Abstract. This text is presented to provide a lower bound of logistic regression loss by $\rho_t^{\perp}$, which can support our claim that increasing $\rho_t^{\perp}$ can lead to increasing loss.

To lower bound the loss by $\rho_t^{\perp}$, we need introduce the new techniques, which is originally from [1]. Here is the scaffolding assumption.

Assumption 0.1 (Non-degenerate data)**.** Let ℓ be ℓ_{log} and $\hat{\mathbf{w}}$ be the SVM solution as presented in Definition 5. And let $\mathcal{S}$ is the support vector set of the dataset. Assume that there exists $\alpha_i > 0, \forall i \in \mathcal{S}$ such that $\hat{\mathbf{w}} = \sum_{i\in\mathcal{S}} \alpha_i y_i \mathbf{x}_i$.

Remark 0.2. It is worth to mention that this assumption is mild. It holds almost surely for all linearly separable dataset that is sampled from continuous distributions (refer to Appendix B of [2]).

Here is the scaffolding lemma.

Lemma 0.3 (Margin Offset)**.** *Suppose that Assumption 0.1 and Assumption 4 hold. There exists the margin offset $b > 0$ such that*

$$-b := \max_{\mathbf{w}\in span^{\perp}\{\hat{\mathbf{w}}\}\cap span\{\mathbf{x}_1,\cdots,\mathbf{x}_n\}} \min_{i\in[n]} y_i \cdot \langle \mathbf{x}_i, \mathbf{w}\rangle / \|\mathbf{w}\|$$

We mention that Definition 5, Assumption 4 and Eq. (1) are in the original paper.

Remark 0.4. This immediately implies that: for any $\mathbf{w} \in \operatorname{span}\{\mathbf{x}_1,\cdots,\mathbf{x}_n\}$ such that $\mathbf{w}^T\hat{\mathbf{w}} = 0$, there exist $i \in [n]$ such that $y_i \cdot \langle \mathbf{x}_i, \mathbf{w}\rangle \le -b \cdot \|\mathbf{w}\|$.

Proof. Assumption 4 assumes the data overparameterazation setting, which suggests that all sample $\mathbf{x}_i$ in the dataset is support vector. Therefore we have $\operatorname{span}\{\mathbf{x}_1,\cdots,\mathbf{x}_n\} = \operatorname{span}\{\mathbf{x}_i|\mathbf{x}_i \in \mathcal{S}\}$

If Assumption 0.1 holds, we can prove for any $\mathbf{v} \in \operatorname{span}\{\mathbf{x}_1,\cdots,\mathbf{x}_n\}$ such that $\langle \mathbf{v}, \hat{\mathbf{w}}\rangle = 0$, there exist $i, j \in S$, such that $y_i \cdot \langle \mathbf{x}_i, \mathbf{v}\rangle < 0$, $y_j \cdot \langle \mathbf{x}_j, \mathbf{v}\rangle > 0$. To see this, we have

$$0 = \langle \mathbf{v}, \hat{\mathbf{w}}\rangle = \sum_{i\in\mathcal{S}} \alpha_i y_i \langle \mathbf{v}, \mathbf{x}_i\rangle$$

Because $\mathbf{v} \in \operatorname{span}\{\mathbf{x}_1,\cdots,\mathbf{x}_n\} = \operatorname{span}\{\mathbf{x}_i|\mathbf{x}_i \in \mathcal{S}\}$, there must exist $i, j \in [n]$, such that $y_i \cdot \langle \mathbf{x}_i, \mathbf{v}\rangle < 0$, $y_j \cdot \langle \mathbf{x}_j, \mathbf{v}\rangle > 0$. Therefore, the constant b as described in this lemma always exists. □

This lemma introduces an important constant b that helps us lower bound loss value.

Lemma 0.5 (Lower Bound of Logistic Loss)**.** *Let ℓ be ℓ_{log} and $\hat{\mathbf{w}}$ be the SVM solution as presented in Definition 5. Suppose Assumption 4 and 0.1 holds. Consider the gradient descent (1), for all $t \ge 0$, if $\alpha_t > 0$, it holds that*

$$\mathcal{R}_t \ge \frac{1}{n}\ell\left(\frac{\alpha_t}{\sqrt{\lambda_{\min}}}\left(\rho_t\gamma^2 - \rho_t^{\perp} b\gamma\right)\right)$$

Proof. Consider $\mathcal{R}_t$, we have

$$\mathcal{R}_t = \frac{1}{n}\sum_{i=1}^{n} \ell\left(\alpha_t \frac{\langle \mathbf{w}_t, \mathbf{x}_i y_i\rangle}{\|\mathbf{w}_t\|_{\mathbf{\Sigma}}}\right)$$

By Lemma 0.3, there exists a $j \in [n]$ such that

$$\begin{aligned}
\left\langle y_j\mathbf{x}_j, \left(\mathbf{I} - \frac{\hat{\mathbf{w}}\hat{\mathbf{w}}^T}{\|\hat{\mathbf{w}}\|^2}\right)\mathbf{w}_t\right\rangle &\le -b\left\|\left(\mathbf{I} - \frac{\hat{\mathbf{w}}\hat{\mathbf{w}}^T}{\|\hat{\mathbf{w}}\|^2}\right)\mathbf{w}_t\right\| \\
&= -b\frac{\left\|\left(\mathbf{I} - \frac{\hat{\mathbf{w}}\hat{\mathbf{w}}^T}{\|\hat{\mathbf{w}}\|^2}\right)\mathbf{w}_t\right\|}{\left\|\left(\mathbf{I} - \frac{\mathbf{w}_t{\mathbf{w}_t}^T}{\|\mathbf{w}_t\|^2}\right)\hat{\mathbf{w}}\right\|}\left\|\left(\mathbf{I} - \frac{\mathbf{w}_t{\mathbf{w}_t}^T}{\|\mathbf{w}_t\|^2}\right)\hat{\mathbf{w}}\right\| \\
&= -b\frac{\|\mathbf{w}_t\|}{\|\hat{\mathbf{w}}\|}\rho_t^{\perp} = -\rho_t^{\perp} b\gamma\|\mathbf{w}_t\|
\end{aligned}$$

Then, we have

$$
\begin{aligned}
\langle \mathbf{w}_t, \mathbf{x}_i y_i \rangle &= \left\langle \left(\frac{\hat{\mathbf{w}}\hat{\mathbf{w}}^T}{\|\hat{\mathbf{w}}\|^2} \right) \mathbf{w}_t, \mathbf{x}_i y_i \right\rangle + \left\langle \left(\mathbf{I} - \frac{\hat{\mathbf{w}}\hat{\mathbf{w}}^T}{\|\hat{\mathbf{w}}\|^2} \right) \mathbf{w}_t, \mathbf{x}_i y_i \right\rangle \\
&\leq \gamma^2 \mathbf{w}_t^T \hat{\mathbf{w}} - \rho_t^{\perp} b \gamma \|\mathbf{w}_t\| \\
&= \gamma^2 \|\mathbf{w}_t\| \rho_t - \rho_t^{\perp} b \gamma \|\mathbf{w}_t\|
\end{aligned}
$$

Then

$$
\begin{aligned}
\mathcal{R}_t \geq & \frac{1}{n} \ell \left(\alpha_t \frac{\langle \mathbf{w}_t, \mathbf{x}_j y_j \rangle}{\|\mathbf{w}_t\|_{\mathbf{\Sigma}}} \right) \\
\geq & \frac{1}{n} \ell \left(\alpha_t \frac{\|\mathbf{w}_t\|}{\|\mathbf{w}_t\|_{\mathbf{\Sigma}}} \left(\rho_t \gamma^2 - \rho_t^{\perp} b \gamma \right) \right) \\
\geq & \frac{1}{n} \ell \left(\frac{\alpha_t}{\sqrt{\lambda_{\min}}} \left(\rho_t \gamma^2 - \rho_t^{\perp} b \gamma \right) \right)
\end{aligned}
$$

□

Appendix

[1] Wu J, Braverman V, Lee J D. Implicit bias of gradient descent for logistic regression at the edge of stability[J]. Advances in Neural Information Processing Systems, 2023, 36: 74229-74256.

[2] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. The Journal of Machine Learning Research, 19(1): 2822–2878, 2018.