Appendix

Theorem 6.1. *Given a set of n training data instances lying in a ball of radius 1, the sensitivity of regularized logistic regression classifier is at most $\frac{2}{n\lambda}$. If $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ are classifiers trained on adjacent datasets of size n with regularization parameter λ,*

$$\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_1 \leq \frac{2}{n\lambda}.$$

Lemma 6.2. *Let $G(\boldsymbol{w})$ and $g(\boldsymbol{w})$ be two differentiable, convex functions of $\boldsymbol{w}$. If $\boldsymbol{w}_1 = \arg\min_{\boldsymbol{w}} G(\boldsymbol{w})$ and $\boldsymbol{w}_2 = \arg\min_{\boldsymbol{w}} G(\boldsymbol{w}) + g(\boldsymbol{w})$, then $\|\boldsymbol{w}_1 - \boldsymbol{w}_2\| \leq \frac{g_1}{G_2}$ where $g_1 = \max_{\boldsymbol{w}} \|\nabla g(\boldsymbol{w})\|$ and $G_2 = \min_{\boldsymbol{v}} \min_{\boldsymbol{w}} \boldsymbol{v}^T \nabla^2 G(\boldsymbol{w}) \boldsymbol{v}$ for any unit vector $\boldsymbol{v} \in \mathbb{R}^d$.*

Proof of Theorem 4.2. We formulate the problem of estimating the individual classifiers $\hat{\mathbf{w}}_j$ and the classifier $\mathbf{w}^*$ trained over the entire training data in terms of minimizing the two differentiable and convex functions $g(\mathbf{w})$ and $G(\mathbf{w})$.

$$\begin{aligned}
\hat{\mathbf{w}}_j &= \operatorname*{argmin}_w J(\mathbf{w}, \mathbf{x}|_j, \mathbf{y}|_j) = \operatorname*{argmin}_w G(\mathbf{w}), \\
\mathbf{w}^* &= \operatorname*{argmin}_w J(\mathbf{w}, \mathbf{x}, \mathbf{y}) = \operatorname*{argmin}_w L(\mathbf{w}, \mathbf{x}|_j, \mathbf{y}|_j) + \sum_{l \in [K]-j} L(\mathbf{w}, \mathbf{x}|_l, \mathbf{y}|_l) + \lambda \|\mathbf{w}\|^2 \\
&= \operatorname*{argmin}_w J(\mathbf{w}, \mathbf{x}|_j, \mathbf{y}|_j) + \sum_{l \in [K]-j} L(\mathbf{w}, \mathbf{x}|_l, \mathbf{y}|_l) = \operatorname*{argmin}_w G(\mathbf{w}) + g(\mathbf{w}).
\end{aligned}$$

$$\begin{aligned}
\nabla g(\mathbf{w}) &= \sum_{l \in [K]-j} \frac{1}{n_l} \sum_{i=1}^{n_l} \frac{-e^{-y_i|_l w^T x_i|_l}}{1 + e^{-y_i|_l w^T x_i|_l}} y_i x_i^T, \\
\|\nabla g(\mathbf{w})\| &= \left\| \sum_l \frac{1}{n_l} \sum_i \frac{-e^{-y_i|_l w^T x_i|_l}}{1 + e^{-y_i|_l w^T x_i|_l}} y_i x_i^T \right\| \leq \sum_l \frac{1}{n_l} \left\| \sum_i \frac{-e^{-y_i|_l w^T x_i|_l}}{1 + e^{-y_i|_l w^T x_i|_l}} y_i x_i^T \right\| \leq \sum_l \frac{1}{n_l}, \\
g_1 &= \max_{\mathbf{w}} \|\nabla g(\mathbf{w})\| \leq \sum_l \frac{1}{n_l}. \qquad (5)
\end{aligned}$$

$$\nabla^2 G(\mathbf{w}) = \frac{1}{n_j} \sum_i \frac{e^{y_i|_j \mathbf{w}^T \mathbf{x}_i|_j}}{1 + e^{y_i|_j \mathbf{w}^T \mathbf{x}_i|_j}} \mathbf{x}_i|_j \mathbf{x}_i|_j^T + \lambda \mathbf{1}, \;\; G_2 \geq \lambda. \qquad (6)$$

Substituting the bounds on g_1 and G_2 in Lemma 6.2,

$$\|\hat{\mathbf{w}}_j - \mathbf{w}^*\| \leq \frac{1}{\lambda} \sum_l \frac{1}{n_l}. \qquad (7)$$

Applying triangle inequality,

$$\begin{aligned}
\|\hat{\mathbf{w}} - \mathbf{w}^*\| &= \left\| \frac{1}{K} \sum_j \hat{\mathbf{w}}_j - \mathbf{w}^* \right\| = \frac{1}{K} \left\| \sum_j \hat{\mathbf{w}}_j - K\mathbf{w}^* \right\| = \frac{1}{K} \|(\hat{\mathbf{w}}_1 - \mathbf{w}^*) + \ldots + (\hat{\mathbf{w}}_K - \mathbf{w}^*)\| \\
&\leq \frac{1}{K} \sum_j \|\hat{\mathbf{w}}_j - \mathbf{w}^*\| = \frac{1}{K\lambda} \sum_j \sum_{l \in [K]-j} \frac{1}{n_l} = \frac{K-1}{K\lambda} \sum_j \frac{1}{n_j} \leq \frac{K-1}{n_{(1)}\lambda}.
\end{aligned}$$

where $n_{(1)} = \min_j n_j$. □

Proof of Theorem 4.3. We use the Taylor series expansion of the function J to have

$$J(\hat{\mathbf{w}}^s) = J(\mathbf{w}^*) + (\hat{\mathbf{w}}^s - \mathbf{w}^*)^T \nabla J(\mathbf{w}^*) + \frac{1}{2} (\hat{\mathbf{w}}^s - \mathbf{w}^*)^T \nabla^2 J(\mathbf{w}) (\hat{\mathbf{w}}^s - \mathbf{w}^*)$$

for some $\mathbf{w} \in \mathbb{R}^d$. By definition, $\nabla J(\mathbf{w}^*) = 0$.

Taking ℓ_2 norm on both sides and applying Cauchy-Schwarz inequality,

$$|J(\hat{\mathbf{w}}^s) - J(\mathbf{w}^*)| \leq \frac{1}{2}\|\hat{\mathbf{w}}^s - \mathbf{w}^*\|^2 \, \|\nabla^2 J(\mathbf{w})\|. \tag{8}$$

The second gradient of the regularized loss function for logistic regression is

$$\nabla^2 J(\mathbf{w}) = \frac{1}{n}\sum_i \frac{e^{y_i \mathbf{w}^T \mathbf{x}_i}}{1 + e^{y_i \mathbf{w}^T \mathbf{x}_i}} \mathbf{x}_i \mathbf{x}_i^T + \lambda \mathbf{1} = \frac{1}{n}\sum_i \frac{1}{1 + e^{-y_i \mathbf{w}^T \mathbf{x}_i}} \mathbf{x}_i \mathbf{x}_i^T + \lambda \mathbf{1}.$$

Since the logistic function term is always less than one and all $\mathbf{x}_i$ lie in a unit ball, $\|\nabla^2 J(\mathbf{w})\| \leq \lambda + 1$. Substituting this into Equation 8 and using the fact that $J(\mathbf{w}^*) \leq J(\mathbf{w}), \forall \mathbf{w} \in \mathbb{R}^d$,

$$J(\hat{\mathbf{w}}^s) \leq J(\mathbf{w}^*) + \frac{\lambda + 1}{2}\|\hat{\mathbf{w}}^s - \mathbf{w}^*\|^2. \tag{9}$$

The classifier $\hat{\mathbf{w}}^s$ is the perturbed aggregate classifier, *i.e.*, $\hat{\mathbf{w}}^s = \hat{\mathbf{w}} + \boldsymbol{\eta}$, with the noise term $\boldsymbol{\eta} \sim \text{Lap}\left(\frac{2}{n_{(1)}\epsilon\lambda}\right)$. We apply Lemma 6.3 to bound $\|\boldsymbol{\eta}\|$ with probability at least $1 - \delta$. Substituting this into Equation 9, we have

$$\begin{aligned} J(\hat{\mathbf{w}}^s) &\leq J(\mathbf{w}^*) + \frac{1}{2}\|\hat{\mathbf{w}} - \mathbf{w}^* + \boldsymbol{\eta}\|^2 = J(\mathbf{w}^*) + \frac{\lambda + 1}{2}\left[\|\hat{\mathbf{w}} - \mathbf{w}^*\|^2 + \|\boldsymbol{\eta}\|^2 + 2(\hat{\mathbf{w}} - \mathbf{w}^*)^T \boldsymbol{\eta}\right] \\ &\leq J(\mathbf{w}^*) + \frac{\lambda + 1}{2}\left[\frac{(K-1)^2}{n_{(1)}^2 \lambda^2} + \frac{4d^2}{n_{(1)}^2 \epsilon^2 \lambda^2} \log^2\left(\frac{d}{\delta}\right)\right] + (\lambda + 1)|(\hat{\mathbf{w}} - \mathbf{w}^*)^T \boldsymbol{\eta}|. \end{aligned}$$

Using the Cauchy-Schwarz inequality on the last term,

$$\begin{aligned} J(\hat{\mathbf{w}}^s) &\leq J(\mathbf{w}^*) + \frac{\lambda + 1}{2}\left[\frac{(K-1)^2}{n_{(1)}^2 \lambda^2} + \frac{4d^2}{n_{(1)}^2 \epsilon^2 \lambda^2} \log^2\left(\frac{d}{\delta}\right)\right] + (\lambda + 1)\|\hat{\mathbf{w}} - \mathbf{w}^*\|\|\boldsymbol{\eta}\| \\ &\leq J(\mathbf{w}^*) + \frac{(K-1)^2(\lambda + 1)}{2n_{(1)}^2 \lambda^2} + \frac{2d^2(\lambda + 1)}{n_{(1)}^2 \epsilon^2 \lambda^2} \log^2\left(\frac{d}{\delta}\right) + \frac{2d(K-1)(\lambda + 1)}{n_{(1)}^2 \epsilon \lambda^2} \log\left(\frac{d}{\delta}\right). \end{aligned}$$

□

Proof of Theorem 4.4. Let $\mathbf{w}^r$ be the classifier minimizing the true risk $\tilde{J}(\mathbf{w})$. By rearranging the terms,

$$\tilde{J}(\hat{\mathbf{w}}^s) = \tilde{J}(\mathbf{w}^*) + [\tilde{J}(\hat{\mathbf{w}}^s) - \tilde{J}(\mathbf{w}^r)] + [\tilde{J}(\mathbf{w}^r) - \tilde{J}(\mathbf{w}^*)] \leq \tilde{J}(\mathbf{w}^*) + [\tilde{J}(\hat{\mathbf{w}}^s) - \tilde{J}(\mathbf{w}^r)].$$

Sridharan, et al. [14] present a bound between the true excess risk of any classifier as an expression of bound on the regularized empirical risk for that classifier and the classifier minimizing the regularized empirical risk. With probability at least $1 - \delta$,

$$\tilde{J}(\hat{\mathbf{w}}^s) - \tilde{J}(\mathbf{w}^r) \leq 2[J(\hat{\mathbf{w}}^s) - J(\mathbf{w}^*)] + \frac{16}{\lambda n}\left[32 + \log\left(\frac{1}{\delta}\right)\right]. \tag{10}$$

Substituting the bound from Theorem 4.3,

$$\begin{aligned} \tilde{J}(\hat{\mathbf{w}}^s) - \tilde{J}(\mathbf{w}^r) \leq{}& \frac{2(K-1)^2(\lambda + 1)}{2n_{(1)}^2 \lambda^2} + \frac{4d^2(\lambda + 1)}{n_{(1)}^2 \epsilon^2 \lambda^2} \log^2\left(\frac{d}{\delta}\right) && (11) \\ &+ \frac{4d(K-1)(\lambda + 1)}{n_{(1)}^2 \epsilon \lambda^2} \log\left(\frac{d}{\delta}\right) + \frac{16}{\lambda n}\left[32 + \log\left(\frac{1}{\delta}\right)\right]. && (12) \end{aligned}$$

Substituting this bound into Equation 10 gives us a bound on the true excess risk of the classifier $\hat{\mathbf{w}}^s$ over the classifier $\mathbf{w}^*$. □

Lemma 6.3. *Given a d-dimensional random variable $\boldsymbol{\eta} \sim$ Lap(β)* i.e., *$P(\boldsymbol{\eta}) = \frac{1}{2\beta} e^{-\frac{\|\boldsymbol{\eta}\|_1}{\beta}}$, with probability at least $1 - \delta$, the ℓ_2 norm of the random variable is bounded as*

$$\|\boldsymbol{\eta}\| \leq 2d\beta \log\left(\frac{d}{\delta}\right).$$

The proof is similar to Lemma 5 of [8].