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

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

Define \( R : \mathcal{F} \to \mathbb{R} \) by

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

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

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

By definition of \( g \) and \( g^i \), we have

\[
d_R(g^i, g) = R(g^i) - R(g) - \langle g^i - g, \nabla R(g) \rangle \leq R(g^i) - R(g),
\]

and

\[
d_R^i(g^i, g) = R^i(g) - R^i(g^i) - \langle g - g^i, \nabla R^i(g^i) \rangle \leq R^i(g) - R^i(g^i).
\]

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

\[
d_R(g^i, g) + d_R^i(g, g^i) \leq R(g^i) - R(g) + R^i(g) - R^i(g^i)
\]

\[
= \frac{1}{N} \left( L(g^i, s_i) - L(g, s_i) + L(g, s'_i) - L(g^i, s'_i) \right).
\]

Since \( d_{A+B} = d_A + d_B \), we have

\[
2\lambda \|g - g^i\|^2_k
= \lambda d_x \|g^i, g\|^2 + \lambda d_y \|g^i, g\|^2
= d_R(g^i) - d_R^i(g, g^i) + \sum_{j \neq i} L(s_j, s_i) (g^i - g)
+ d_R(g^i, g) + \sum_{i \in [N]} L(s_i, s_i) (g^i - g)
= d_R(g^i, g) + d_R(g^i, g)
\]

\[
\leq d_R(g^i, g) + d_R(g^i, g)
\]

\[
\leq \frac{1}{N} \left( L(g^i, s_i) - L(g, s_i) + L(g, s'_i) - L(g^i, s'_i) \right).
\]

This completes the proof.

B. Proof of Theorem 3.7

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

\[
2\lambda \|v - v^i\|^2_k
\leq \frac{1}{N} \left( L(g^i, s_i) - L(g, s_i) + L(g, s'_i) - L(g^i, s'_i) \right).
\]

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

\[
L(f, s) - L(f', s) \leq \langle \nabla \alpha L(f, s) - \alpha - \alpha', \phi(s) \rangle
\]

(Concavity of \( L(\cdot, s) \))

\[
\leq \| \nabla \alpha L(\alpha, s) \|_2 \cdot \| \alpha - \alpha' \|_2
\]

(Defn. of \( L \)).

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

\[
\|v - v^i\|^2_k
\leq \frac{1}{2\lambda N} \left( L(g^i, s_i) - L(g, s_i) + L(g, s'_i) - L(g^i, s'_i) \right)
\]

(Concavity of \( L(\cdot, s) \))

\[
\leq \| \nabla \alpha L(\alpha, s) \|_2 \cdot \| \alpha - \alpha' \|_2
\]

(Defn. of \( G \)).

It implies that \( \|v - v^i\|^2 \leq \frac{G^2}{\lambda N} \). Combining with Inequality (6), we have for any \( s \in \mathcal{D} \),

\[
L(g, s) - L(g^i, s) \leq G \|v - v^i\|^2 \leq G^2 \frac{1}{\lambda N}.
\]

This completes the proof of 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 \( (x, f) \) and \( L(f(x), y) = I[f(x) \neq y] \),\(^5\) then we have that

\[
\|L(f(x), y) - L(f(x'), y)\|
\]

\[
= I[f(x) \neq y] - I[f(x') \neq y]
\]

\[
= I[f(x) \neq f(x')] = \frac{1}{2} |f(x) - f(x')|,
\]

which is \( \frac{1}{2} \)-admissible.

\(^5\)Here, \( I[\cdot] \) is the indicator function.
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.,

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

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

\[
L(f, s) = (1 - yf(x))_{+}^6
\]

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

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

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 \( A \) 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 \( A \) is \( \frac{\kappa^2}{\lambda N} \)-uniformly stable.

- Let \( G = \sup_{f = \alpha \phi(\cdot) \in \mathcal{F} : \Omega(f) \leq 0} \sup_{s \in \mathcal{D}} \| \nabla^2_{\alpha} L(f, s) \|_2 \). Then \( A \) is \( \frac{G^2}{\kappa^2} \)-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.

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

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

**Corollary D.2.** Suppose the learning algorithm \( A \) 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 \( A \) is \( \frac{(2B + 2)^2 \kappa^2}{\lambda N} \)-uniformly stable.

- Let \( G = \sup_{f = \alpha \phi(\cdot) \in \mathcal{F} : \Omega(f) \leq 0} \sup_{s \in \mathcal{D}} \| \nabla^2_{\alpha} L(f, s) \|_2 \). Then \( A \) is \( \frac{G^2}{\kappa^2} \)-uniformly stable.

Proof. We only need to verify that \( L(\cdot, \cdot) \) is \( (2B + 2)\)-
Stable and Fair Classification

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

\[
\left| (f(x) - y)^2 - (f(x') - y)^2 \right| \\
= |(f(x) - f(x')) \cdot (f(x) + f(x') - 2y)| \\
\leq |f(x)| + |f(x')| + 2 \cdot |(f(x) - f(x'))| \\
\leq (2B + 2) |(f(x) - f(x'))|.
\]

This completes the proof.

**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.

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

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 \( A \) computes a minimizer \( A_S \) of Program (LR).

- If \( k(x_i, x_j) \leq \kappa^2 < \infty \) for each \( i \in [N] \), then \( A \) 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 \( A \) is \( \frac{G^2}{\lambda N} \)-uniformly stable.