International Journal of Machine Learning and Cybernetics manuscript No.
(will be inserted by the editor)

Supplementary Material of "Universal Consistency of Twin Support Vector Machines"

Weixia Xu · Dingjiang Huang · Shuigeng Zhou

the date of receipt and acceptance should be inserted later

1 Proofs of Lemmas and Theorems

1.1 Proof of Theorem 1

Proof By the definition of infimum, there exists $f_{1,\varepsilon} \in H$ such that

$$R^{reg}_{1,P,c_1}(f_{1,\varepsilon}) \leq \inf_{f_1 \in H} R^{reg}_{1,P,c_1}(f_1) + \varepsilon$$

for any $\varepsilon \in (0, L_1(1,0,\lambda_1)+L_1(-1,0,\lambda_1)]$. Note the value $L_1(1,0,\lambda_1)+L_1(-1,0,\lambda_1)$ is an upper bound with $f_1 = 0$ for the loss function L_1, since it measures the largest distance between y and $f_1(x)$. Thus,

$$\begin{aligned}\Omega(c_1, ||f_{1,\varepsilon}||_H) \leq R^{reg}_{1,P,c_1}(f_{1,\varepsilon}) &\leq \inf_{f_1 \in H} R^{reg}_{1,P,c_1}(f_1) + \varepsilon \leq R^{reg}_{1,P,c_1}(0) + \varepsilon \\ &= R_{1,P}(0) + \varepsilon \leq 2(L_1(1,0,\lambda_1) + L_1(-1,0,\lambda_1)).\end{aligned}$$

The regularization term Ω is bounded for $f_{1,\varepsilon}$, and there exists $\delta_1 > 0$ such that $||f_{1,\varepsilon}||_H \leq \delta_1$. By the Eberlein-Smulyan theorem and the Bolzano-Weierstrass theorem, there exist $f_{1,P,c_1} \in \delta_1 B_H$, $b \in [0, \delta_1]$ and a sequence (f_{1,ε_m}) such that

W.X. Xu
School of Information Management, Shanghai Lixin University of Accounting and Finance, Shanghai 201209, China
E-mail: 20190092@lixin.edu.cn

D.J. Huang
School of Data Science and Engineering, East China Normal University, Shanghai 200062, China
E-mail: djhuang@dase.ecnu.edu.cn

S.G. Zhou
School of Computer Science and Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai 200433, China
E-mail: sgzhou@fudan.edu.cn

$||f_{1,\varepsilon_m}||_H \to b$ and $f_{1,\varepsilon_m} \to f_{1,P,c_1}$ weakly with m tending to infinity. Considering the continuity of L_1 and the reproducing property of H, we have

$$L_1(y, f_{1,\varepsilon_m}(x), \lambda_1) \longrightarrow L_1(y, f_{1,P,c_1}(x), \lambda_1)$$

with m tending to infinity for any $(x, y) \in X \times Y$. Taking the expectation on both sides, it derives that

$$R_{1,P}(f_{1,\varepsilon_m}) \longrightarrow R_{1,P}(f_{1,P,c_1}). \tag{1}$$

Then for any $\rho > 0$, there exists m_0 such that for any $m > m_0$ we have $\varepsilon_m \leq \rho$ and

$$\begin{aligned} R_{1,P}(f_{1,P,c_1}) + \Omega(c_1, ||f_{1,\varepsilon_m}||_H) - \rho &\leq R_{1,P}(f_{1,\varepsilon_m}) + \Omega(c_1, ||f_{1,\varepsilon_m}||_H) \\ &\leq R_{1,P}(f_{1,P,c_1}) + \Omega(c_1, ||f_{1,P,c_1}||_H) + \varepsilon_m. \end{aligned}$$

For the arbitrary smallness of ρ, it indicates that

$$\lim_{m\to\infty} \Omega(c_1, ||f_{1,\varepsilon_m}||_H) \leq \Omega(c_1, ||f_{1,P,c_1}||_H).$$

On the other hand, according to Corollary 2.6 in [2] we have

$$||f_{1,P,c_1}||_H \leq \liminf_{m\to\infty} ||f_{1,\varepsilon_m}||_H = b.$$

Then, it concludes that

$$\Omega(c_1, ||f_{1,P,c_1}||_H) \leq \Omega(c_1, b) = \lim_{m\to\infty} \Omega(c_1, ||f_{1,\varepsilon_m}||_H).$$

Therefore, it derives from Eq. (1) that

$$R^{reg}_{1,P,c_1}(f_{1,\varepsilon_m}) \longrightarrow R^{reg}_{1,P,c_1}(f_{1,P,c_1}), \;\; m \to \infty.$$

Since the construction of f_{1,ε_m} yields the relationship

$$R^{reg}_{1,P,c_1}(f_{1,\varepsilon_m}) \longrightarrow \inf_{f_1\in H} R^{reg}_{1,P,c_1}(f_1), \;\; m \to \infty,$$

we have the following equaiton

$$R^{reg}_{1,P,c_1}(f_{1,P,c_1}) = \inf_{f_1\in H} R^{reg}_{1,P,c_1}(f_1).$$

Moreover, considering the definition of δ_{c_1}, it is obvious that $\delta_{c_1} \geq \delta_1$ and that $||f_{1,P,c_1}||_H \leq b \leq \delta_1 \leq \delta_{c_1}$, which completes the proof of the first part of the theorem. Since the loss functions L_1 and L_2 are constructed in a symmetric way, we would have the symmetric results

$$R^{reg}_{2,P,c_2}(f_{2,P,c_2}) = \inf_{f_2\in H} R^{reg}_{2,P,c_2}(f_2), \; ||f_{2,P,c_2}||_H \leq \delta_{c_2}.$$

The theorem has been proven.

1.2 Proof of Lemma 1

Proof Since $R_{1,P}$ and $R_{2,P}$ are constructed in a symmetric way, the results for $R_{1,P}$ and $R_{2,P}$ could be derived in a symmetric way. Here we only need to get the result for $R_{1,P}$. Let (α_i) be a dense sequence in $[0,1]$ with $\alpha_1 = 0$ and $\alpha_i \neq \alpha_j$ if $i \neq j$. Let t_i be the solution of

$$C_1(\alpha, t_\alpha, \lambda_1) = \min_{t\in\mathbb{R}} C_1(\alpha, t, \lambda_1), \tag{2}$$

for each α_i, $i \geq 1$. Let

$$f_{1,m}(\alpha) := \min\{t_1, \ldots, t_m\},$$
$$\hat{f}_1(\alpha) = \lim \inf_{m\to\infty} f_{1,m}(\alpha).$$

Select a real number $\hat{\alpha} \in (0,1)$ at which M_1 is continuous. There exist two subsequences (α_{i_j}) and (t_{i_j}) such that $(\alpha_{i_j}) \to \hat{\alpha}$ and $(t_{i_j}) \to \hat{f}_1(\hat{\alpha})$ with j tending to infinity. It follows from the continuity of L_1 that

$$\begin{aligned} M_1(\hat{\alpha}, \lambda_1) &= \lim_{j\to\infty} M_1(\alpha_{i_j}, \lambda_1) \\ &= \lim_{j\to\infty} [\alpha_{i_j} L_1(1, t_{i_j}, \lambda_1) + (1-\alpha_{i_j}) L_1(-1, t_{i_j}, \lambda_1)] \\ &= \hat{\alpha} L_1(1, \hat{f}_1(\hat{\alpha}), \lambda_1) + (1-\hat{\alpha}) L_1(-1, \hat{f}_1(\hat{\alpha}), \lambda_1), \end{aligned}$$

which implies that $\hat{f}_1(\hat{\alpha})$ is a solution of Eq. (2) for $\hat{\alpha}$. Since M_1 is concave, it is also continuous for all but at most countably many points of $[0,1]$ by Theorem 1.16 in [4]. Then in order to construct f_1^* such that M_1 is continuous for all points, we only have to modify $\hat{f}_1$ on the countably many points. Therefore,

$$\begin{aligned} R_{1,P} &= \inf\{R_{1,P}(f_1) | f_1 : X \to \mathbb{R} \text{ measurable}\} \\ &= \inf_{f_1} \int_{(x,y)\sim P} L_1(y, f_1(x), \lambda_1) dP(x,y) \\ &= \inf_{f_1} \int_X \Big[P(1|x) L_1(1, f_1(x), \lambda_1) + (1-P(1|x)) L_1(-1, f_1(x), \lambda_1)\Big] P_X(dx) \\ &= \inf_{f_1} \int_X C_1(P(1|x), f_1(x), \lambda_1) P_X(dx) = \int_X M_1(f_1^*(P(1|x)), \lambda_1) P_X(dx). \end{aligned}$$

The lemma has been proven.

1.3 Proof of Theorem 2

Proof Since $R_{1,P}$ and $R_{2,P}$ are constructed in a symmetric way, the results for $R_{1,P}$ and $R_{2,P}$ could be derived in a symmetric way. Here we only need to get the result for $R_{1,P}$.

By the definition of infimum, there exists $f_{1,\varepsilon} \in H$ such that

$$R_{1,P}(f_{1,\varepsilon}) \leq \inf_{f_1 \in H} R_{1,P}(f_1) + \varepsilon$$

for any $\varepsilon > 0$. Since the regularization term $\Omega(\cdot, ||f_{1,\varepsilon}||_H)$ is continuous in 0, there exists $\hat{c}_1$ such that for any $c_1 < \hat{c}_1$ we have $\Omega(c_1, ||f_{1,\varepsilon}||_H) \leq \varepsilon$. It follows from Theorem 1 that

$$\lim_{c_1 \to 0} R^{reg}_{1,P,c_1}(f_{1,P,c_1}) = \lim_{c_1 \to 0} \inf_{f_1 \in H} R^{reg}_{1,P,c_1}(f_1) = \inf_{f_1 \in H} R_{1,P}(f_1).$$

In order to prove the theorem, it suffice to give the proof that

$$\inf_{f_1 \in H} R_{1,P}(f_1) = \inf\{R_{1,P}(f_1) : f_1 \in L_\infty(P_X)\} = R_{1,P}. \tag{3}$$

Since k is a universal kernel, for any $\epsilon, \delta > 0$ and any bounded measurable function $h : X \to \mathbb{R}$, there exists $f_1 \in H$ such that

$$\begin{aligned} &P_X(\{x \in X : |h(x) - f_1(x)| \geq \epsilon\}) \leq \delta, \\ &||f_1||_\infty \leq ||h||_\infty \in L_\infty(P_X). \end{aligned}$$

Since L_1 is uniformly continuous on $Y \times [-||h||_\infty, ||h||_\infty]$, the first equality of Eq. (3) is valid.

Now we begin to show the validity for the last equality of Eq. (3). Define two functions $f_{1,m} : [0,1] \to \mathbb{R}$ and $M_{1,m}(\alpha, \lambda_1)$ as follows:

$$\begin{aligned} f_{1,m}(\alpha) &= \begin{cases} f_1^*(\alpha), & |f_1^*(\alpha)| \leq m, \\ 0, & \text{otherwise}, \end{cases} \\ M_{1,m}(\alpha, \lambda_1) &= \alpha L_1(1, f_{1,m}(\alpha), \lambda_1) + (1-\alpha) L_1(-1, f_{1,m}(\alpha), \lambda_1), \end{aligned}$$

where f_1^* is constructed as in Lemma 1. For any $\alpha \in [0,1]$ with $|f^*(\alpha)| < \infty$, we have

$$\begin{aligned} M_{1,m}(\alpha, \lambda_1) =& \mathbf{1}_{[-m,m]}(f_1^*(\alpha)) M_1(\alpha, \lambda_1) \\ &+ \mathbf{1}_{\mathbb{R}\setminus[-m,m]}(f_1^*(\alpha))(\alpha L_1(1, 0, \lambda_1) + (1-\alpha) L_1(-1, 0, \lambda_1)). \end{aligned}$$

Moreover as m tends to infinity, $M_{1,m}(\alpha, \lambda_1)$ is monotonically decreasing bounded in $M_1(\alpha, \lambda_1)$ with respect to α , according to the definition of $M_1(\alpha, \lambda_1)$, i.e.,

$$|M_{1,m}(\alpha, \lambda_1) - M_1(\alpha, \lambda_1)| \to 0.$$

Therefore, it derives from Lemma 1 that

$$\begin{aligned} R_{1,P} &= \int_X M_1(P(1|x), \lambda_1) P_X(dx) \\ &= \lim_{m \to \infty} \int_X M_{1,m}(P(1|x), \lambda_1) P_X(dx) \\ &= \inf\{R_{1,P}(h) | h : X \to \mathbb{R} \text{ bounded, measurable}\}. \end{aligned}$$

The theorem has been proven.

1.4 Proof of Theorem 3

Proof The set of points misclassified by f is defined as

$$E_f := \{x \in X : P(1|x) < \frac{1}{2},\ f(x) > 0\} \bigcup \{x \in X : P(1|x) > \frac{1}{2},\ f(x) < 0\}.$$

Given a measurable functions $f_1^* : [0,1] \to \mathbb{R}$ according to Lemma 1, we have

$$\begin{aligned}
\delta_1 \geq & R_{1,P}(f_1) - R_{1,P} \\
= & \int_{X \setminus E_f} C_1(P(1|x), f_1(x), \lambda_1) P_X(dx) + \int_{E_f} C_1(P(1|x), f_1(x), \lambda_1) P_X(dx) \\
& - \int_{X \setminus E_f} C_1(P(1|x), f_1^*(P(1|x)), \lambda_1) P_X(dx) \\
& - \int_{E_f} C_1(P(1|x), f_1^*(P(1|x)), \lambda_1) P_X(dx) \\
\geq & \int_{E_f} C_1(P(1|x), f_1(P(1|x)), \lambda_1) P_X(dx) - \int_{E_f} M_1(P(1|x), \lambda_1) P_X(dx).
\end{aligned}$$

In order to estimate the first term on the right-hand side, define a novel function $\tilde{f}_1 : [0,1] \to \mathbb{R}$ such that

$$C_1(\alpha, \tilde{f}_1(\alpha), \lambda_1) = \begin{cases} \inf\{C_1(\alpha, t, \lambda_1) : t \geq 0\}, & \alpha < \dfrac{1}{2}, \\ \inf\{C_1(\alpha, t, \lambda_1) : t \leq 0\}, & \alpha > \dfrac{1}{2}. \end{cases}$$

Note, $\tilde{f}_1$ is assumed to be measurable by the technique in the proof of Lemma 1. Since L_1 is admissible, we have $C_1(\alpha, \tilde{f}_1(\alpha), \lambda_1) - M_1(\alpha, \lambda_1) > 0$ for any $\alpha \in [0,1] \backslash \{1/2\}$. Indeed, for $\alpha < 1/2$, we have $t_\alpha < 0$, implying that

$$C_1(\alpha, \tilde{f}_1(\alpha), \lambda_1) = \inf\{C_1(\alpha, t, \lambda_1) : t \geq 0\} > C_1(\alpha, t_\alpha, \lambda_1) = M_1(\alpha, \lambda_1).$$

For $\alpha > 1/2$, we have $t_\alpha > 0$, implying that

$$C_1(\alpha, \tilde{f}_1(\alpha), \lambda_1) = \inf\{C_1(\alpha, t, \lambda_1) : t \leq 0\} > C_1(\alpha, t_\alpha, \lambda_1) = M_1(\alpha, \lambda_1).$$

Define $\Delta_1 : X \to \mathbb{R}$ as follows:

$$\Delta_1(x) = C_1(P(1|x), \tilde{f}_1(P(1|x)), \lambda_1) - M_1(P(1|x), \lambda_1).$$

Δ_1 is a strictly positive function on $\hat{X} := \{x \in X : P(1|x) \neq 1/2\}$, thus

$$0 < \int_{E_f} \Delta_1(x) dx \leq \delta_1.$$

Similarly, we could get another symmetric result

$$0 < \int_{E_f} \Delta_2(x) dx \leq \delta_2$$

for function $\Delta_2 : X \to \mathbb{R}$ as follows:

$$\Delta_2(x) = C_2(P(1|x), \tilde{f}_2(P(1|x)), \lambda_2) - M_2(P(1|x), \lambda_2).$$

As δ_1, δ_2 tends to 0, it derives that $P\{x \in E_f\} \to 0$. Note, Theorem 2.6 in [3] indicates that

$$R_P(f) - R_P = \int_{E_f} (1 - 2\eta(x)) P_X(dx),$$

where $\eta(x) = \min\{P(1|x), P(-1|x)\}$. The function $1 - 2\eta(x)$ is also strictly positive on X. Thus, the equation $R_P(f) - R_P \longrightarrow 0$ holds true. The theorem has been proven.

1.5 Proof of Lemma 2

Proof Since the hyper-planes f_1 and f_2 are constructed in a symmetric way, the results for f_1 and f_2 could also be derived in a similar way. We only need to get the result for f_1.

Denote $\mathcal{F}_1 = \{L_1(\cdot, f_1(\cdot), \lambda_1) : f_1 \in \delta_{c_1} I(B_H)\}$. Let $\{f_1^1, \ldots, f_1^n\}$ be the ϵ-net of $\delta_{c_1} I(B_H)$. Then $\{f_1^1, \ldots, f_1^n\}$ is shown to be the $w(L_{1,c_1}, \epsilon)$-net of $\mathcal{F}_1$ by virtue of the definition of the modulus of continuity. Thus, suppose there exists a minimal $w^{-1}(L_{1,c_1}, \epsilon)$-net for $\delta_{c_1} I(B_H)$, then there exists a ϵ-net for $\mathcal{F}_1$ by virtue of the definition of the modulus of continuity. Therefore, $\mathcal{N}(\mathcal{F}_1, \epsilon) \leq \mathcal{N}(\delta_{c_1} I, w^{-1}(L_{1,c_1}, \epsilon))$. Since $\mathcal{F}_1$ is a subset of of nonnegative functions that are bounded by $||L_{1,c_1}||_\infty$, it follows from Hoeffding's inequality for Theorem 8.1 in [3] that

$$\begin{aligned}
&Pr\Big\{S : \sup_{f_1 \in \delta_{c_1} I(B_H)} |R_{1,S}(f_1) - R_{1,P}(f_1)| \geq \epsilon\Big\} \\
\leq &2\mathcal{N}(\mathcal{F}_1, \epsilon/3) \exp\Big\{ - \frac{2\epsilon^2 m}{9||L_{1,c_1}||_\infty^2}\Big\} \\
\leq &2 \exp\Big\{\mathcal{H}(\delta_{c_1} I, \omega^{-1}(L_{1,c_1}, \epsilon/3)) - \frac{2\epsilon^2 m}{9\|L_{1,c_1}\|_\infty^2}\Big\}.
\end{aligned}$$

Since Theorem 1 guarantees that $f_{1,S,c_1} \in \delta_{c_1} I(B_H)$, we have

$$\begin{aligned}
&Pr\{S : |R_{1,S}(f_{1,S,c_1}) - R_{1,P}(f_{1,S,c_1})| \geq \epsilon\} \\
\leq &Pr\Big\{S : \sup_{f_1 \in \delta_{c_1} I(B_H)} |R_{1,S}(f_1) - R_{1,P}(f_1)| \geq \epsilon\Big\} \\
\leq &2 \exp\Big\{\mathcal{H}(\delta_{c_1} I, \omega^{-1}(L_{1,c_1}, \epsilon/3)) - \frac{2\epsilon^2 m}{9\|L_{1,c_1}\|_\infty^2}\Big\}.
\end{aligned}$$

The lemma has been proven.

1.6 Proof of Theorem 4

Proof Define a variable

$$g_1 := \sup\{|L_1(y,t,\lambda_1) - L_1(y,t^{'},\lambda_1)| : y \in Y, t, t^{'} \in [-\hat{\delta}_1 K, \hat{\delta}_1 K]\}.$$

Let $\epsilon \in (0, g_1)$, and fix δ_1 as in Theorem 3 satisfying $R_{1,P}(f_1) \leq R_{1,P} + \delta_1$ for any measurable function $f_1 : X \to \mathbb{R}$. Applying Theorem 2, there exists $m_0 \in \mathbb{N}$ such that for any $m > m_0$, we have

$$|R^{reg}_{1,P,c_1(m)}(f_{1,P,c_1(m)}) - R_{1,P}| \leq \delta_1/3.$$

Note, the quantity

$$\sup\{w^{-1}(L_{1,c_1(m)}, \epsilon) : c_1(m) \in (0,1], \epsilon \in (0, g_1)\} < \infty$$

ensures that there exists ρ_1 such that for any $m > m_0$, we have

$$\mathcal{H}(\delta_{c_1(m)} I, w^{-1}(L_{1,c_1(m)}, \epsilon)) \geq \rho_1,$$

i.e., there are at least two points to cover $\delta_{c_1(m)} I(B_H)$. Thus it follows from the condition on $c_1(m)$ that

$$\frac{\|L_{1,c_1(m)}\|^2_\infty}{m} \longrightarrow 0, \;\; m \longrightarrow \infty$$
$$\exp\Big\{\mathcal{H}(\delta_{c_1(m)} I, \omega^{-1}(L_{1,c_1(m)}, \epsilon/3)) - \frac{2\epsilon^2 m}{9\|L_{1,c_1(m)}\|^2_\infty}\Big\} \longrightarrow 0, \;\; m \longrightarrow \infty.$$

Using Lemma 2 together with Hoeffding's inequality, we have

$$Pr\Big\{S : |R_{1,S}(f_{1,S,c_1(m)}) - R_{1,P}(f_{1,S,c_1(m)})| > \frac{\delta_1}{3}$$
$$\bigcup |R_{1,S}(f_{1,P,c_1(m)}) - R_{1,P}(f_{1,P,c_1(m)})| > \frac{\delta_1}{3}\Big\} \leq \epsilon,$$

for any $m > m_0$. Furthermore,

$$\begin{aligned} R_{1,P}(f_{1,S,c_1(m)}) &\leq \Omega(c_1(m), ||f_{1,S,c_1(m)}||_H) + R_{1,P}(f_{1,S,c_1(m)}) \\ &\leq \Omega(c_1(m), ||f_{1,S,c_1(m)}||_H) + R_{1,S}(f_{1,S,c_1(m)}) + \delta_1/3 \\ &\leq \Omega(c_1(m), ||f_{1,P,c_1(m)}||_H) + R_{1,S}(f_{1,P,c_1(m)}) + \delta_1/3 \\ &\leq \Omega(c_1(m), ||f_{1,P,c_1(m)}||_H) + R_{1,P}(f_{1,P,c_1(m)}) + 2\delta_1/3 \\ &\leq R_{1,P} + \delta_1. \end{aligned}$$

Since the hyper-planes f_1 and f_2 are constructed in a symmetric way, the results for f_1 and f_2 could be derived in a similar way. Analogously, we have the result

$$R_{2,P}(f_{2,S,c_2(m)}) \leq R_{2,P} + \delta_2,$$

where $c_2(m)$ is defined the same as $c_1(m)$ and δ_2 is defined as in Theorem 3. It concludes from Theorem 3 that $R_P(f_S) \leq R_P + \epsilon$, where $f_S(\cdot) = |f_{2,S,c_2(m)}(\cdot)| -$

$|f_{1,S,c_1(m)}(\cdot)|$, implying the optimal problem Eq. (3) (in the main paper) is universally consistent.

Now we start to show the strongly universal consistency. The conditions

$$\begin{aligned}\sum_{m=1}^{\infty}\exp\{-\epsilon m/||L_{1,c_1(m)}||_{\infty}^2\}<\infty,\\ \sum_{m=1}^{\infty}\exp\{-\epsilon m/||L_{2,c_2(m)}||_{\infty}^2\}<\infty,\end{aligned}\tag{4}$$

ensure that

$$\begin{aligned}\exp\{-\epsilon m/||L_{1,c_1(m)}||_{\infty}^2\}\longrightarrow 0,\ \ m\longrightarrow\infty,\\ \exp\{-\epsilon m/||L_{2,c_2(m)}||_{\infty}^2\}\longrightarrow 0,\ \ m\longrightarrow\infty.\end{aligned}$$

Furthermore,

$$\begin{aligned}\exp\Big\{\mathcal{H}(\delta_{c_1}I,\omega^{-1}(L_{1,c_1},\epsilon/3))-\frac{2\epsilon^2 m}{9\|L_{1,c_1}\|_{\infty}^2}\Big\}\longrightarrow 0,\ \ m\longrightarrow\infty,\\ \exp\Big\{\mathcal{H}(\delta_{c_2}I,\omega^{-1}(L_{2,c_2},\epsilon/3))-\frac{2\epsilon^2 m}{9\|L_{2,c_2}\|_{\infty}^2}\Big\}\longrightarrow 0,\ \ m\longrightarrow\infty.\end{aligned}$$

Thus the optimal problem is strongly universally consistent.

1.7 Proof of Lemma 3

Proof Since the hyper-planes f_1 and f_2 are constructed in a symmetric way, the results for f_1 and f_2 could be derived in a similar way. We only need to get the result for f_1.

Denote $\mathcal{F}_1=\{L_1(\cdot,f(\cdot),\lambda_1):f\in\delta_{c_1}I(B_H)\}$. As discussed in the proof of Lemma 2, it derives for localized covering number that

$$\mathcal{N}(\mathcal{F}_1,2m,\epsilon)\leq\mathcal{N}(\delta_{c_1}I,2m,w^{-1}(L_{1,c_1},\epsilon)).$$

It follows from Lemma 3.4 in [1] that

$$\begin{aligned}&Pr\Big\{S:\sup_{f_1\in\delta_{c_1}I(B_H)}|R_{1,S}(f_1)-R_{1,P}(f_1)|\geq\epsilon\Big\}\\ \leq&12m\mathcal{N}(\mathcal{F}_1,2m,\epsilon/6)\exp\Big\{-\frac{\epsilon^2 m}{36||L_{1,c_1}||_{\infty}^2}\Big\}\\ \leq&12m\exp\Big\{\mathcal{H}(\delta_{c_1}I,2m,w^{-1}(L_{1,c_1},\epsilon/6))-\frac{\epsilon^2 m}{36||L_{1,c_1}||_{\infty}^2}\Big\}.\end{aligned}$$

Since Lemma 1 guarantees that $f_{1,S,c_1}\in\delta_{c_1}I(B_H)$, we have

$$\begin{aligned}&Pr\Big\{S:|R_{1,S}(f_{1,S,c_1})-R_{1,P}(f_{1,S,c_1})|\geq\epsilon\Big\}\\ \leq&12m\exp\Big\{\mathcal{H}(\delta_{c_1}I,2m,w^{-1}(L_{1,c_1},\epsilon/6))-\frac{\epsilon^2 m}{36||L_{1,c_1}||_{\infty}^2}\Big\}.\end{aligned}$$

The lemma has been proven.

1.8 Proof of Theorem 5

Proof Let g_1, ϵ, δ_1 and m_0 be defined the same as that in the proof of Theorem 4. Note that, the quantity

$$\sup\{w^{-1}(L_{1,c_1(m)},\epsilon) : c_1(m) \in (0,1], \epsilon \in (0,g_1)\} < \infty$$

ensures there exists ρ_1 such that for any $m > m_0$, we have

$$\mathcal{H}(\delta_{c_1(m)}I, 2m, w^{-1}(L_{1,c_1(m)},\epsilon)) \geq \rho_1,$$

i.e., there are at least two points to cover $\delta_{c_1(m)}I(B_H)$. Thus, it follows from the condition on $c_1(m)$ that

$$\frac{\|L_{1,c_1(m)}\|_\infty^2}{m} \longrightarrow 0, \;\; m \longrightarrow \infty,$$
$$\exp\Big\{\mathcal{H}(\delta_{c_1}I, 2m, \omega^{-1}(L_{1,c_1},\epsilon/6)) - \frac{\epsilon^2 m}{36\|L_{1,c_1}\|_\infty^2}\Big\} \longrightarrow 0, \;\; m \longrightarrow \infty.$$

Using Lemma 3 together with Hoeffding's inequality, for any $m > m_0$ we have

$$Pr\Big(S : |R_{1,S}(f_{1,S,c_1(m)}) - R_{1,P}(f_{1,S,c_1(m)})| > \delta_1/3 \bigcup |R_{1,S}(f_{1,P,c_1(m)}) - R_{1,P}(f_{1,P,c_1(m)})| > \delta_1/3\Big) \leq \epsilon.$$

According to the procedures in the proof of Theorem 4, it derives that

$$R_{1,P}(f_{1,S,c_1(m)}) \leq R_{1,P} + \delta_1,$$

Analogously, we have the result for $f_{2,S,c_2(m)}$ that

$$R_{2,P}(f_{2,S,c_2(m)}) \leq R_{2,P} + \delta_2,$$

where $c_2(m)$ and δ_2 are defined the same as $c_1(m)$ and δ_1. It yields by Theorem 3 that $R_P(f_S) \leq R_P + \epsilon$, where $f_S(\cdot) = |f_{2,S,c_2(m)}(\cdot)| - |f_{1,S,c_1(m)}(\cdot)|$, implying the optimal problem Eq. (3) (in the main paper) is universally consistent.

Now we start to show the strongly universal consistency. The conditions

$$\sum_{m=1}^{\infty} \exp\{-\epsilon m/||L_{1,c_1(m)}||_\infty^2\} < \infty,$$
$$\sum_{m=1}^{\infty} \exp\{-\epsilon m/||L_{2,c_2(m)}||_\infty^2\} < \infty,$$

ensures that

$$\exp\Big\{\mathcal{H}(\delta_{c_1}I, 2m, \omega^{-1}(L_{1,c_1},\epsilon/6)) - \frac{\epsilon^2 m}{36\|L_{1,c_1}\|_\infty^2}\Big\} \longrightarrow 0, \;\; m \longrightarrow \infty,$$
$$\exp\Big\{\mathcal{H}(\delta_{c_2}I, 2m, \omega^{-1}(L_{2,c_2},\epsilon/6)) - \frac{\epsilon^2 m}{36\|L_{2,c_2}\|_\infty^2}\Big\} \longrightarrow 0, \;\; m \longrightarrow \infty.$$

Thus, the optimal problem is strongly universally consistent.

1.9 Proof of Lemma 4

Proof Since the lemma is valid for differentiable loss functions by virtue of Theorem 12.4 in [6], we only consider the lemma for non-differentiable loss functions below. The idea of the proof mainly follows the approach of [8].

Let $\hat{S} = S_{i,(x,y)}$ for some fixed $i = 1, \ldots, m$. Let $E_S f$ be the expectation of f with respect to the empirical measure induced by S. Recall for a convex, continuous function $f_1 : X \to \mathbb{R}$, the sub-differential of f_1 in $x \in X$ is defined by

$$\partial f_1(x) := \{x^* \in X : < x^*, x^{'} - x > \leq f_1(x^{'}) - f_1(x), \forall x^{'} \in X\}.$$

The convexity of L_1 implies that $L_{1,c_1(m)}$ is locally 1-Hölder-continuous. Actually, we only have to consider the case for finite dimensional subspaces of H. Theorem 23.8 and 23.9 in [5] show that

$$R^{reg}_{1,S,c_1(m)}(f_1) = 2c_1(m)f_1 + D,$$

where

$$D = \{E_S h\Phi : h(x_i, y_i) \in \partial L_1(y_i, f_1(x_i), \lambda_1), \forall i = 1, \ldots, m\},$$

and the sub-differential ∂L_1 is only with respect to the second variable of L_1. Since $f_{1,S,c_1(m)}$ is the element minimizing $R^{reg}_{1,S,c_1(m)}$, and $0 \in \partial R^{reg}_{1,S,c_1(m)}(f_{1,S,c_1(m)})$, there exists $h(x_i, y_i) \in \partial L_1(y_i, f_{1,S,c_1(m)}(x_i), \lambda_1)$, $i = 1, \ldots, m$, such that

$$0 \leq 2c_1(m)f_{1,S,c_1(m)} + E_S h\Phi = R^{reg}_{1,S,c_1(m)}(f_{1,S,c_1(m)}) \leq R^{reg}_{1,S,c_1(m)}(0) = 0.$$

Note, $f_{1,S,c_1(m)}$ has an upper bound by virtue of the proof of Lemma 1. It follows from the Lipschitz continuity of $L_{1,c_1(m)}$ that $||h||_\infty \leq |L_{1,c_1(m)}|_1$. Thus,

$$\begin{aligned} &h(x_i, y_i)(f_{1,\hat{S},c_1(m)}(x_i) - f_{1,S,c_1(m)}(x_i)) \\ \leq &L_1(y_i, f_{1,\hat{S},c_1(m)}(x_i), \lambda_1) - L_1(y_i, f_{1,S,c_1(m)}(x_i), \lambda_1), i = 1, \ldots, m. \end{aligned}$$

Taking expectation with respect to the empirical measure of $\hat{S}$ on both sides, it derives

$$\begin{aligned} &< f_{1,\hat{S},c_1(m)} - f_{1,S,c_1(m)}, E_{\hat{S}} h\Phi > \\ \leq &E_{\hat{S}} L_1(\cdot, f_{1,\hat{S},c_1(m)}(\cdot), \lambda_1) - E_{\hat{S}} L_1(\cdot, f_{1,S,c_1(m)}(\cdot), \lambda_1), \end{aligned}$$

for the reproducing property of Φ. Since

$$\begin{aligned} &||f_{1,S,c_1(m)}||^2 + 2 < f_{1,\hat{S},c_1(m)} - f_{1,S,c_1(m)}, f_{1,S,c_1(m)} > + ||f_{1,S,c_1(m)} - f_{1,\hat{S},c_1(m)}||^2 \\ = &||f_{1,\hat{S},c_1(m)}||^2, \end{aligned}$$

we have

$$\begin{aligned}
&R^{reg}_{1,\hat{S},c_1(m)}(f_{1,S,c_1(m)}) + c_1(m)||f_{1,S,c_1(m)} - f_{1,\hat{S},c_1(m)}||^2 \\
&+ < f_{1,\hat{S},c_1(m)} - f_{1,S,c_1(m)}, E_{\hat{S}}h\Phi + 2c_1(m)f_{1,S,c_1(m)} > \\
=&E_{\hat{S}}L_1(\cdot, f_{1,S,c_1(m)}(\cdot), \lambda_1) + c_1(m)||f_{1,S,c_1(m)}||^2 \\
&+ c_1(m)||f_{1,S,c_1(m)} - f_{1,\hat{S},c_1(m)}||^2 + < f_{1,\hat{S},c_1(m)} - f_{1,S,c_1(m)}, E_{\hat{S}}h\Phi > \\
&+ 2c_1(m) < f_{1,\hat{S},c_1(m)} - f_{1,S,c_1(m)}, f_{1,S,c_1(m)} > \\
=&E_{\hat{S}}L_1(\cdot, f_{1,S,c_1(m)}(\cdot), \lambda_1) + < f_{1,\hat{S},c_1(m)} - f_{1,S,c_1(m)}, E_{\hat{S}}h\Phi > + ||f_{1,\hat{S},c_1(m)}||^2 \\
\leq&E_{\hat{S}}L_1(\cdot, f_{1,\hat{S},c_1(m)}(\cdot), \lambda_1) + ||f_{1,\hat{S},c_1(m)}||^2 \\
=&R^{reg}_{1,\hat{S},c_1(m)}(f_{1,\hat{S},c_1(m)}).
\end{aligned}$$

Since $f_{1,\hat{S},c_1(m)}$ minimizes $R^{reg}_{1,\hat{S},c_1(m)}$, it follows that

$$R^{reg}_{1,\hat{S},c_1(m)}(f_{1,S,c_1(m)}) \geq R^{reg}_{1,\hat{S},c_1(m)}(f_{1,\hat{S},c_1(m)}).$$

Thus, we have

$$\begin{aligned}
&c_1(m)||f_{1,S,c_1(m)} - f_{1,\hat{S},c_1(m)}||^2 \\
\leq& < f_{1,S,c_1(m)} - f_{1,\hat{S},c_1(m)}, E_{\hat{S}}h\Phi + 2c_1(m)f_{1,S,c_1(m)} > \\
=&||f_{1,S,c_1(m)} - f_{1,\hat{S},c_1(m)}||\ ||E_{\hat{S}}h\Phi + 2c_1(m)f_{1,S,c_1(m)}||.
\end{aligned}$$

Furthermore, we have

$$||f_{1,S,c_1(m)} - f_{1,\hat{S},c_1(m)}|| \leq \frac{1}{c_1(m)}||E_{\hat{S}}h\Phi - E_S h\Phi|| \leq \frac{2K|L_{1,c_1(m)}|_1}{mc_1(m)}.$$

Therefore,

$$\begin{aligned}
&|L_1(\cdot, f_{1,S,c_1(m)}(\cdot), \lambda_1) - L_1(\cdot, f_{1,\hat{S},c_1(m)}(\cdot), \lambda_1)| \\
\leq&K|L_{1,c_1(m)}|_1\ ||f_{1,S,c_1(m)} - f_{1,\hat{S},c_1(m)}|| \\
=&\frac{2K^2|L_{1,c_1(m)}|_1^2}{mc_1(m)}.
\end{aligned}$$

Since the hyper-planes f_1 and f_2 are constructed in a symmetric way, the results for f_1 and f_2 could be derived in a similar way. Thus,

$$|L_2(\cdot, f_{2,S,c_2(m)}(\cdot), \lambda_2) - L_2(\cdot, f_{2,\hat{S},c_2(m)}(\cdot), \lambda_2)| \leq \frac{2K^2|L_{2,c_2(m)}|_1^2}{mc_2(m)}.$$

The above two inequalities satisfy the definition of stability for the optimal problem Eq. (3) (in the main paper). The lemma has been proven.

1.10 Proof of Lemma 5

Proof The classifiers f_1 and f_2 are both stable [7] according to the definition of stability for the classifier f. It follows from Lemma 3.21 in [7] that

$$\begin{aligned} &Pr\{S : |R_{1,S}(f_{1,S,c_1}) - R_{1,P}(f_{1,S,c_1})| > \epsilon + \beta_1(m)\} \\ &\qquad \leq 2\exp\Big\{-\frac{\epsilon^2 m}{2(m\beta_1(m) + \|L_{1,c_1(m)}\|_\infty)^2}\Big\}, \\ &Pr\{S : |R_{2,S}(f_{2,S,c_2}) - R_{2,P}(f_{2,S,c_2})| > \epsilon + \beta_2(m)\} \\ &\qquad \leq 2\exp\Big\{-\frac{\epsilon^2 m}{2(m\beta_2(m) + \|L_{2,c_2(m)}\|_\infty)^2}\Big\}. \end{aligned}$$

The lemma has been proven.

1.11 Proof of Theorem 6

Proof By virtue of the definitions of δ_{c_1} and L_{1,c_1}, there exist constants a and b such that

$$||L_{1,c_1(m)}||_\infty \leq a\delta_{c_1(m)}|L_{1,c_1(m)}|_1,\ \delta_{c_1(m)} \leq \frac{b}{\sqrt{c_1(m)}},$$

for any $m \geq 1$ and for the sequence $(c_1(m))$. Thus, with m tending to infinity, we have

$$\frac{||L_{1,c_1(m)}||_\infty^2}{\sqrt{m}} \leq \frac{a^2\delta_{c_1(m)}^2|L_{1,c_1(m)}|_1^2}{\sqrt{m}} \leq \frac{a^2b^2|L_{1,c_1(m)}|_1^2}{\sqrt{m}c_1(m)} \longrightarrow 0.$$

$\beta_1(m)$ is defined as in Lemma 5. It derives from Lemmas 4 and 5 that

$$\sqrt{m}\beta_1(m) = \frac{\sqrt{m}2K^2|L_{1,c_1(m)}|_1^2}{mc_1(m)} = \frac{2K^2|L_{1,c_1(m)}|_1^2}{\sqrt{m}c_1(m)} \longrightarrow 0,\ \ m \longrightarrow \infty,$$

since K is a bounded constant. Therefore,

$$\exp\{-\frac{\epsilon^2 m}{2(m\beta_1(m) + \|L_{1,c_1(m)}\|_\infty)^2}\} \longrightarrow 0,\ \ m \longrightarrow \infty.$$

Fix $\delta_1 = \epsilon + \beta_1(m)$, the proof of Theorem 4 yields that

$$R_{1,P}(f_{1,S,c_1(m)}) \leq R_{1,P} + \delta_1.$$

Analogously, we have the result for $f_{2,S,c_2(m)}$ that

$$R_{2,P}(f_{2,S,c_2(m)}) \leq R_{2,P} + \delta_2$$

where $c_2(m)$ is defined the same as $c_1(m)$ and $\delta_2 = \epsilon + \beta_2(m)$. It follows from Theorem 3 that $R_P(f_S) \leq R_P + \epsilon$, where $f_S(\cdot) = |f_{2,S,c_2(m)}(\cdot)| - |f_{1,S,c_1(m)}(\cdot)|$, implying the optimal problem Eq. (3) (in the main paper) is universally consistent.

References

1. Alon, N., Ben-David, S., Cesa-Bianchi, N., Haussler, D.: Scale-sensitive dimensions, uniform convergence, and learnability. In: 34th Annual Symposium on Foundations of Computer Science, pp. 292–301. IEEE Computer Society (1993)
2. Barbu, V., Precupanu, T.: Convexity and Optimization in Banach Spaces, fourth edn. Springer Monographs in Mathematics. Springer (2012)
3. Devroye, L., Györfi, L., Lugosi, G.: A probabilistic theory of pattern recognition, *Stochastic Modelling and Applied Probability*, vol. 31. Springer (1996)
4. Phelps, R.R.: Convex Functions, Monotone Operators and Differentiability, *Lecture Notes in Mathematics*, vol. 1364. Springer (1993)
5. Rockafellar, R.T.: Convex Analysis, *Princeton Landmarks in Mathematics and Physics*, vol. 36. Princeton University Press (1970)
6. Schölkopf, B., Smola, A.J.: Learning with kernels: support vector machines, regularization, optimization, and beyond. Adaptive computation and machine learning series. MIT Press (2002)
7. Steinwart, I.: Consistency of support vector machines and other regularized kernel classifiers. IEEE Transactions on Information Theory **51**(1), 128–142 (2005)
8. Zhang, T.: Convergence of large margin separable linear classification. In: T.K. Leen, T.G. Dietterich, V. Tresp (eds.) Advances in Neural Information Processing Systems 13, pp. 357–363. MIT Press (2001)