APPENDIX A
SMO ALGORITHM

Sequential Minimal Optimization (SMO) is a simple algorithm that can quickly solve the SVM QP problem without any extra matrix storage and without using time-consuming numerical QP optimization steps [1]. SMO decomposes the overall QP problem into the smallest possible optimization problem. This sub-problem can be solved analytically. An appropriate variant of SMO to solve (7) is detailed below following [2].

Given $\boldsymbol{\alpha}$, the algorithm optimizes two variables of $\boldsymbol{\alpha}$ with other variables fixed. Two variables to be optimized should be chosen from $\{\alpha_i \mid i \in I_-\}$ or $\{\alpha_i \mid i \in I_+\}$. Otherwise, the variables which we are trying to optimize cannot change since the other variables are fixed and due to the constraints $\sum_{i\in I_-} \alpha_i = 1$ and $\sum_{i\in I_+} \alpha_i = 1$. Suppose that we choose two variables from $\{\alpha_i \mid i \in I_+\}$. For notational convenience, assume the two variables are α_1 and α_2 and $1, 2 \in I_+$. Then, (7) reduces to

$$\begin{aligned} \min_{\alpha_1,\alpha_2} \quad & \frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{2}\alpha_i\alpha_j Q_{ij} + \sum_{i=1}^{2} d_i\alpha_i + D \\ \text{s.t} \quad & \alpha_1, \alpha_2 \geq 0, \quad \sum_{i=1}^{2}\alpha_i = \Delta \end{aligned}$$

where $D = \frac{1}{2}\sum_{i=3}^{n}\sum_{j=3}^{n}\alpha_i\alpha_j Q_{ij} - \sum_{i=3}^{n} c_i\alpha_i$ and

$$d_i = \sum_{j=3}^{n}\alpha_j Q_{ij} - c_i, \quad \Delta = 1 - \sum_{i\in I_+\setminus\{1,2\}}\alpha_i.$$

We discard D, which is independent of α_1 and α_2, and eliminate α_1 to obtain

$$\begin{aligned} \min_{\alpha_2} \quad & \frac{1}{2}\left(\Delta - \alpha_2\right)^2 Q_{11} + \alpha_2\left(\Delta - \alpha_2\right) Q_{12} \\ & + \frac{1}{2}\alpha_2^2 Q_{22} + \left(\Delta - \alpha_2\right) d_1 + \alpha_2 d_2 \\ \text{s.t} \quad & 0 \leq \alpha_2 \leq \Delta. \end{aligned} \tag{12}$$

Since the objective function is quadratic and convex in one variable α_2, we can take the derivative of (12) and set it equal to zero. Then,

$$\alpha_2 = \frac{\Delta\left(Q_{11} - Q_{12}\right) + d_1 - d_2}{Q_{11} - 2Q_{12} + Q_{22}}. \tag{13}$$

Let $\boldsymbol{\alpha}^*$ denote the value before the optimization step. If we define $O_i := Q_{i1}\alpha_1^* + Q_{i2}\alpha_2^* + d_i = \sum_{j=1}^{n}\alpha_i^* Q_{ij} - c_i$, then (13) can be expressed as the update equation

$$\alpha_2 = \alpha_2^* + \frac{O_1 - O_2}{Q_{11} - 2Q_{12} + Q_{22}}. \tag{14}$$

If α_2 is outside $[0, \Delta]$, we truncate it so that it is within $[0, \Delta]$. After finding α_2, α_1 can be recovered from $\alpha_1 = \Delta - \alpha_2$.

The optimality condition and the choice of α_i's can be found in the following way. There are three cases when choosing α_1 and α_2 : (a) Both are zero, (b) One is positive and the other is zero, (c) Both are positive.

Case (a): α_1 and α_2 are not updated because of nonnegativity constraints.

Case (b): Assume that α_2 is zero. From (14), α_2 is updated only when $O_1 - O_2 > 0$ and so is α_1

Case (c): α_1 and α_2 are updated only when $O_1 \neq O_2$.

The objective value will strictly decrease if and only if α_1 and α_2 are updated after optimization step. Therefore, the optimal solution should satisfy

$$O_i \geq O_j \quad \text{for} \quad \alpha_i = 0,\, \alpha_j > 0 \tag{15}$$

$$O_i = O_j \quad \text{for} \quad \alpha_i, \alpha_j > 0. \tag{16}$$

The convergence to the global minimum is thus guaranteed by choosing two α_i's which do not satisfy (15) or (16) for each optimization step. The optimization procedure for two variables from $\{\alpha_i \in I_-\}$ is similar.

APPENDIX B
PROOF OF LEMMA 1

Note that for any given i, $(k_\sigma(\mathbf{X}_j, \mathbf{X}_i))_{j \neq i}$ are independent and bounded by $M = 1/\left(\sqrt{2\pi}\sigma\right)^d$. For random vectors $\mathbf{Z} \sim f_+(\mathbf{x})$ and $\mathbf{W} \sim f_-(\mathbf{x})$, $h(\mathbf{X}_i)$ in (6) can be expressed as

$$h(\mathbf{X}_i) = \mathbf{E}\left[k_\sigma(\mathbf{Z}, \mathbf{X}_i) \mid \mathbf{X}_i\right] - \gamma \mathbf{E}\left[k_\sigma(\mathbf{W}, \mathbf{X}_i) \mid \mathbf{X}_i\right].$$

Since $\mathbf{X}_i \sim f_+(\mathbf{x})$ for $i \in I_+$ and $\mathbf{X}_i \sim f_-(\mathbf{x})$ for $i \in I_-$, it can be easily shown that

$$\mathbf{E}\left[\widehat{h}_i \mid \mathbf{X}_i\right] = h(\mathbf{X}_i).$$

For $i \in I_+$,

$$\begin{aligned}
&\mathbf{P}\left\{\left|\widehat{h}_i - h(\mathbf{X}_i)\right| > \epsilon \,\middle|\, \mathbf{X}_i = \mathbf{x}, E\right\} \\
&\quad\leq \mathbf{P}\left\{\left|\frac{1}{n_+ - 1}\sum_{j \in I_+, j \neq i} k_\sigma(\mathbf{X}_j, \mathbf{X}_i) - \mathbf{E}\left[k_\sigma(\mathbf{Z}, \mathbf{X}_i) \mid \mathbf{X}_i\right]\right| > \frac{\epsilon}{1+\gamma} \,\middle|\, \mathbf{X}_i = \mathbf{x}\right\} \\
&\qquad + \mathbf{P}\left\{\left|\frac{\gamma}{n_-}\sum_{j \in I_-} k_\sigma(\mathbf{X}_j, \mathbf{X}_i) - \gamma\mathbf{E}\left[k_\sigma(\mathbf{W}, \mathbf{X}_i) \mid \mathbf{X}_i\right]\right| > \frac{\gamma\epsilon}{1+\gamma} \,\middle|\, \mathbf{X}_i = \mathbf{x}\right\} \qquad (17)
\end{aligned}$$

Since we are conditioning on E, the first term in (17) is

$$\begin{aligned}
&\mathbf{P}\left\{\left|\sum_{j \in I_+, j \neq i} k_\sigma(\mathbf{X}_j, \mathbf{X}_i) - (n_+ - 1)\mathbf{E}\left[k_\sigma(\mathbf{Z}, \mathbf{X}_i) \mid \mathbf{X}_i\right]\right| > \frac{(n_+ - 1)\epsilon}{1+\gamma} \,\middle|\, \mathbf{X}_i = \mathbf{x}\right\} \\
&= \mathbf{P}\left\{\left|\sum_{j \in I_+, j \neq i} k_\sigma(\mathbf{X}_j, \mathbf{X}_i) - \mathbf{E}\left[\sum_{j \in I_+, j \neq i} k_\sigma(\mathbf{X}_j, \mathbf{X}_i) \mid \mathbf{X}_i\right]\right| > \frac{(n_+ - 1)\epsilon}{(1+\gamma)} \,\middle|\, \mathbf{X}_i = \mathbf{x}\right\} \\
&= \mathbf{P}\left\{\left|\sum_{j \in I_+, j \neq i} k_\sigma(\mathbf{X}_j, \mathbf{X}_i) - \mathbf{E}\left[\sum_{j \in I_+, j \neq i} k_\sigma(\mathbf{X}_j, \mathbf{X}_i) \mid \mathbf{X}_i\right]\right| > \frac{(n_+ - 1)\epsilon}{(1+\gamma)} \,\middle|\, \mathbf{X}_i = \mathbf{x}\right\} \\
&\leq 2e^{-2(n_+ - 1)\epsilon^2/(1+\gamma)^2 M^2}.
\end{aligned}$$

where the last inequality holds by Hoeffding's inequality [3]. The second term in (17) is

$$\begin{aligned}
&\mathbf{P}\left\{\left|\sum_{j \in I_-} k_\sigma(\mathbf{X}_j, \mathbf{X}_i) - n_-\mathbf{E}\left[k_\sigma(\mathbf{W}, \mathbf{X}_i) \mid \mathbf{X}_i\right]\right| > \frac{n_-\epsilon}{1+\gamma} \,\middle|\, \mathbf{X}_i = \mathbf{x}\right\} \\
&\quad\leq \mathbf{P}\left\{\left|\sum_{j \in I_-} k_\sigma(\mathbf{X}_j, \mathbf{X}_i) - \mathbf{E}\left[\sum_{j \in I_-} k_\sigma(\mathbf{X}_j, \mathbf{X}_i) \mid \mathbf{X}_i\right]\right| > \frac{n_-\epsilon}{1+\gamma} \,\middle|\, \mathbf{X}_i = \mathbf{x}\right\} \\
&\quad\leq 2e^{-2n_-\epsilon^2/(1+\gamma)^2 M^2} \leq 2e^{-2(n_- - 1)\epsilon^2/(1+\gamma)^2 M^2}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbf{P}\left\{\left|\widehat{h}_i - h(\mathbf{X}_i)\right| > \epsilon\right\} &= \sum_{\mathbf{x}} \mathbf{P}\left\{\mathbf{X}_i = \mathbf{x}\right\} \cdot \mathbf{P}\left\{\left|\widehat{h}_i - h(\mathbf{X}_i)\right| > \epsilon \,\middle|\, \mathbf{X}_i = \mathbf{x}\right\} \\
&\leq \sum_{\mathbf{x}} \mathbf{P}\left\{\mathbf{X}_i = \mathbf{x}\right\}\left(2e^{-2(n_+ - 1)\epsilon^2/(1+\gamma)^2 M^2} + 2e^{-2(n_- - 1)\epsilon^2/(1+\gamma)^2 M^2}\right) \\
&= 2e^{-2(n_+ - 1)\epsilon^2/(1+\gamma)^2 M^2} + 2e^{-2(n_- - 1)\epsilon^2/(1+\gamma)^2 M^2}.
\end{aligned}$$

In a similar way, it can be shown that for $i \in I_-$,

$$\mathbf{P}\left\{\left|\widehat{h}_i - h(\mathbf{X}_i)\right| > \epsilon\right\} \leq 2e^{-2(n_+ - 1)\epsilon^2/(1+\gamma)^2 M^2} + 2e^{-2(n_- - 1)\epsilon^2/(1+\gamma)^2 M^2}.$$

Then,

$$
\begin{aligned}
&\mathbf{P}\left\{\sup_{\boldsymbol{\alpha}\in A}\left|H_n\left(\boldsymbol{\alpha}\right)-H\left(\boldsymbol{\alpha}\right)\right|>\epsilon\right\}=\mathbf{P}\left\{\sup_{\boldsymbol{\alpha}\in A}\left|\sum_{i=1}^{n}\alpha_i Y_i\left(\widehat{h}_i-h\left(\mathbf{X}_i\right)\right)\right|>\epsilon\right\}\\
\leq\ & \mathbf{P}\left\{\sup_{\boldsymbol{\alpha}\in A}\sum_{i=1}^{n}\alpha_i\left|Y_i\right|\left|\widehat{h}_i-h\left(\mathbf{X}_i\right)\right|>\epsilon\right\}\\
=\ & \mathbf{P}\left\{\sup_{\boldsymbol{\alpha}\in A}\sum_{i\in I_+}^{n}\alpha_i\left|\widehat{h}_i-h\left(\mathbf{X}_i\right)\right|+\sum_{i\in I_-}^{n}\alpha_i\gamma\left|\widehat{h}_i-h\left(\mathbf{X}_i\right)\right|>\epsilon\right\}\\
\leq\ & \mathbf{P}\left\{\sup_{\boldsymbol{\alpha}\in A}\sum_{i\in I_+}^{n}\alpha_i\left|\widehat{h}_i-h\left(\mathbf{X}_i\right)\right|>\frac{\epsilon}{1+\gamma}\right\}+\mathbf{P}\left\{\sup_{\boldsymbol{\alpha}\in A}\sum_{i\in I_-}^{n}\alpha_i\gamma\left|\widehat{h}_i-h\left(\mathbf{X}_i\right)\right|>\frac{\gamma\epsilon}{1+\gamma}\right\}\\
=\ & \mathbf{P}\left\{\max_{i\in I_+}\left|\widehat{h}_i-h\left(\mathbf{X}_i\right)\right|>\frac{\epsilon}{1+\gamma}\right\}+\mathbf{P}\left\{\max_{i\in I_-}\left|\widehat{h}_i-h\left(\mathbf{X}_i\right)\right|>\frac{\epsilon}{1+\gamma}\right\}\\
=\ & \mathbf{P}\left\{\bigcup_{i\in I_+}\left\{\left|\widehat{h}_i-h\left(\mathbf{X}_i\right)\right|>\frac{\epsilon}{1+\gamma}\right\}\right\}+\mathbf{P}\left\{\bigcup_{i\in I_-}\left\{\left|\widehat{h}_i-h\left(\mathbf{X}_i\right)\right|>\frac{\epsilon}{1+\gamma}\right\}\right\}\\
\leq\ & \sum_{i\in I_+}\mathbf{P}\left\{\left|\widehat{h}_i-h\left(\mathbf{X}_i\right)\right|>\frac{\epsilon}{1+\gamma}\right\}+\sum_{i\in I_-}\mathbf{P}\left\{\left|\widehat{h}_i-h\left(\mathbf{X}_i\right)\right|>\frac{\epsilon}{1+\gamma}\right\}\\
\leq\ & n_+\left(2e^{-2(n_+-1)\epsilon^2/(1+\gamma)^4M^2}+2e^{-2(n_--1)\epsilon^2/(1+\gamma)^4M^2}\right)\\
& +n_-\left(2e^{-2(n_+-1)\epsilon^2/(1+\gamma)^4M^2}+2e^{-2(n_--1)\epsilon^2/(1+\gamma)^4M^2}\right)\\
=\ & n\left(2e^{-2(n_+-1)\epsilon^2/(1+\gamma)^4M^2}+2e^{-2(n_--1)\epsilon^2/(1+\gamma)^4M^2}\right).
\end{aligned}
$$

APPENDIX C
PROOF OF THEOREM 2

Define $\mathbf{u}=(u_1,\ldots,u_n)$ such that $u_i=1/n_+$ for $i\in I_+$ and $u_i=1/n_-$ for $i\in I_-$. By the similar argument for the convergence of MISE of kernel density estimate [4], it can be shown, using a multivariate Taylor series, that

$$
\begin{aligned}
&MISE\left(\mathbf{u};n_+,n_-\right)=\mathbf{E}\left[ISE\left(\mathbf{u}\right)\right]\\
&=\int Var\left(\widehat{d}_\gamma\left(\mathbf{x};\mathbf{u}\right)\right)+bias^2\left(\widehat{d}_\gamma\left(\mathbf{x};\mathbf{u}\right)\right)d\mathbf{x}\\
&=\left\{\frac{1}{n_+\sigma^d}+\frac{\gamma^2}{n_-\sigma^d}\right\}R\left(k\right)+\frac{1}{4}\sigma^4R\left(tr\left\{\mathcal{H}_{d_\gamma}\right\}\right)+o\left(n_+^{-1}\sigma^{-d}+n_-^{-1}\sigma^{-d}+\sigma^4\right)
\end{aligned}
$$

where $R\left(f\right)=\int f^2\left(\mathbf{x}\right)d\mathbf{x}$ and $\mathcal{H}_f$ represent the Hessian matrix of f. Therefore, $ISE\left(\mathbf{u}\right)$ converges to 0 in probability since $\sigma\to0$, $n_+\sigma^d\to\infty$ and $n_+\sigma^d\to\infty$ as $n\to\infty$. Furthermore,

$$
\begin{aligned}
\mathbf{P}\left\{ISE\left(\widehat{\boldsymbol{\alpha}}\right)>\epsilon\right\}\ &=\ \mathbf{P}\left\{ISE\left(\widehat{\boldsymbol{\alpha}}\right)>\epsilon,\ ISE\left(\mathbf{u}\right)>\frac{\epsilon}{2}\right\}+\mathbf{P}\left\{ISE\left(\widehat{\boldsymbol{\alpha}}\right)>\epsilon,\ ISE\left(\mathbf{u}\right)\leq\frac{\epsilon}{2}\right\}\\
&\leq\ \mathbf{P}\left\{ISE\left(\mathbf{u}\right)>\frac{\epsilon}{2}\right\}+\mathbf{P}\left\{ISE\left(\widehat{\boldsymbol{\alpha}}\right)>ISE\left(\mathbf{u}\right)+\frac{\epsilon}{2}\right\}.
\end{aligned}
$$

From the consistency of $ISE\left(\mathbf{u}\right)$ and the oracle inequality stated in Theorem 1, $ISE\left(\widehat{\boldsymbol{\alpha}}\right)$ converges to 0 in probability.

APPENDIX D
PROOF OF THEOREM 3

First note that in the previous analyses we treat N_+,N_- and γ as deterministic variables but now we turn to the case where these variables are random. Thus, some of the previous results should be restated considering this.

Lemma 2: γ converges to γ^* with probability 1.

Proof: Note that N_+ and N_- are binomial random variables with (n,p) and (n,q) where $q=1-p$. From the Hoeffding's inequality, we know that for $\forall\epsilon>0$

$$
\begin{aligned}
\mathbf{P}\left\{\frac{N_+}{n}-p>\epsilon\right\}\leq e^{-2n\epsilon^2},\quad \mathbf{P}\left\{\frac{N_+}{n}-p<-\epsilon\right\}\leq e^{-2n\epsilon^2}\\
\mathbf{P}\left\{\frac{N_-}{n}-q>\epsilon\right\}\leq e^{-2n\epsilon^2},\quad \mathbf{P}\left\{\frac{N_-}{n}-q<-\epsilon\right\}\leq e^{-2n\epsilon^2}.
\end{aligned}
$$

Then, for any $\epsilon > 0$

$$\begin{aligned}
&\mathbf{P}_n(\epsilon) \triangleq \mathbf{P}\left\{\left|\frac{N_-}{N_+} - \frac{q}{p}\right| > \epsilon\right\} = \mathbf{P}\left\{|pN_- - qN_+| > \epsilon p N_+\right\} \\
&= \mathbf{P}\left\{|pN_- - qN_+| > \epsilon p N_+, N_+ \geq \frac{np}{2}\right\} + \mathbf{P}\left\{|pN_- - qN_+| > \epsilon p N_+, N_+ < \frac{np}{2}\right\} \\
&\leq \mathbf{P}\left\{|pN_- - qN_+| > \epsilon p \cdot \frac{np}{2}\right\} + \mathbf{P}\left\{N_+ < \frac{np}{2}\right\} \\
&\leq \mathbf{P}\left\{|pN_- - pqn + pqn - qN_+| > \frac{n\epsilon p^2}{2}\right\} + \mathbf{P}\left\{N_+ - pn < -\frac{np}{2}\right\} \\
&\leq \mathbf{P}\left\{|pN_- - pqn| > \frac{n\epsilon p^3}{2}\right\} + \mathbf{P}\left\{|qN_+ - pqn| > \frac{n\epsilon p^2 q}{2}\right\} + \mathbf{P}\left\{N_+ - pn < -\frac{np}{2}\right\} \\
&= \mathbf{P}\left\{\left|\frac{N_-}{n} - q\right| > \frac{\epsilon p^2}{2}\right\} + \mathbf{P}\left\{\left|\frac{N_+}{n} - p\right| > \frac{\epsilon p^2}{2}\right\} + \mathbf{P}\left\{\frac{N_+}{n} - p < -\frac{p}{2}\right\} \\
&\leq 4\exp\left(-\frac{n\epsilon^2 p^4}{2}\right) + \exp\left(-\frac{np^2}{2}\right).
\end{aligned}$$

Since $\sum_{n=1}^{\infty} \mathbf{P}_n(\epsilon) < \infty$ for all $\epsilon > 0$, γ converges to γ^* with probability 1. $\square$

Lemma 3: Suppose the assumptions in Theorem 3 are satisfied. For any $\epsilon' > 0$, $\mathbf{P}\{ISE(\widehat{\boldsymbol{\alpha}}) > \inf_{\boldsymbol{\alpha}\in A} ISE(\boldsymbol{\alpha}) + \epsilon'\}$ converges to 0.

Proof: We need to restate Theorem 1 as follows. For any $\delta > 0$,

$$\mathbf{P}\left\{ISE(\widehat{\boldsymbol{\alpha}}) > \inf_{\boldsymbol{\alpha}\in A} ISE(\boldsymbol{\alpha}) + 4\sqrt{\frac{\ln(2n/\delta)}{c[\min(N_+, N_-) - 1]}} \,\middle|\, N_+ = n_+, N_- = n_-\right\} \leq \delta$$

since

$$\sqrt{\frac{\ln(2n/\delta)}{c[\min(n_+, n_-) - 1]}} \leq \epsilon \leq \sqrt{\frac{\ln(2n/\delta)}{c[\max(n_+, n_-) - 1]}}.$$

Let us define $c' = 2\left(\sqrt{2\pi}\sigma\right)^{2d} / (1 + 2\gamma^*)^4$ and an event $D = \left\{N_+ \geq \frac{np}{2}, N_- \geq \frac{n(1-p)}{2}, \gamma \leq 2\gamma^*\right\}$. Then,

$$\begin{aligned}
&\mathbf{P}\left\{ISE(\widehat{\boldsymbol{\alpha}}) > \inf_{\boldsymbol{\alpha}\in A} ISE(\boldsymbol{\alpha}) + 4\sqrt{\frac{2\ln(2n/\delta)}{c'[\min(np, n(1-p)) - 1]}}\right\} \\
&\leq \mathbf{P}\{D^c\} + \mathbf{P}\{D\} \cdot \mathbf{P}\left\{ISE(\widehat{\boldsymbol{\alpha}}) > \inf_{\boldsymbol{\alpha}\in A} ISE(\boldsymbol{\alpha}) + 4\sqrt{\frac{2\ln(2n/\delta)}{c'[\min(np, n(1-p)) - 1]}} \,\middle|\, D\right\}.
\end{aligned}$$

The first term converges to 0 from the strong law of large numbers and Lemma 2. The second term becomes

$$\begin{aligned}
&\mathbf{P}\left\{ISE(\widehat{\boldsymbol{\alpha}}) > \inf_{\boldsymbol{\alpha}\in A} ISE(\boldsymbol{\alpha}) + 4\sqrt{\frac{2\ln(2n/\delta)}{c'[\min(np, n(1-p)) - 1]}} \,\middle|\, D\right\} \\
&\leq \mathbf{P}\left\{ISE(\widehat{\boldsymbol{\alpha}}) > \inf_{\boldsymbol{\alpha}\in A} ISE(\boldsymbol{\alpha}) + 4\sqrt{\frac{\ln(2n/\delta)}{c[\min(N_+, N_-) - 1]}} \,\middle|\, D\right\} \\
&= \sum \mathbf{P}\left\{ISE(\widehat{\boldsymbol{\alpha}}) > \inf_{\boldsymbol{\alpha}\in A} ISE(\boldsymbol{\alpha}) + 4\sqrt{\frac{\ln(2n/\delta)}{c[\min(N_+, N_-) - 1]}} \,\middle|\, D, N_+ = n_+, N_- = n_-\right\} \\
&\quad \cdot \mathbf{P}\{N_+ = n_+, N_- = n_-\} \\
&\leq \sum \delta \mathbf{P}\{N_+ = n_+, N_- = n_-\} = \delta.
\end{aligned}$$

For any $\delta > 0$, we can make $4\sqrt{\frac{2\ln(2n/\delta)}{c'[\min(np, n(1-p)) - 1]}}$ smaller than ϵ' as $n \to \infty$, provided that $\ln n / n\sigma^d \to 0$ as $n \to 0$. Therefore, $\mathbf{P}\{ISE(\widehat{\boldsymbol{\alpha}}) > \inf_{\boldsymbol{\alpha}\in A} ISE(\boldsymbol{\alpha}) + \epsilon'\}$ converges to 0. $\square$

Lemma 4: Suppose the assumptions in Theorem 3 are satisfied. Then, $ISE(\mathbf{u})$ converges to 0 in probability.

Proof: Define an event $D = \left\{N_+ \geq \frac{np}{2}, N_- \geq \frac{n(1-p)}{2}, \gamma \leq 2\gamma^*\right\}$. For any $\epsilon > 0$,

$$\mathbf{P}\{ISE(\mathbf{u}) > \epsilon\} \leq \mathbf{P}\{D^c\} + \mathbf{P}\{ISE(\mathbf{u}) > \epsilon,\ D\}.$$

The first term converges to 0 from the strong law of large numbers and Lemma 2. Let define a set $S = \left\{(n_+, n_-) \,\middle|\, n_+ \geq \frac{np}{2}, n_- \geq \frac{n(1-p)}{2}, \frac{n_-}{n_+} \leq 2\gamma^*\right\}$. Then,

$$\begin{aligned}
&\mathbf{P}\left\{ISE\left(\mathbf{u}\right) > \epsilon,\, D\right\} \\
&= \sum \mathbf{P}\left\{ISE\left(\mathbf{u}\right) > \epsilon,\, D \,\middle|\, N_+ = n_+, N_- = n_-\right\} \cdot \mathbf{P}\left\{N_+ = n_+, N_- = n_-\right\} \\
&= \sum_{(n_+,n_-)\in S} \mathbf{P}\left\{ISE\left(\mathbf{u}\right) > \epsilon \,\middle|\, N_+ = n_+, N_- = n_-\right\} \cdot \mathbf{P}\left\{N_+ = n_+, N_- = n_-\right\} \\
&\leq \sum_{(n_+,n_-)\in S} \frac{\mathbf{E}\left[ISE\left(\mathbf{u}\right) \,\middle|\, N_+ = n_+, N_- = n_-\right]}{\epsilon} \cdot \mathbf{P}\left\{N_+ = n_+, N_- = n_-\right\} \\
&\leq \frac{1}{\epsilon} \sum_{(n_+,n_-)\in S} \left[\frac{1}{n\sigma^d}\left(\frac{2}{p} + \frac{8{\gamma^*}^2}{1-p}\right) R\left(k\right) + \frac{1}{4}\sigma^4 R\left(tr\left\{\mathcal{H}_{d_\gamma}\right\}\right) + o\left(n^{-1}\sigma^{-d} + \sigma^4\right)\right] \\
&\quad \cdot \mathbf{P}\left\{N_+ = n_+, N_- = n_-\right\} \\
&\leq \frac{1}{\epsilon}\left(\frac{1}{n\sigma^d}\left\{\frac{2}{p} + \frac{2{\gamma^*}^2}{1-p}\right\} R\left(k\right) + \frac{1}{4}\sigma^4 R\left(tr\left\{\mathcal{H}_{d_\gamma}\right\}\right) + o\left(n^{-1}\sigma^{-d} + \sigma^4\right)\right)
\end{aligned}$$

where the second to the last step, we used $MISE\left(\mathbf{u}; n_+, n_-\right)$ formula in explained in Appendix C and the fact that for $(n_+, n_-) \in S$,

$$\frac{1}{n_+\sigma^d} + \frac{1}{n_-\sigma^d} \leq \frac{2}{np\sigma^d} + \frac{2}{n(1-p)\sigma^d} = \frac{1}{n\sigma^d}\left(\frac{2}{p} + \frac{2}{1-p}\right)$$

Therefore, $ISE\left(\mathbf{u}\right)$ converges to 0 since $\sigma \to 0$ and $n\sigma^d \to \infty$ as $n \to \infty$. □

Now let's prove Theorem 3. From Theorem 3 in [5], it suffices to show that

$$\int \left(\widehat{d}_\gamma\left(\mathbf{x}; \widehat{\boldsymbol{\alpha}}\right) - d_{\gamma^*}\left(\mathbf{x}\right)\right)^2 d\mathbf{x} \to 0$$

in probability. Note that

$$\begin{aligned}
\|\widehat{d}_\gamma\left(\mathbf{x}; \widehat{\boldsymbol{\alpha}}\right) - d_{\gamma^*}\left(\mathbf{x}\right)\|_{L^2} &= \|\widehat{d}_\gamma\left(\mathbf{x}; \widehat{\boldsymbol{\alpha}}\right) - d_\gamma\left(\mathbf{x}\right) + \left(\gamma - \gamma^*\right) f_-\left(\mathbf{x}\right)\|_{L^2} \\
&\leq \|\widehat{d}_\gamma\left(\mathbf{x}; \widehat{\boldsymbol{\alpha}}\right) - d_\gamma\left(\mathbf{x}\right)\|_{L_2} + \|\left(\gamma - \gamma^*\right) f_-\left(\mathbf{x}\right)\|_{L^2} \\
&= \sqrt{ISE\left(\widehat{\boldsymbol{\alpha}}\right)} + |\gamma - \gamma^*| \cdot \|f_-\left(\mathbf{x}\right)\|_{L^2}.
\end{aligned} \tag{18}$$

For the first term in (18), $\mathbf{P}\left\{ISE\left(\widehat{\boldsymbol{\alpha}}\right) > \epsilon\right\}$ converges to 0 in probability since

$$\mathbf{P}\left\{SE\left(\widehat{\boldsymbol{\alpha}}\right) > \epsilon\right\} \leq \mathbf{P}\left\{ISE\left(\widehat{\boldsymbol{\alpha}}\right) > ISE\left(\mathbf{u}\right) + \frac{\epsilon}{2}\right\} + \mathbf{P}\left\{ISE\left(\mathbf{u}\right) > \frac{\epsilon}{2}\right\}$$

and from Lemma 3 and 4, . The second term in (18) also converges to 0 in probability from Lemma 2. This proves the theorem.

REFERENCES

[1] John C.Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," *Technical Report MSR-TR-98-14*, April 2001.
[2] Mark Girolami and Chao He, "Probability density estimation from optimally condensed data samples," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, pp. 1253–1264, OCT 2003.
[3] L. Devroye and G. Lugosi, "Combinatorial methods in density estimation," 2001.
[4] D. W. Scott, *Multivariate Density Estimation*, Wiley, New York, 1992.
[5] Charles T. Wolverton and Terry J. Wagner, "Asymptotically optimal discriminant fucntions for pattern classification," *IEEE Trans. Info. Theory*, vol. 15, no. 2, pp. 258–265, Mar 1969.