## A. Proof of Lemma 3.5

*Proof.* Given a differentiable function $F:\mathcal{F}\times\mathcal{F}\to\mathbb{R}$, we define the Bregman divergence by

$$d_F(f,f') = F(f) - F(f') - \langle f-f', \nabla F(f')\rangle,\ \forall f,f'\in\mathcal{F}.$$

Define $R:\mathcal{F}\to\mathbb{R}$ by

$$R(f) := \frac{1}{N}\sum_{i\in[N]} L(f,s_i) + \lambda\|f\|_k^2,\ \forall f\in\mathcal{F}.$$

Also define $R^i:\mathcal{F}\to\mathbb{R}$ by

$$R^i(f) := \frac{1}{N}\left(\sum_{j\neq i} L(f,s_j) + L(f,s_i')\right) + \lambda\|f\|_k^2,\ \forall f\in\mathcal{F}.$$

By definition of $g$ and $g^i$, we have

$$\begin{aligned} d_R(g^i,g) =& R(g^i) - R(g) - \langle g^i - g, \nabla R(g)\rangle \\ \leq& R(g^i) - R(g), \end{aligned} \tag{5}$$

and

$$\begin{aligned} d_{R^i}(g,g^i) =& R^i(g) - R^i(g^i) - \langle g-g^i, \nabla R^i(g^i)\rangle \\ \leq& R^i(g) - R^i(g^i). \end{aligned} \tag{6}$$

By Inequalities (5) and (6), we have

$$\begin{aligned} & d_R(g^i,g) + d_{R^i}(g,g^i) \\ \leq& R(g^i) - R(g) + R^i(g) - R^i(g^i) \\ =& \frac{1}{N}\left(L(g^i,s_i) - L(g,s_i) + L(g,s_i') - L(g^i,s_i')\right). \end{aligned} \tag{7}$$

Since $d_{A+B} = d_A + d_B$, we have

$$\begin{aligned} & 2\lambda\|g-g^i\|_k^2 \\ =& \lambda d_{\|\cdot\|_k^2}(g,g^i) + \lambda d_{\|\cdot\|_k^2}(g^i,g) \\ & (\text{Defn. of } \|\cdot\|_k^2) \\ =& d_{R^i}(g,g^i) - d_{\sum_{j\neq i} L(\cdot,s_j)}(g,g^i) \\ & + d_R(g^i,g) - d_{\sum_{i\in[N]} L(\cdot,s_i)}(g^i,g) \\ & (d_{A+B} = d_A + d_B) \\ \leq& d_{R^i}(g,g^i) + d_R(g^i,g) \\ & (\text{nonnegativity of } d_F) \\ \leq& \frac{1}{N}\left(L(g^i,s_i) - L(g,s_i) + L(g,s_i') - L(g^i,s_i')\right) \\ \leq& \frac{\sigma}{N}\left(|g(x_i) - g^i(x_i)| + |g(x_i') - g^i(x_i')|\right). \\ & (L(\cdot,\cdot) \text{ is } \sigma\text{-admissible}) \end{aligned} \tag{8}$$

This completes the proof. □

## B. Proof of Theorem 3.7

*Proof.* By Inequality (8) in the proof of Lemma 3.5, we have

$$\begin{aligned} & 2\lambda\|v-v^i\|_2^2 \\ \leq& \frac{1}{N}\left(L(g^i,s_i) - L(g,s_i) + L(g,s_i') - L(g^i,s_i')\right). \end{aligned} \tag{9}$$

Moreover, we have for any $f = \alpha\cdot\phi(\cdot), f' = \alpha'\cdot\phi(\cdot)\in\mathcal{F}$ and $s\in\mathcal{D}$,

$$\begin{aligned} L(f,s) - L(f',s) \leq& \langle\nabla_\alpha L(f,s), \alpha-\alpha'\rangle \\ & (\text{Convexity of } L(\cdot,s)) \\ \leq& \|\nabla_\alpha L(\alpha,s)\|_2\cdot\|\alpha-\alpha'\|_2 \\ \leq& G\|\alpha-\alpha'\|_2 \\ & (\text{Defn. of } G). \end{aligned} \tag{10}$$

Combining with Inequalities (9) and (6), we have

$$\begin{aligned} & \|v-v^i\|_2^2 \\ \leq& \frac{1}{2\lambda N}\left(L(g^i,s_i) - L(g,s_i) + L(g,s_i') - L(g^i,s_i')\right) \\ & (\text{Ineq. (9)}) \\ \leq& \frac{1}{2\lambda N}\left(G\|v-v^i\|_2 + G\|v-v^i\|_2\right) \\ & (\text{Ineq. (10)}) \\ =& \frac{G}{\lambda N}\|v-v^i\|_2. \end{aligned}$$

It implies that $\|v-v^i\|_2 \leq \frac{G}{\lambda N}$. Combining with Inequality (6), we have for any $s\in\mathcal{D}$,

$$L(g,s) - L(g^i,s) \leq G\|v-v^i\|_2 \leq \frac{G^2}{\lambda N}.$$

This completes the proof for the stability guarantee. For the sacrifice in the empirical risk, the argument is the same as that of Theorem 3.2. □

## C. Details of Remark 3.3

- Prediction error: $f(x)\in\{-1,1\}$ for any pair $(f,x)$ and $L(f(x),y) = \mathrm{I}[f(x)\neq y]$,[5] then we have that

$$\begin{aligned} & |L(f(x),y) - L(f(x'),y)| \\ =& |\mathrm{I}[f(x)\neq y] - \mathrm{I}[f(x')\neq y]| \\ =& \mathrm{I}[f(x)\neq f(x')] = \frac{1}{2}|f(x) - f(x')|, \end{aligned}$$

  which is $\frac{1}{2}$-admissible.

[5] Here, $\mathrm{I}[\cdot]$ is the indicator function.

- Soft margin SVM: $L(f,s) = (1-yf(x))_+$,[6] then we have that

$$\begin{aligned} & |L(f(x),y) - L(f(x'),y)| \\ = & |(1-yf(x))_+ - (1-yf(x'))_+| \\ \leq & |yf(x) - yf(x')| \\ = & |f(x) - f(x')|, \end{aligned}$$

which is 1-admissible.

- Least Squares regression: $L(f,s) = (f(x)-y)^2$. Suppose $f(x) \in [-1,1]$ for any $x \in \mathcal{X}$, then we have that

$$\begin{aligned} & |L(f(x),y) - L(f(x'),y)| \\ = & \left|(f(x)-y)^2 - (f(x')-y)^2\right| \\ = & |(f(x)+f(x')-2y)(f(x)-f(x'))| \\ \leq & 4\,|f(x)-f(x')|, \end{aligned}$$

which is 4-admissible.

- Logistic regression: $L(f,s) = \ln(1+e^{-yf(x)})$. Note that we have for any $x \in \mathcal{X}$ and $y \in \{-1,1\}$,

$$\begin{aligned} & \left|\nabla_{f(x)} \ln(1+e^{-yf(x)})\right| \\ = & \left|\frac{-ye^{-yf(x)}}{1+e^{-yf(x)}}\right| = \left|\frac{e^{-yf(x)}}{1+e^{-yf(x)}}\right| \leq 1. \end{aligned}$$

Hence, the loss function $L(f,s) = \ln(1+e^{-yf(x)})$ is 1-admissible.

# D. Analysis of Our Framework in Specified Settings

Next, we show the stability guarantee of our framework in several specified models. We mainly analyze three commonly-used models: soft margin SVMs, least squares regression, and logistic regression.

**Soft margin SVMs.** Recall that $S = \{s_i = (x_i, z_i, y_i)\}_{i\in[N]}$ is the given training set. We first have a kernel function $k(\cdot,\cdot)$ that defines values $k(x_i,x_j)$. Then each classifier $f$ is a linear combination of $k(x_i,\cdot)$, i.e.,

$$f(\cdot) = \sum_{i\in[N]} \alpha_i k(x_i,\cdot)$$

for some $\alpha \in \mathbb{R}^N$ In the soft margin SVM model, we consider the following loss function

$$L(f,s) = (1-yf(x))_+$$

[6] $(a)_+ = a$ if $a \geq 0$ and otherwise $(a)_+ = 0$.

which is 1-admissible. Then Program (Stable-Fair) can be rewritten as follows.

$$\begin{aligned} & \min_{\alpha\in\mathbb{R}^N} \sum_{i\in[N]} \left(1 - y_i \sum_{j\in[N]} \alpha_j k(x_j,x_i)\right)_+ \\ & + \lambda \| \sum_{i,j\in[N]} \alpha_i\alpha_j k(x_i,x_j)\|_k^2 \quad s.t. \\ & \Omega(f) \leq 0. \end{aligned} \tag{SVM}$$

This model has been considered in (Zafar et al., 2017b;a) that aims to avoid disparate impact/disparate mistreatment. Applying Theorems 3.2 and 3.7, and the fact that $L(\cdot,\cdot)$ is 1-admissible (Remark 3.3), we directly have the following corollary.

**Corollary D.1.** *Suppose the learning algorithm $\mathcal{A}$ computes a minimizer $\mathcal{A}_S$ of Program* (SVM).

- *If $k(x_i,x_i) \leq \kappa^2 < \infty$ for each $i \in [N]$, then $\mathcal{A}$ is $\frac{\kappa^2}{\lambda N}$-uniformly stable.*
- *Let $G = \sup_{f=\alpha^\top\phi(\cdot)\in\mathcal{F}:\Omega(f)\leq 0} \sup_{s\in\mathcal{D}} \|\nabla_\alpha L(f,s)\|_2$. Then $\mathcal{A}$ is $\frac{G^2}{\lambda N}$-uniformly stable.*

**Least square regression.** The only difference from soft margin SVM is the loss function, which is defined as follows.

$$L(f,s) = (f(x)-y)^2.$$

Then Program (Stable-Fair) can be rewritten as follows.

$$\begin{aligned} & \min_{\alpha\in\mathbb{R}^N} \sum_{i\in[N]} \left(y_i - \sum_{j\in[N]} \alpha_j k(x_j,x_i)\right)^2 \\ & + \lambda \| \sum_{i,j\in[N]} \alpha_i\alpha_j k(x_i,x_j)\|_k^2 \quad s.t. \\ & \Omega(f) \leq 0. \end{aligned} \tag{LS}$$

Applying Theorems 3.2 and 3.7, we have the following corollary.

**Corollary D.2.** *Suppose the learning algorithm $\mathcal{A}$ computes a minimizer $\mathcal{A}_S$ of Program* (LS).

- *If $B = \max_{x\in\mathcal{X}} |f(x)|$ and $k(x_i,x_i) \leq \kappa^2 < \infty$ for each $i \in [N]$, then $\mathcal{A}$ is $\frac{(2B+2)^2\kappa^2}{\lambda N}$-uniformly stable.*
- *Let $G = \sup_{f=\alpha^\top\phi(\cdot)\in\mathcal{F}:\Omega(f)\leq 0} \sup_{s\in\mathcal{D}} \|\nabla_\alpha L(f,s)\|_2$. Then $\mathcal{A}$ is $\frac{G^2}{\lambda N}$-uniformly stable.*

*Proof.* We only need to verify that $L(\cdot,\cdot)$ is $(2B+2)$-

admissible. For any $x, x' \in \mathcal{X}$ and $y \in \{-1, 1\}$, we have

$$\begin{aligned}
&\left|(f(x)-y)^2-(f(x')-y)^2\right| \\
=&\left|(f(x)-f(x'))\cdot(f(x)+f(x')-2y)\right| \\
\leq&(|f(x)|+|f(x')|+2)\cdot|(f(x)-f(x'))| \\
\leq&(2B+2)\left|(f(x)-f(x'))\right|.
\end{aligned}$$

This completes the proof. $\square$

**Logistic regression.** Again, the only difference from soft margin SVM is the loss function, which is defined as follows.

$$L(f, s) = \ln(1 + e^{-yf(x)}).$$

This model has been widely used in the literature (Zafar et al., 2017b;a; Goel et al., 2018). Then Program (Stable-Fair) can be rewritten as follows.

$$\begin{aligned}
&\min_{\alpha \in \mathbb{R}^N} \sum_{i \in [N]} \ln\left(1 + y_i \cdot e^{-\sum_{j \in [N]} \alpha_j k(x_j, x_i)}\right) \\
&+ \lambda \| \sum_{i,j \in [N]} \alpha_i \alpha_j k(x_i, x_j) \|_k^2 \quad s.t. \\
&\Omega(f) \leq 0.
\end{aligned} \tag{LR}$$

Applying Theorem 3.2 and 3.7, and the fact that $L(\cdot, \cdot)$ is 1-admissible (Remark 3.3), we have the following corollary.

**Corollary D.3.** *Suppose the learning algorithm $\mathcal{A}$ computes a minimizer $\mathcal{A}_S$ of Program* (LR).

- *If $k(x_i, x_i) \leq \kappa^2 < \infty$ for each $i \in [N]$, then $\mathcal{A}$ is $\frac{\kappa^2}{\lambda N}$-uniformly stable.*
- *Let $G = \sup_{f=\alpha^\top \phi(\cdot) \in \mathcal{F}: \Omega(f) \leq 0} \sup_{s \in \mathcal{D}} \|\nabla_\alpha L(f, s)\|_2$. Then $\mathcal{A}$ is $\frac{G^2}{\lambda N}$-uniformly stable.*