb{O}} p_i(\boldsymbol{w}_C)l_i + \sum_{j\in\mathbb{O}} p_j(\boldsymbol{w}_C)l_j}$$

Without loss of generality assume that the outliers are ordered as follows: $\|\boldsymbol{w}_{b_1}-\boldsymbol{w}^*\| \leq \|\boldsymbol{w}_{b_2}-\boldsymbol{w}^*\| \leq \cdots \leq \|\boldsymbol{w}_{b_{|\mathbb{O}|}}-\boldsymbol{w}^*\|$.

Now $\widetilde{\boldsymbol{w}}$ be some point of intersection of function in the set of clean samples and a function in the set of outliers to $\boldsymbol{w}^*$. Let θ_j be the angle between the line connecting $\boldsymbol{w}_{b_j}$ and $\boldsymbol{w}^*$ to the line connecting $\boldsymbol{w}_C$ to $\boldsymbol{w}^*$. For any two curves with Lipschitz constants l_i and l_j, the halfspaces passing through the weighted mean are also the region where both functions have equal values.

Thus,

$$\widetilde{\boldsymbol{w}} = \frac{\sqrt{l_i}\boldsymbol{w}^* + \sqrt{l_j}\boldsymbol{w}_{b_j}}{\sqrt{l_i}+\sqrt{l_j}}$$

.

$$\|\widetilde{\boldsymbol{w}}-\boldsymbol{w}^*\| = \frac{\sqrt{l_j}\|\boldsymbol{w}_{b_j}-\boldsymbol{w}^*\|}{\sqrt{l_j}+\sqrt{l_i}}$$

Let γ denote the following ratio:

$$\gamma = \frac{\min_{j\in\mathbb{O}}\|\boldsymbol{w}_{b_j}-\boldsymbol{w}^*\|}{\max_{j\in\mathbb{O}}\|\boldsymbol{w}_{b_j}-\boldsymbol{w}^*\|} = \frac{2\delta}{\delta_{max}}$$

Now, we want:

$$\frac{\sum_{j\in\mathbb{O}} p_j(\boldsymbol{w}_C) l_j \left\|\boldsymbol{w}_{b_j} - \boldsymbol{w}^*\right\|}{\sum_{i\notin\mathbb{O}} p_i(\boldsymbol{w}_C) l_i + \sum_{j\in\mathbb{O}} p_j(\boldsymbol{w}_C) l_j} \leq \frac{\sqrt{l_{t_j}}}{\sqrt{l_{t_j}} + \sqrt{l_g}} \frac{\left\|\boldsymbol{w}_{b_j} - \boldsymbol{w}^*\right\|}{\cos\theta_j} = \frac{\|\widetilde{\boldsymbol{w}} - \boldsymbol{w}^*\|}{\cos\theta_j}$$

$$\frac{\sum_{j\in\mathbb{O}} p_j(\boldsymbol{w}_C) l_j \left\|\boldsymbol{w}_{b_j} - \boldsymbol{w}^*\right\|}{\sum_{i\notin\mathbb{O}} p_i(\boldsymbol{w}_C) l_i + \sum_{j\in\mathbb{O}} p_j(\boldsymbol{w}_C) l_j} \leq \frac{\sum_{j\in\mathbb{O}} p_j(\boldsymbol{w}_C) l_j \left\|\boldsymbol{w}_{b_{|\mathbb{O}|}} - \boldsymbol{w}^*\right\|}{\sum_{i\notin\mathbb{O}} p_i(\boldsymbol{w}_C) l_i + \sum_{j\in\mathbb{O}} p_j(\boldsymbol{w}_C) l_j} \leq \frac{\sqrt{l_{t_j}}}{\sqrt{l_{t_j}} + \sqrt{l_g}} \frac{\left\|\boldsymbol{w}_{b_j} - \boldsymbol{w}^*\right\|}{\cos\theta_j}$$

$$\frac{\sum_{j\in\mathbb{O}} p_j(\boldsymbol{w}_C) l_j}{\sum_{i\notin\mathbb{O}} p_i(\boldsymbol{w}_C) l_i + \sum_{j\in\mathbb{O}} p_j(\boldsymbol{w}_C) l_j} \leq \frac{\sqrt{l_{t_j}}}{\sqrt{l_{t_j}} + \sqrt{l_g}} \frac{\left\|\boldsymbol{w}_{b_j} - \boldsymbol{w}^*\right\|}{\cos\theta_j \left\|\boldsymbol{w}_{b_{|\mathbb{O}|}} - \boldsymbol{w}^*\right\|}$$

$$\frac{\sum_{j\in\mathbb{O}} p_j(\boldsymbol{w}_C) l_j}{\sum_{i\notin\mathbb{O}} p_i(\boldsymbol{w}_C) l_i + \sum_{j\in\mathbb{O}} p_j(\boldsymbol{w}_C) l_j} \leq \frac{\sqrt{l_{t_j}}}{\sqrt{l_{t_j}} + \sqrt{l_g}} \frac{\gamma}{\cos\theta_j}$$

For simplicity, $\Gamma = \dfrac{\gamma}{\cos\theta_j}$, then we have:

$$\frac{\sum_{j\in\mathbb{O}} p_j(\boldsymbol{w}_C) l_j}{\sum_{i\notin\mathbb{O}} p_i(\boldsymbol{w}_C) l_i + \sum_{j\in\mathbb{O}} p_j(\boldsymbol{w}_C) l_j} \leq \frac{\sqrt{l_{t_j}}}{\sqrt{l_{t_j}} + \sqrt{l_g}} \Gamma$$

$$\frac{1}{\Gamma}\left(\frac{\sqrt{l_g}}{\sqrt{l_{t_j}}} + 1\right) - 1 \leq \frac{(1-\hat{p}) l_m}{\hat{p} l_M} \leq \frac{\sum_{i\notin\mathbb{O}} p_i(\boldsymbol{w}_C) l_i}{\sum_{j\in\mathbb{O}} p_j(\boldsymbol{w}_C) l_j}$$

$$\frac{\hat{p}}{1-\hat{p}} \leq \frac{\frac{l_m}{l_M}}{\frac{1}{\Gamma} - 1 + \frac{1}{\Gamma}\frac{\sqrt{l_g}}{\sqrt{l_{t_j}}}}$$

$$\hat{p} \leq \frac{1}{1 + \kappa\left(\frac{1}{\Gamma} - 1 + \frac{\sqrt{\kappa}}{\Gamma}\right)} \leq \frac{1}{1 + \frac{l_M}{l_m}\left(\frac{1}{\Gamma} - 1 + \frac{1}{\Gamma}\frac{\sqrt{l_g}}{\sqrt{l_{t_j}}}\right)}$$

Replacing $\Gamma = \dfrac{\gamma}{\cos\theta_j}$, and let $q = \frac{\cos\theta_j}{\gamma} - 1 + \frac{\cos\theta_j\sqrt{\kappa}}{\gamma}$ the condition to guarantee that bad local minima do no exist is $\hat{p} \leq \dfrac{1}{1+\kappa q}$ and $q > 0$.

Note: In the vector case, for example there exists a fine tradeoff between how large θ_j can be and if for large θ_j, the loss corresponding to the outlier will be one of the lowest. Understanding that tradeoff is beyond the scope of this paper.

Note that, the lemma 2 leads to a very strong worst-case guarantee. It states that the farthest optimum will always be within a bowl of distance r from $\boldsymbol{w}^*$ no matter where we initialize. Moreover, as long as the condition is satisfied no matter where the outliers lie (can be adversarially chosen), MKL-SGD always has the propensity to bring the iterates to a ball of radius r around $\boldsymbol{w}^*$. However, when the necessary conditions for its convergence are violated, the guarantees are initialization dependent. Thus, all the discussions in the rest of this section will be with respect to these worst case guarantees. However, as we see in the experimental section for both neural networks and linear regression, random initialization also seems to perform better than SGD.

Effect of κ A direct result of Lemma 2 is that higher the condition number of the set of quadratic loss functions, lower is the fraction ϵ of outliers the MKL-SGD can tolerate. This is because large κ results in a small value of $\dfrac{1}{1+\kappa q}$. This implies that $\hat{p}$ has to be small which in turn requires smaller fractions fo corruptions, ϵ.

Effect of γ: The relative distance of the outliers from $\boldsymbol{w}^*$ plays a critical role in the condition for Lemma 2. We know that $\gamma \in (0, 1]$. $\gamma = 1$ implies the outliers are equidistant from the optimum $\boldsymbol{w}^*$. Low values

of γ lead to a large q leading to the violation of the condition with $\hat{p}$ (since RHS in the condition is very small), which implies that one bad outlier can guarantee that the condition in Lemma 2 are violated. The guarantees in the above lemma are only when the outliers are not adversarially chosen to lie at very high relative distances from $\boldsymbol{w}^*$. One way to avoid the set of outliers far far away from the optimum is to have a filtering step at the start of the algorithm like the one in Diakonikolas et al. [2018]. We will refer this in Experiments.

Effect of $\cos\theta_{j,\bar{\boldsymbol{w}}}$: At first glance, it may seem that $\cos\theta_{j,\bar{\boldsymbol{w}}} = 0$ may cause $1 + \kappa q < 0$ and since $\hat{p}(\boldsymbol{w}) > 0$, the condition in Lemma 2 may never be satisfied. Since, the term $\cos\theta_{j,\bar{\boldsymbol{w}}}$ shows up in the denominator of the loss associated with outlier centered at $\boldsymbol{w}_{b_j}$. Thus, low values of $\cos\theta_{j,\bar{\boldsymbol{w}}}$ implies high value of loss associated with the function centered at $\boldsymbol{w}_{b_j}$ which in turn implies the maximum probability attained by that sample can never be in the top-$|\mathbb{O}|$ probabilities for that $\bar{\boldsymbol{w}}$.

7.2.5 Proof of Lemma 3

Proof. At $\bar{\boldsymbol{w}}_{\boldsymbol{SGD}}$, $\nabla\widetilde{F}(\bar{\boldsymbol{w}}_{\boldsymbol{SGD}}) = 0$. Then,

$$\begin{aligned}
\sum_{i\notin\mathbb{O}} \nabla f_i(\bar{\boldsymbol{w}}_{\boldsymbol{SGD}}) &= -\sum_{i\in\mathcal{O}} \nabla f_i(\bar{\boldsymbol{w}}_{\boldsymbol{SGD}}) \\
\left\|\sum_{i\notin\mathbb{O}} \nabla f_i(\bar{\boldsymbol{w}}_{\boldsymbol{SGD}})\right\| &= \left\|\sum_{i\in\mathcal{O}} \nabla f_i(\bar{\boldsymbol{w}}_{\boldsymbol{SGD}})\right\| \\
\left\|\sum_{i\notin\mathbb{O}} \nabla f_i(\bar{\boldsymbol{w}}_{\boldsymbol{SGD}})\right\| &\leq \sum_i \|\nabla f_i(\bar{\boldsymbol{w}}_{\boldsymbol{SGD}})\| \\
&\leq \sum_i L\|\bar{\boldsymbol{w}}_{\boldsymbol{SGD}} - \boldsymbol{w}^*\| \\
&= (1-\epsilon)nL\|\bar{\boldsymbol{w}}_{\boldsymbol{SGD}} - \boldsymbol{w}^*\| \qquad (18) \\
\left\|\sum_{i\in\mathcal{O}} \nabla f_i(\bar{\boldsymbol{w}}_{\boldsymbol{SGD}})\right\| &\leq \sum_{i\in\mathcal{O}} \|\nabla f_i(\bar{\boldsymbol{w}}_{\boldsymbol{SGD}})\| \\
&\leq \sum_{i\in\mathcal{O}} G(\bar{\boldsymbol{w}}_{\boldsymbol{SGD}}) \\
&\leq \epsilon G(\bar{\boldsymbol{w}}_{\boldsymbol{SGD}}) \qquad (19)
\end{aligned}$$

$$\left\|\sum_{i\in\mathcal{O}} \nabla f_i(\bar{\boldsymbol{w}}_{\boldsymbol{SGD}})\right\| = \min\left(\epsilon nG(\bar{\boldsymbol{w}}_{\boldsymbol{SGD}}), (1-\epsilon)nL\|\bar{\boldsymbol{w}}_{\boldsymbol{SGD}} - \boldsymbol{w}^*\|\right)$$

□

7.2.6 Proof of Lemma 4

Proof. At $\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}$, $\nabla \widetilde{F}(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}) = 0$. This implies

$$\sum_{i \notin \mathbb{O}} p_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}) \nabla f_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}) = -\sum_{i \in \mathcal{O}} p_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}) \nabla f_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}})$$

Multiplying both sides by $(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}} - \boldsymbol{w}^*)$

$$\sum_{i \notin \mathbb{O}} p_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}) < \nabla f_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}), \bar{\boldsymbol{w}}_{\boldsymbol{MKL}} - \boldsymbol{w}^* > = -\sum_{i \in \mathcal{O}} p_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}) < \nabla f_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}), \bar{\boldsymbol{w}}_{\boldsymbol{MKL}} - \boldsymbol{w}^* > \quad (20)$$

$$< \nabla \widetilde{F}_{\mathcal{G}}(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}), \bar{\boldsymbol{w}}_{\boldsymbol{MKL}} - \boldsymbol{w}^* > = -\sum_{i \in \mathcal{O}} p_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}) < \nabla f_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}), \bar{\boldsymbol{w}}_{\boldsymbol{MKL}} - \boldsymbol{w}^* >$$

Lower bounding the LHS using Lemma 1 and $m = m(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}})$[5] ,

(21)

$$m\|\bar{\boldsymbol{w}}_{\boldsymbol{MKL}} - \boldsymbol{w}^*\|^2 \leq \left\| < \nabla \widetilde{F}_{\mathcal{G}}(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}), \bar{\boldsymbol{w}}_{\boldsymbol{MKL}} - \boldsymbol{w}^* > \right\| = LHS \quad (22)$$

$$RHS \leq \left\| -\sum_{i \in \mathcal{O}} p_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}) < \nabla f_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}), \bar{\boldsymbol{w}}_{\boldsymbol{MKL}} - \boldsymbol{w}^* > \right\|$$

$$m\|\bar{\boldsymbol{w}}_{\boldsymbol{MKL}} - \boldsymbol{w}^*\|^2 \leq \sum_{i \in \mathcal{O}} p_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}) \| < \nabla f_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}), \bar{\boldsymbol{w}}_{\boldsymbol{MKL}} - \boldsymbol{w}^* > \|$$

$$m\|\bar{\boldsymbol{w}}_{\boldsymbol{MKL}} - \boldsymbol{w}^*\|^2 \leq \sum_{i \in \mathcal{O}} p_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}) \|\nabla f_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}})\| \|\bar{\boldsymbol{w}}_{\boldsymbol{MKL}} - \boldsymbol{w}^*\|$$

$$m\|\bar{\boldsymbol{w}}_{\boldsymbol{MKL}} - \boldsymbol{w}^*\|^2 \leq \sum_{i \in \mathcal{O}} p_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}) \|\bar{\boldsymbol{w}}_{\boldsymbol{MKL}} - \boldsymbol{w}^*\| G(\bar{\boldsymbol{w}}_{\boldsymbol{SGD}})$$

$$m\|\bar{\boldsymbol{w}}_{\boldsymbol{MKL}} - \boldsymbol{w}^*\| \leq \epsilon^k G(\bar{\boldsymbol{w}}_{\boldsymbol{SGD}}) \quad (23)$$

□

7.2.7 Proof of Theorem 2

Proof. There exists an $\epsilon' \leq \epsilon$ such that in Lemma 3, we have

$$(1-\epsilon) L \|\bar{\boldsymbol{w}}_{\boldsymbol{SGD}} - \boldsymbol{w}^*\| \geq \epsilon G(\bar{\boldsymbol{w}}_{\boldsymbol{SGD}})$$

Combining above equation with Lemma 4, we get

$$(1-\epsilon) L \|\bar{\boldsymbol{w}}_{\boldsymbol{SGD}} - \boldsymbol{w}^*\| \geq \epsilon G(\bar{\boldsymbol{w}}_{\boldsymbol{SGD}}) \geq \epsilon \frac{\lambda}{\epsilon^2} \|\bar{\boldsymbol{w}}_{\boldsymbol{MKL}} - \boldsymbol{w}^*\|$$

$$\Rightarrow \|\bar{\boldsymbol{w}}_{\boldsymbol{MKL}} - \boldsymbol{w}^*\| \leq \frac{(1-\epsilon) L \epsilon^{k-1}}{\lambda} \|\bar{\boldsymbol{w}}_{\boldsymbol{SGD}} - \boldsymbol{w}^*\|$$

Picking a large enough k, we can guarantee that $\frac{(1-\epsilon) L \epsilon^{k-1}}{\lambda} < 1$ □

7.2.8 Proof of Lemma 6

Proof. From the definition of good samples in the noiseless setting, we know that $f_i(\boldsymbol{w}^*) = 0 \ \forall \ i \notin \mathbb{O}$. Similarly, for samples belonging to the outlier set, $f_i(\boldsymbol{w}^*) > 0 \ \forall \ i \in \mathcal{O}$. There exists a ball around the optimum of radius r such that $f_i(\boldsymbol{w}) \leq f_j(\boldsymbol{w}) \ \forall i \notin \mathbb{O}, j \in \mathcal{O}, \boldsymbol{w} \in \mathbb{O}_r(\boldsymbol{w}^*)$. Assume that $N' \geq N$ and $\epsilon' \leq \epsilon$, such that $\|\bar{\boldsymbol{w}}_{\boldsymbol{MKL}} - \boldsymbol{w}^*\| \leq r$.

At $\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}$, $\nabla \widetilde{F}(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}) = 0$. This implies

$$\sum_{i \notin \mathbb{O}} p_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}) \nabla f_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}) = -\sum_{i \in \mathcal{O}} p_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}) \nabla f_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}})$$

$$\left\| \sum_{i \notin \mathbb{O}} p_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}) \nabla f_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}) \right\| = \left\| \sum_{i \in \mathcal{O}} p_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}) \nabla f_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}) \right\|$$

$$\begin{aligned}
\left\| \sum_{i \notin \mathbb{O}} p_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}) \nabla f_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}) \right\| &\leq \sum_i p_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}) \|\nabla f_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}})\| \\
&\leq \sum_i p_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}) L \|\bar{\boldsymbol{w}}_{\boldsymbol{MKL}} - \boldsymbol{w}^*\| \\
&= (1 - \epsilon^k) L \|\bar{\boldsymbol{w}}_{\boldsymbol{MKL}} - \boldsymbol{w}^*\| \qquad (24)
\end{aligned}$$

$$\begin{aligned}
\left\| \sum_{i \in \mathcal{O}} p_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}) \nabla f_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}) \right\| &\leq \sum_{i \in \mathcal{O}} p_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}) \|\nabla f_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}})\| \\
&\leq \sum_{i \in \mathcal{O}} p_i(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}) G(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}) \\
&\leq \epsilon^k G(\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}) \qquad (25)
\end{aligned}$$

□

7.2.9 Proof of Lemma 7

Problem Setting: We will assume that without loss of generality all the outliers will lie on the same side of w^*. If that's not the case, the bounds which show in the subsequent part will be even stronger. Without loss of generality, assume $0 < w_{b_1} < w_{b_2} < \cdots < w_{b_{|\mathcal{B}|}}$.

The loss functions and $\widetilde{w}$ are redefined as follows:

$$f_i(w) = \begin{cases} l_i(w - w^*)^2 & \forall\ i \notin \mathbb{O} \\ l_i(w - w_{b_i})^2 & \forall\ i \in \mathbb{O}, \end{cases} \qquad (26)$$

$$\widetilde{w} := \left\{ w \,\middle|\, w = \min_{\lambda \in (0,1)} \left[\lambda w^* + (1 - \lambda) \min_j w_{b_j} \right], f_{l_m}(w) = f_{l_M}(w) \right\} \qquad (27)$$

Once again by simple analysis of different points of intersection, we can describe a closed form expression of $\widetilde{\boldsymbol{w}}$ as follows:

$$\widetilde{w} = \frac{\sqrt{l_m} w^* + \sqrt{l_M} w_{b_M}}{\sqrt{l_m} + \sqrt{l_M}}$$

where $\widetilde{M} = \arg\max_{j \in \mathbb{O}} \sqrt{l_j} w_{b_j}$, $M = \arg\max_{i \in \mathbb{O}} l_i$, $m = \arg\max_{i \notin \mathbb{O}} l_i$.

Here, $|\boldsymbol{\mathcal{G}}| = g$ and $|\mathbb{O}| = b$ such that $n = g + b$. Let $l_m = \min_{i \notin \mathbb{O}} l_i$ and Let $l_M = \max_{i \in \mathbb{O}} l_i$ and $l_{max} = \min_{i \in [n]} l_i$, $l_{min} = \min_{i \in [n]} l_i$. Let us define $\kappa = \dfrac{l_{max}}{l_{min}} \geq \dfrac{l_M}{l_m}$. Let $\gamma = \dfrac{w_{b_1}}{w_{b_{|B|}}}$

Proof. Assume we start at w_{b_1}, such that the outlier functions will have the highest weights. If any one of the outlier functions does not have the top $|B|$ probabilities at w_{b_1}, then $\hat{p}$ in the subsequent bounds will be smaller and will still satisfy the final condition. Without loss of generality, the probability of picking sample j in the bad set is $p_{m_j(w_{b_1})}(w_{b_1})$. Note that, we assume that $\widetilde{w}$ lies between w^* and w_{b_1} and the outlier functions have the highest weights at w_{b_1}. If that's not the case, our bounds are still satisfied since there will be atleast one term which will not be a part of $\hat{p}$ and so $\bar{w}$ will be even closer to w^*.

Let $\bar{w}$ indicate the stationary point of MKL-SGD assuming fixed $p_i(w_B)$ centered at w_B. The probabilities will not change until we reach $\widetilde{w}$. At $\bar{w}$, we have:

$$\sum_{i \notin \mathbb{O}} \nabla f_i(\bar{w}) = -\sum_{j \in \mathbb{O}} \nabla f_j(\bar{w})$$

$$\sum_{i \notin \mathbb{O}} p_i(w_{b_1}) l_i(\bar{w} - w^*) = -\sum_{j \in \mathbb{O}} p_j(w_{b_1}) l_j(\bar{w} - w_{b_1})$$

$$\bar{w} = \frac{\sum_{i \notin \mathbb{O}} p_i(w_{b_1}) l_i w^* + \sum_{j \in \mathbb{O}} p_j(w_{b_1}) l_j w_{b_1}}{\sum_{i \notin \mathbb{O}} p_i(w_{b_1}) l_i + \sum_{j \in \mathbb{O}} p_j(w_{b_1}) l_j}$$

If $\bar{w}$ is closer to w_{b_1} then $\widetilde{w}$, the local minima $\bar{w}$ exists, since the probabilities don't change before we reach $\widetilde{w}$ starting from w_{b_1} moving towards w^*.

When is $\bar{w} < \widetilde{w}$

A sufficient condition for that is:

$$\frac{\sum_{j \in \mathbb{O}} p_j(w_{b_1}) l_j w_{b_j}}{\sum_{i \notin \mathbb{O}} p_i(w_{b_1}) l_i + \sum_{j \in \mathbb{O}} p_j(w_{b_1}) l_j} \leq \frac{\sum_{j \in \mathbb{O}} p_j(w_{b_1}) l_j w_{b_M}}{\sum_{i \notin \mathbb{O}} p_i(w_{b_1}) l_i + \sum_{j \in \mathbb{O}} p_j(w_{b_1}) l_j} \leq \frac{\sqrt{l_{\tilde{M}}} w_{b_{\tilde{M}}}}{\sqrt{l_m} + \sqrt{l_{\tilde{M}}}}$$

$$\frac{1}{1 + \frac{\sum_{i \notin \mathbb{O}} p_i(w_{b_1}) l_i}{\sum_{j \in \mathbb{O}} p_j(w_{b_1}) l_j}} \leq \frac{\gamma}{1 + \frac{\sqrt{l_m}}{\sqrt{l_{\tilde{M}}}}} \leq \frac{\frac{w_{b_{\tilde{M}}}}{w_{b_M}}}{1 + \frac{\sqrt{l_m}}{\sqrt{l_{\tilde{M}}}}}$$

$$\frac{\sqrt{l_m}}{\sqrt{l_{\tilde{M}}}} \leq \frac{\sum_{i \notin \mathbb{O}} p_i(w_{b_1}) l_i}{\sum_{j \in \mathbb{O}} p_j(w_{b_1}) l_j}$$

We have,

$$1 + \frac{\sqrt{l_m}}{\sqrt{l_{\tilde{M}}}} \leq \gamma \left(1 + \frac{(1-\hat{p}) l_m}{\hat{p} l_M}\right) \leq \gamma \left(1 + \frac{\sum_{i \notin \mathbb{O}} p_i(w_{b_1}) l_i}{\sum_{j \in \mathbb{O}} p_j(w_{b_1}) l_j}\right)$$

$$\frac{1}{\gamma}\left(1 + \frac{\sqrt{l_m}}{\sqrt{l_{\tilde{M}}}}\right) - 1 \leq \frac{(1-\hat{p}) l_m}{\hat{p} l_M}$$

$$\frac{\hat{p} l_M}{(1-\hat{p}) l_m} \leq \frac{1}{\frac{1}{\gamma}\left(1 + \frac{\sqrt{l_m}}{\sqrt{l_{\tilde{M}}}}\right) - 1}$$

$$\frac{\hat{p}}{(1-\hat{p})} \leq \frac{l_m}{l_M} \frac{1}{\frac{1}{\gamma}\left(1 + \frac{\sqrt{l_m}}{\sqrt{l_{\tilde{M}}}}\right) - 1}$$

$$\hat{p} \leq \frac{1}{1 + \frac{l_M}{l_m}\left(\frac{1}{\gamma}\left(1 + \frac{\sqrt{l_m}}{\sqrt{l_{\tilde{M}}}}\right) - 1\right)}$$

$$\leq \frac{1}{1 + \frac{l_M}{l_m}\left(\frac{1}{\gamma} - 1 + \frac{\sqrt{l_m}}{\gamma\sqrt{l_{\tilde{M}}}}\right)}$$

Thus, for the minima to not exist, we need: $\hat{p} \leq \frac{1}{1 + \kappa\left(\left(\frac{1}{\gamma} - 1\right) + \frac{\sqrt{\kappa}}{\gamma}\right)}$

The condition for any other pair of l_i, l_j is similar. This is because if we replace m by m_1 in the square root,

the final step of the analysis simplifies to:

$$\hat{p} \leq \frac{1}{1+\frac{l_M}{l_m}\left(\left(\frac{1}{\gamma}-1\right)+\sqrt{\frac{l_{m_1}}{l_{\tilde{M}}}}\frac{1}{\gamma}\right)} \tag{28}$$

If $\hat{p} \leq \frac{1}{1+\kappa\left(\left(\frac{1}{\gamma}-1\right)+\frac{\sqrt{\kappa}}{\gamma}\right)}$, then the above equation is satisfied and thus the equivalent condition still remains the same. □

7.3 Additional results and proofs for Section 5

Consider the sample size n with bad set(outlier) $\mathbb{O}$ and good set $\mathcal{G}$ such that $|\mathcal{G}| = n - |\mathbb{O}|$. Define

$$F_{good}(\boldsymbol{w}) = \frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}} f_i(\boldsymbol{w}).$$

We assume:
(1) (Stationary Point) Assume $\boldsymbol{w}^*$ is the solution for the average loss function of good sample such that

$$\nabla F_{good}(\boldsymbol{w}^*) = 0 \quad \text{ but } \nabla f_i(\boldsymbol{w}^*) \neq 0, \forall i \in \mathbb{O}$$

(2) (Strong Convexity) $F_{good}(\boldsymbol{w})$ is strongly convex with parameters λ_{good} i.e.,

$$\langle \nabla F_{good}(\boldsymbol{w}) - \nabla F_{good}(\boldsymbol{w}^*), \boldsymbol{w} - \boldsymbol{w}^* \rangle \ \geq \lambda_{good}\|\boldsymbol{w} - \boldsymbol{w}^*\|^2$$

(3) (Gradient Lipschitz) $f_i(\boldsymbol{w})$ has L_i Liptchitz gradient i.e.,

$$\|\nabla f_i(\boldsymbol{w}) - \nabla f_i(\boldsymbol{w}^*)\| \ \leq L_i\|\boldsymbol{w} - \boldsymbol{w}^*\|$$

Theorem 4. ***(Distance to $\boldsymbol{w}^*$)***

$$\mathbb{E}_i\left[\|\boldsymbol{w_{t+1}} - \boldsymbol{w}^*\|^2 | \boldsymbol{w_t}\right] \leq \left(1 - 2\eta_t\lambda_{good}(1-\eta_t\sup_i L_i)\min_{i\in\mathcal{G}} p_i(\boldsymbol{w_t})\right)\|\boldsymbol{w_t} - \boldsymbol{w}^*\|^2 + R_t \tag{29}$$

where

$$\begin{aligned} R_t =& -2\eta_t\sum_{i\in\mathcal{G}} p_i(\boldsymbol{w_t})\langle \boldsymbol{w_t} - \boldsymbol{w}^*, \nabla f_i(\boldsymbol{w}^*)\rangle \\ &+ 2\eta_t^2\sum_{i\in\mathcal{G}} p_i(\boldsymbol{w_t})\|\nabla f_i(\boldsymbol{w}^*)\|^2 + \eta_t^2\sum_{i\in\mathbb{O}} p_i(\boldsymbol{w_t})\|\nabla f_i(\boldsymbol{w_t})\|^2 + 2\eta_t\sum_{i\in\mathbb{O}} p_i(\boldsymbol{w_t})\left(f_i(\boldsymbol{w}^*) - f_i(\boldsymbol{w_t})\right) \end{aligned}$$

Proof. Observe first that for each component function i.e. ,

$$\langle \boldsymbol{w} - \boldsymbol{v}, \nabla f_i(\boldsymbol{w}) - \nabla f_i(\boldsymbol{v})\rangle \geq \frac{1}{L_i}\|f_i(\boldsymbol{w}) - f_i(\boldsymbol{v})\|^2$$

For detailed proof, see Lemma A.1 in [Needell et al., 2014].

For each individual component function $f_i(\boldsymbol{w})$, we have

$$\begin{aligned} \|\boldsymbol{w_{t+1}} - \boldsymbol{w}^*\|^2 =& \|\boldsymbol{w_t} - \boldsymbol{w}^*\|^2 + \eta_t^2\|\nabla f_i(\boldsymbol{w_t})\|^2 - 2\eta_t\langle \boldsymbol{w_t} - \boldsymbol{w}^*, \nabla f_i(\boldsymbol{w_t})\rangle \\ \leq& \|\boldsymbol{w_t} - \boldsymbol{w}^*\|^2 + 2\eta_t^2\|\nabla f_i(\boldsymbol{w_t}) - \nabla f_i(\boldsymbol{w}^*)\|^2 + 2\eta_t^2\|\nabla f_i(\boldsymbol{w}^*)\|^2 - 2\eta_t\langle \boldsymbol{w_t} - \boldsymbol{w}^*, \nabla f_i(\boldsymbol{w_t})\rangle \\ \leq& \|\boldsymbol{w_t} - \boldsymbol{w}^*\|^2 + 2\eta_t^2 L_i\langle \boldsymbol{w_t} - \boldsymbol{w}^*, \nabla f_i(\boldsymbol{w_t}) - \nabla f_i(\boldsymbol{w}^*)\rangle + 2\eta_t^2\|\nabla f_i(\boldsymbol{w}^*)\|^2 \\ & - 2\eta_t\langle \boldsymbol{w_t} - \boldsymbol{w}^*, \nabla f_i(\boldsymbol{w_t})\rangle \\ =& \|\boldsymbol{w_t} - \boldsymbol{w}^*\|^2 - 2\eta_t(1-\eta_t\sup_i L_i)\langle \boldsymbol{w_t} - \boldsymbol{w}^*, \nabla f_i(\boldsymbol{w_t}) - \nabla f_i(\boldsymbol{w}^*)\rangle + 2\eta_t^2\|\nabla f_i(\boldsymbol{w}^*)\|^2 \\ & - 2\eta_t\langle \boldsymbol{w_t} - \boldsymbol{w}^*, \nabla f_i(\boldsymbol{w}^*)\rangle \end{aligned}$$

We next take an expectation with respect to the choice of i conditional on $\boldsymbol{w_t}$

$$\begin{aligned}\mathbb{E}_i\left[\|\boldsymbol{w_{t+1}}-\boldsymbol{w^*}\|^2|\boldsymbol{w_t}\right] \leq & \|\boldsymbol{w_t}-\boldsymbol{w^*}\|^2 - 2\eta_t(1-\eta_t\sup_i L_i)\underbrace{\left\langle \boldsymbol{w_t}-\boldsymbol{w^*}, \sum_{i\in\mathcal{G}} p_i(\boldsymbol{w_t})\left(\nabla f_i(\boldsymbol{w_t})-\nabla f_i(\boldsymbol{w^*})\right)\right\rangle}_{Term1} \\ & - 2\eta_t\langle \boldsymbol{w_t}-\boldsymbol{w^*}, \sum_{i\in\mathcal{G}} p_i(\boldsymbol{w_t})\nabla f_i(\boldsymbol{w^*})\rangle + 2\eta_t^2\sum_{i\in\mathcal{G}} p_i(\boldsymbol{w_t})\|\nabla f_i(\boldsymbol{w^*})\|^2 \\ & + \eta_t^2\sum_{i\in\mathbb{O}} p_i(\boldsymbol{w_t})\|\nabla f_i(\boldsymbol{w_t})\|^2 + 2\eta_t\underbrace{\langle \boldsymbol{w^*}-\boldsymbol{w_t}, \sum_{i\in\mathbb{O}} p_i(\boldsymbol{w_t})\nabla f_i(\boldsymbol{w_t})\rangle}_{Term2} \end{aligned} \tag{30}$$

Now we first bound $Term1$ as follows

$$\begin{aligned} Term1 &\leq \min_{i\in\mathcal{G}} p_i(\boldsymbol{w_t})\sum_{i\in\mathcal{G}}\left\langle \boldsymbol{w_t}-\boldsymbol{w^*}, \nabla f_i(\boldsymbol{w_t})-\nabla f_i(\boldsymbol{w^*})\right\rangle \\ &\leq \min_{i\in\mathcal{G}} p_i(\boldsymbol{w_t})\lambda_{good}\|\boldsymbol{w_{t+1}}-\boldsymbol{w^*}\|^2 \end{aligned}$$

For $Term2$ we apply the property of the convex function $\langle\nabla f_i(\boldsymbol{v}), \boldsymbol{w}-\boldsymbol{v}\rangle \leq f_i(\boldsymbol{w}) - f_i(\boldsymbol{v})$

$$Term2 \leq \sum_{i\in\mathbb{O}} p_i(\boldsymbol{w_t})\left(f_i(\boldsymbol{w^*})-f_i(\boldsymbol{w_t})\right)$$

Putting the upper bound of $Term1$ and $Term2$ back to (30) gives

$$\mathbb{E}_i\left[\|\boldsymbol{w_{t+1}}-\boldsymbol{w^*}\|^2|\boldsymbol{w_t}\right] \leq \left(1-2\eta_t\lambda_{good}(1-\eta_t\sup_i L_i)\min_{i\in\mathcal{G}} p_i(\boldsymbol{w_t})\right)\|\boldsymbol{w_t}-\boldsymbol{w^*}\|^2 + R_t \tag{31}$$

where

$$\begin{aligned} R_t = & -2\eta_t\sum_{i\in\mathcal{G}} p_i(\boldsymbol{w_t})\langle \boldsymbol{w_t}-\boldsymbol{w^*}, \nabla f_i(\boldsymbol{w^*})\rangle \\ & + 2\eta_t^2\sum_{i\in\mathcal{G}} p_i(\boldsymbol{w_t})\|\nabla f_i(\boldsymbol{w^*})\|^2 + \eta_t^2\sum_{i\in\mathbb{O}} p_i(\boldsymbol{w_t})\|\nabla f_i(\boldsymbol{w_t})\|^2 + 2\eta_t\sum_{i\in\mathbb{O}} p_i(\boldsymbol{w_t})\left(f_i(\boldsymbol{w^*})-f_i(\boldsymbol{w_t})\right) \end{aligned}$$

□

We have the following corollary that for noiseless setting, if we can have some good initialization, MKL-SGD is always better than SGD even the corrupted data is greater than half. For noisy setting, we can also perform better than SGD with one more condition: the noise is not large than the distance $\|\Delta_t\|^2$. This condition is not mild in the sense that $\|\boldsymbol{w}_t-\boldsymbol{w^*}\|^2$ is always greater than $\|\bar{\boldsymbol{w}}_{\boldsymbol{SGD}}-\boldsymbol{w^*}\|^2$ for SGD algorithm and $\|\bar{\boldsymbol{w}}_{\boldsymbol{MKL}}-\boldsymbol{w^*}\|^2$ for MKL-SGD.

Corollary 1. *Suppose we have $|\mathcal{G}| \leq \frac{n}{2}$. At iteration t for $\eta_t \leq \frac{1}{\sup_i L_i}$, the parameter $\boldsymbol{w_t}$ satisfies $\sup_{i\in\mathcal{G}} f_i(\boldsymbol{w_t}) \leq \inf_{j\in\mathbb{O}} f_j(\boldsymbol{w_t})$. Moreover, assume the noise level at optimal $\boldsymbol{w^*}$ satisfies*

$$\textit{either} \quad \|\nabla f_i(\boldsymbol{w^*})\| \leq \frac{\lambda_{good}(1-\eta_t\sup_i L_i)/n}{1+\sqrt{1+\eta_t(1-\eta_t\sup_i L_i)\lambda_{good}/n}}\|\boldsymbol{w_t}-\boldsymbol{w^*}\|, \ \textit{for } i\in\mathcal{G} \tag{32}$$

$$\textit{or} \quad \sum_{i\in\mathcal{G}}\|\nabla f_i(\boldsymbol{w^*})\|^2 \leq \left(\frac{\lambda_{good}(1-\eta_t\sup_i L_i)|\mathcal{G}|/n}{\sqrt{n}+\sqrt{\sqrt{n}+\eta_t(1-\eta_t\sup_i L_i)\lambda_{good}|\mathcal{G}|/n}}\right)^2\|\boldsymbol{w_t}-\boldsymbol{w^*}\|^2. \tag{33}$$

Using the same setup, the vanilla SGD and MKL-SGD (K=2) algorithms yield respectively

$$\textbf{\textit{SGD}} \quad \mathbb{E}_i\left[\|\boldsymbol{w_{t+1}}-\boldsymbol{w^*}\|^2|\boldsymbol{w_t}\right] \leq \left(1-2\eta_t\lambda_{good}(1-\eta_t\sup_i L_i)\frac{|\mathcal{G}|}{n}\right)\|\boldsymbol{w_t}-\boldsymbol{w^*}\|^2 + R_t^{(SGD)}$$

$$\textbf{\textit{MKL-2}} \quad \mathbb{E}_i\left[\|\boldsymbol{w_{t+1}}-\boldsymbol{w^*}\|^2|\boldsymbol{w_t}\right] \leq \left(1-2\eta_t\lambda_{good}(1-\eta_t\sup_i L_i)\frac{|\mathcal{G}|}{n}\right)\|\boldsymbol{w_t}-\boldsymbol{w^*}\|^2 + R_t^{(MKL_2)}$$

where

$$R_t^{(MKL_2)} \leq R_t^{(SGD)}.$$

Proof. Start from the inequality 30 in the proof of Theorem 4. We have $Term1$ as follows:

$$\begin{aligned}
Term1 =& \frac{|\mathcal{G}|}{n}\left\langle \boldsymbol{w_t} - \boldsymbol{w}^*, \sum_{i\in\mathcal{G}} \frac{p_i(\boldsymbol{w_t})}{|\mathcal{G}|/n}\left(\nabla f_i(\boldsymbol{w_t}) - \nabla f_i(\boldsymbol{w}^*)\right)\right\rangle \\
=& \frac{|\mathcal{G}|}{n}\langle \boldsymbol{w_t} - \boldsymbol{w}^*, \nabla F_{good}(\boldsymbol{w_t}) - \nabla F_{good}(\boldsymbol{w}^*)\rangle \\
&+ \frac{|\mathcal{G}|}{n}\sum_{i\in\mathcal{G}}\left(\frac{p_i(\boldsymbol{w_t})}{|\mathcal{G}|/n} - \frac{1}{|\mathcal{G}|}\right)\langle \boldsymbol{w_t} - \boldsymbol{w}^*, \nabla f_i(\boldsymbol{w_t}) - \nabla f_i(\boldsymbol{w}^*)\rangle \\
\geq& \lambda_{good}\frac{|\mathcal{G}|}{n}\|\boldsymbol{w_t} - \boldsymbol{w}^*\|^2 + \sum_{i\in\mathcal{G}}\left(p_i(\boldsymbol{w_t}) - \frac{1}{n}\right)\langle \boldsymbol{w_t} - \boldsymbol{w}^*, \nabla f_i(\boldsymbol{w_t}) - \nabla f_i(\boldsymbol{w}^*)\rangle
\end{aligned}$$

Putting the terms back to (30), we have for $\eta_t \leq 1/(\sup_i L_i)$

$$\mathbb{E}_i\left[\|\boldsymbol{w_{t+1}} - \boldsymbol{w}^*\|^2 | \boldsymbol{w_t}\right] \leq \left(1 - 2\eta_t\lambda_{good}(1 - \eta_t \sup_i L_i)\frac{|\mathcal{G}|}{n}\right)\|\boldsymbol{w_t} - \boldsymbol{w}^*\|^2 + R_t \tag{34}$$

where

$$\begin{aligned}
R_t =& -2\eta_t(1 - \eta_t \sup_i L_i))\sum_{i\in\mathcal{G}}\left(p_i(\boldsymbol{w_t}) - \frac{1}{n}\right)\langle \boldsymbol{w_t} - \boldsymbol{w}^*, \nabla f_i(\boldsymbol{w_t}) - \nabla f_i(\boldsymbol{w}^*)\rangle \\
&- 2\eta_t\sum_{i\in\mathcal{G}} p_i(\boldsymbol{w_t})\langle \boldsymbol{w_t} - \boldsymbol{w}^*, \nabla f_i(\boldsymbol{w}^*)\rangle \\
&+ 2\eta_t^2\sum_{i\in\mathcal{G}} p_i(\boldsymbol{w_t})\|\nabla f_i(\boldsymbol{w}^*)\|^2 + \eta_t^2\sum_{i\in\mathbb{O}} p_i(\boldsymbol{w_t})\|\nabla f_i(\boldsymbol{w_t})\|^2 + 2\eta_t\sum_{i\in\mathbb{O}} p_i(\boldsymbol{w_t})\left(f_i(\boldsymbol{w}^*) - f_i(\boldsymbol{w_t})\right)
\end{aligned}$$

Now we analyse the term R_t for vanilla SGD and MKL-SGD($K = 2$) respectively. For vanilla SGD, we have $p_i(\boldsymbol{w_t}) = \frac{1}{n}$ and $\sum_{i\in\mathcal{G}}\nabla f_i(\boldsymbol{w}^*) = 0$, which results in

$$R_t^{(SGD)} = \frac{2\eta_t^2}{n}\sum_{i\in\mathcal{G}}\|\nabla f_i(\boldsymbol{w}^*)\|^2 + \frac{\eta_t^2}{n}\sum_{i\in\mathbb{O}}\|\nabla f_i(\boldsymbol{w_t})\|^2 + \frac{2\eta_t}{n}\sum_{i\in\mathbb{O}}\left(f_i(\boldsymbol{w}^*) - f_i(\boldsymbol{w_t})\right)$$

Note that MKL-SGD for $K = 2$ have

$$p_{m_i(\boldsymbol{w})}(\boldsymbol{w}) = \frac{2(n-i)}{n(n-1)} \tag{35}$$

where $m_1(\boldsymbol{w}), m_2(\boldsymbol{w}), m_3(\boldsymbol{w}), \ldots m_n(\boldsymbol{w})$ are the indices of data samples for some $\boldsymbol{w}$:

$$f_{m_1(\boldsymbol{w})}(\boldsymbol{w}) \leq f_{m_2(\boldsymbol{w})}(\boldsymbol{w}) \leq \cdots \leq f_{m_n(\boldsymbol{w})}(\boldsymbol{w})$$

Suppose the iteration $\boldsymbol{w_t}$ satisfies that $f_i(\boldsymbol{w_t}) < f_j(\boldsymbol{w_t})$ for $i \in \mathcal{G}, j \in \mathbb{O}$. For $|\mathcal{G}| \leq \frac{n}{2}$, we have for

$$\begin{aligned}
R_t^{(MKL_2)} =& -2\eta_t(1-\eta_t \sup_i L_i)\sum_{i=1}^{|\mathcal{G}|}\frac{(n-2i+1)}{n(n-1)}\langle \boldsymbol{w_t}-\boldsymbol{w^*}, \nabla f_{m_i}(\boldsymbol{w_t}) - \nabla f_{m_i}(\boldsymbol{w^*})\rangle \\
&+ 2\eta_t \sum_{i=1}^{|\mathcal{G}|}\frac{2(n-i)}{n(n-1)}\left(\langle \boldsymbol{w^*}-\boldsymbol{w_t}, \nabla f_i(\boldsymbol{w^*})\rangle + \eta_t\|\nabla f_i(\boldsymbol{w^*})\|^2\right) \\
&+ \eta_t^2 \sum_{i=|\mathcal{G}|+1}^{n}\frac{2(n-i)}{n(n-1)}\|\nabla f_i(\boldsymbol{w_t})\|^2 + 2\eta_t \sum_{i=|\mathcal{G}|+1}^{n}\frac{2(n-i)}{n(n-1)}\left(f_i(\boldsymbol{w^*}) - f_i(\boldsymbol{w_t})\right) \\
\leq& -2\eta_t(1-\eta_t \sup_i L_i)\frac{|\mathcal{G}|\lambda_{good}}{n(n-1)}\|\boldsymbol{w_t}-\boldsymbol{w^*}\|^2 \\
&+ \frac{4\eta_t}{n}\sum_{i=1}^{|\mathcal{G}|}\left(\|\boldsymbol{w^*}-\boldsymbol{w_t}\|\|\nabla f_i(\boldsymbol{w^*})\| + \eta_t\|\nabla f_i(\boldsymbol{w^*})\|^2\right) \\
&+ \sum_{i=|\mathcal{G}|+1}^{n}\frac{\eta_t^2}{n}\|\nabla f_i(\boldsymbol{w_t})\|^2 + \sum_{i=|\mathcal{G}|+1}^{n}\frac{2\eta_t}{n}\left(f_i(\boldsymbol{w^*}) - f_i(\boldsymbol{w_t})\right)
\end{aligned}$$

We will have $R_t^{(MKL_2)} \leq R_t^{(SGD)}$ if the following inequality holds

$$(1-\eta_t \sup_i L_i)\frac{|\mathcal{G}|\lambda_{good}}{(n-1)}\|\boldsymbol{w_t}-\boldsymbol{w^*}\|^2 \geq \sum_{i=1}^{|\mathcal{G}|}\left(2\|\boldsymbol{w^*}-\boldsymbol{w_t}\|\|\nabla f_i(\boldsymbol{w^*})\| + \eta_t\|\nabla f_i(\boldsymbol{w^*})\|^2\right). \tag{36}$$

Indeed, for the noise level $\|\nabla f_i(\boldsymbol{w^*})\|^2$ satisfying (32) we have for $i \in \mathcal{G}$,

$$(1-\eta_t \sup_i L_i)\frac{\lambda_{good}}{(n-1)}\|\boldsymbol{w_t}-\boldsymbol{w^*}\|^2 \geq 2\|\boldsymbol{w^*}-\boldsymbol{w_t}\|\|\nabla f_i(\boldsymbol{w^*})\| + \eta_t\|\nabla f_i(\boldsymbol{w^*})\|^2.$$

Summing up the terms in $i \in \mathcal{G}$, we get (36). For the noise level $\|\nabla f_i(\boldsymbol{w^*})\|^2$ satisfying (33) we have

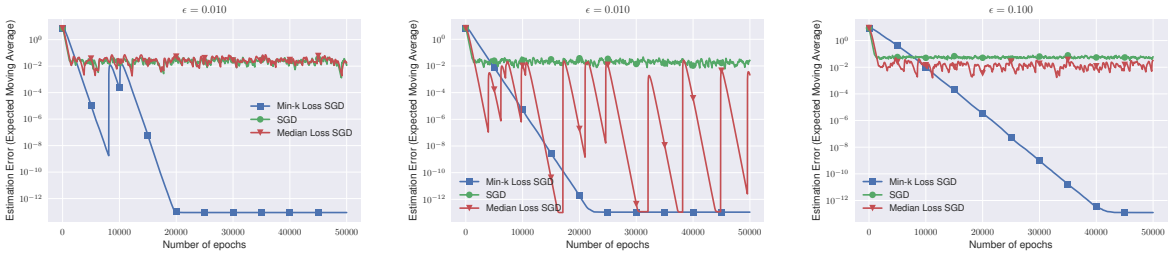
$$\begin{aligned}
(1-\eta_t \sup_i L_i)\frac{\lambda_{good}|\mathcal{G}|}{(n-1)}\|\boldsymbol{w_t}-\boldsymbol{w^*}\|^2 &\geq \left(2\|\boldsymbol{w^*}-\boldsymbol{w_t}\|\sqrt{n\sum_{i\in\mathcal{G}}\|\nabla f_i(\boldsymbol{w^*})\|^2} + \eta_t\sum_{i\in\mathcal{G}}\|\nabla f_i(\boldsymbol{w^*})\|^2\right) \\
&\geq 2\|\boldsymbol{w^*}-\boldsymbol{w_t}\|\sum_{i\in\mathcal{G}}\|\nabla f_i(\boldsymbol{w^*})\| + \eta_t\sum_{i\in\mathcal{G}}\|\nabla f_i(\boldsymbol{w^*})\|^2.
\end{aligned}$$

which results in (36). □

7.4 More experimental results

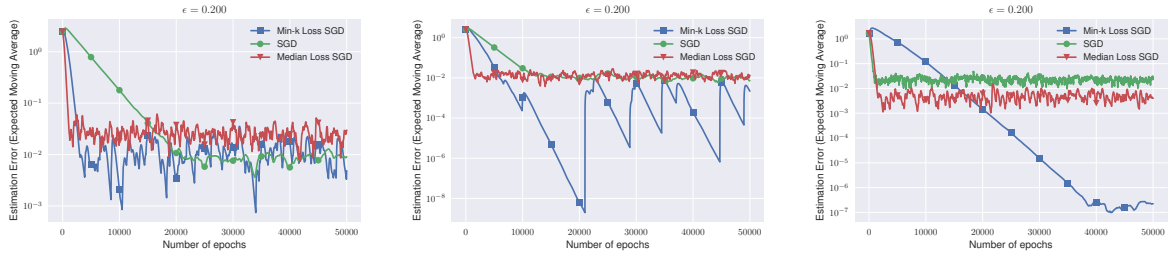
7.4.1 Linear Regression

Here, we show that there exists a trade-off for MKL-SGD between the rate of convergence and robustness the algorithm provides against outliers depending on the value of the parameter k. Larger the k, more robust is the algorithm, but slower is the rate of convergence. The algorithm outperforms median loss SGD and SGD. We also experimented with other order statistics and observed that for most general settings MKL-SGD was the best to pick. Note that the outliers are chosen from $\mathcal{N}(0,1)$ distribution independently of the data sample.



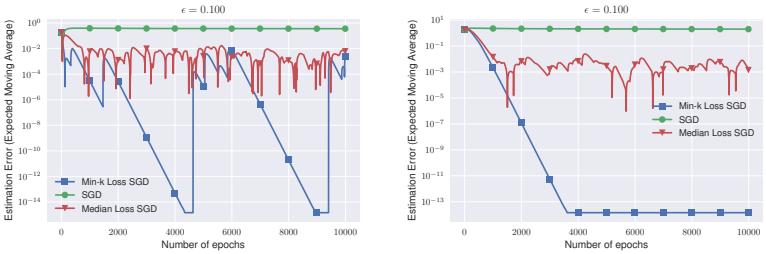
(a) k=2 (b) k=3 (c) k=5

Figure 6: Comparing the performance of MKL-SGD , SGD and Median loss SGD in the noiseless setting, $d = 50$.



(a) k=3 (b) k=5 (c) k=9

Figure 7: Comparing the performance of MKL-SGD , SGD and Median loss SGD in the noisy setting, $d = 10$, Noise variance=0.0001



(a) k=3 (b) k=5

Figure 8: Comparing the performance of MKL-SGD , SGD and Median loss SGD in the noiseless setting, $d = 25$, Noise variance=0.01

7.4.2 Neural Network Experiments

Here, we show that in presence of outliers instead of tuning other hyperparameters like learning rate, tuning over k might lead to significant gains in performances for deep neural networks. To illustrate this we play around with two commonly used noise models: random noise and directed noise. In the random noise model, the outlier label is randomly assigned while for the directed noise model for some class 'a', the outlier is assigned the same label 'b', similarly all the outliers for class 'b' are assigned label 'c' and so on.

Dataset	**MNIST with 2-layer CNN (Directed Noise)**						
Optimizer	**SGD**	**MKL-SGD**					**Oracle**
ϵ \ α	1.0	0.9	0.8	0.7	0.6	0.5	1.0
0.1	96.76	97.23	95.89	**97.47**	96.34	94.54	98.52
0.2	92.54	95.81	95.58	**97.46**	97.03	95.76	98.33
0.3	85.77	91.56	93.59	95.30	**96.54**	95.96	98.16
0.4	71.95	78.68	82.25	85.93	91.29	**94.20**	97.98

Table 2: In this experiments, we train a standard 2 layer CNN on subsampled MNIST (5000 training samples with labels corrupted using random label noise). We train over 80 epochs using an initial learning rate of 0.05 with the decaying schedule of factor 5 after every 30 epochs. The reported accuracy is based on the true validation set. The results of the MNIST dataset are reported as the mean of 5 runs. For the MKL-SGD algorithm, we introduce a more practical variant that evaluates k sample losses and picks a batch of size αk where $k = 10$.

Dataset	**MNIST with 2-layer CNN (Random Noise)**						
Optimizer	**SGD**	**MKL-SGD**					**Oracle**
ϵ \ α	1.0	0.9	0.8	0.7	0.6	0.5	1.0
0.1	96.91	97.9	**98.06**	97.59	96.49	94.43	98.44
0.2	93.94	95.5	96.16	97.02	**97.04**	96.25	98.18
0.3	87.14	90.71	91.60	92.97	94.54	**95.36**	97.8
0.4	71.83	74.31	76.6	78.30	77.58	**80.86**	97.16

Table 3: In this experiments, we train a standard 2 layer CNN on subsampled MNIST (5000 training samples with labels corrupted using random label noise). We train over 80 epochs using an initial learning rate of 0.05 with the decaying schedule of factor 5 after every 30 epochs. The reported accuracy is based on the true validation set. The results of the MNIST dataset are reported as the mean of 5 runs. For the MKL-SGD algorithm, we introduce a more practical variant that evaluates k sample losses and picks a batch of size αk where $k = 10$.

Dataset	**CIFAR-10 with Resnet-18 (Directed Noise)**						
Optimizer	**SGD**	**MKL-SGD**					**Oracle**
ϵ \ α	1.0	0.9	0.8	0.7	0.6	0.5	1.0
0.1	79.1	77.52	79.57	81.00	**81.94**	80.53	84.56
0.2	72.29	69.58	70.17	72.76	77.77	**78.93**	84.40
0.3	63.96	61.43	60.46	61.58	66.49	**69.57**	84.66
0.4	**52.4**	51.53	51.04	51.07	53.57	51.2	84.42

Table 4: In this experiments, we train Resnet 18 on CIFAR-10 (50000 training samples with labels corrupted using directed label noise). We train over 200 epochs using an initial learning rate of 0.05 with the decaying schedule of factor 5 after every 90 epochs. The reported accuracy is based on the true validation set. The results of the CIFAR-10 dataset are reported as the mean of 3 runs. For the MKL-SGD algorithm, we introduce a more practical variant that evaluates k sample losses and picks a batch of size αk where $k = 16$.

7.5 Conclusions and Future Work

7.5.1 Conclusions

In this paper, we propose MKL-SGD that is computationally inexpensive, has linear convergence (upto a certain neighborhood) and is robust against outliers. We analyze MKL-SGD algorithm under noiseless and noisy settings with and without outliers. MKL-SGD outperforms SGD in terms of generalization for both

linear regression and neural network experiments. MKL-SGD opens up a plethora of challenging questions with respect to understanding convex optimization in a non-convex landscape which will be discussed in the Appendix.

7.5.2 Future Work

To ensure consistency, i.e. $\|\bar{\boldsymbol{w}}_{\boldsymbol{MKL}} - \boldsymbol{w}^*\| \to 0$, we require that $k \geq n\epsilon + 1$. In all other cases, there will be a non-zero contribution from the outliers which keeps the MKL-SGD solution from exactly converging to $\boldsymbol{w}^*$. In this paper, we consider unknown ϵ and thus k should be a hyperparameter. For neural network experiments in the Appendix, we show that tuning k as a hyperparameter can lead to significant improvements in performance in presence of outliers.

The obvious question is if it is possible to provide worst case guarantees for a larger subset of problems using smarter initialization techniques. It will be interesting to analyze the tradeoff between better generalization guarantees offered by large k and rates of convergence. The worst case analysis in the noisy setting for standard convex optimization losses remains an open problem. As we show in the previous set of experiments, in presence of noise, tuning the hyperparameter k can provide significant boosts to the performance.