Biometrika, *pp*. 1–13
Printed in Great Britain

Supplementary material for Classification with imperfect training labels

BY TIMOTHY I. CANNINGS

School of Mathematics, University of Edinburgh,
James Clerk Maxwell Building, Edinburgh EH9 3FD, U.K.
timothy.cannings@ed.ac.uk

YINGYING FAN

Department of Data Science and Operations, University of Southern California,
Los Angeles, California 90089, U.S.A.
fanyingy@marshall.usc.edu

AND RICHARD J. SAMWORTH

Statistical Laboratory, University of Cambridge,
Centre for Mathematical Sciences, Cambridge CB3 0WB, U.K.
r.samworth@statslab.cam.ac.uk

SUMMARY

We present the proofs of the theoretical results in 'Classification with imperfect training labels', as well as an illustrative example involving the 1-nn classifier.

A. PROOFS

A·1. *Proofs from Section* 3

Proof of Theorem 1. (i) First, we have that for P_X-almost all $x \in \mathcal{B}$,

$$\begin{aligned}\tilde{\eta}(x) - 1/2 &= \{\eta(x) - 1/2\}\{1 - \rho_0(x) - \rho_1(x)\} + \frac{1}{2}\{\rho_0(x) - \rho_1(x)\} \\ &= \{\eta(x) - 1/2\}\{1 - \rho_0(x) - \rho_1(x)\}\Big(1 - \frac{\rho_1(x) - \rho_0(x)}{\{2\eta(x) - 1\}\{1 - \rho_0(x) - \rho_1(x)\}}\Big). \qquad \text{(A1)}\end{aligned}$$

Thus, for P_X-almost all $x \in \mathcal{B}$, we have $\{\rho_1(x) - \rho_0(x)\}/[\{2\eta(x) - 1\}\{1 - \rho_0(x) - \rho_1(x)\}] < 1$ if and only if

$$\operatorname{sgn}\{\tilde{\eta}(x) - 1/2\} = \operatorname{sgn}\{\eta(x) - 1/2\}.$$

This completes the proof of (3). It follows that, if $P_X(\mathcal{A}^c \cap \mathcal{S}^c) = 0$, then $P_X(\{x \in \mathcal{B} : \tilde{C}^{\mathrm{Bayes}}(x) = C^{\mathrm{Bayes}}(x)\}^c \cap \mathcal{S}^c) = 0$. In other words $P_X(\{x \in \mathcal{S}^c : \tilde{C}^{\mathrm{Bayes}}(x) \neq C^{\mathrm{Bayes}}(x)\}) = 0$, i.e. (2) holds. Here we have used the fact that $\mathcal{A} \subseteq \mathcal{B}$, so if $P_X(\mathcal{A}^c \cap \mathcal{S}^c) = 0$, then $P_X(\mathcal{B}^c \cap \mathcal{S}^c) = 0$.

(ii) For the proof of this part, we apply Proposition 1. First, since (2) holds, we have $\tilde{R}(C^{\mathrm{Bayes}}) = \tilde{R}(\tilde{C}^{\mathrm{Bayes}})$. From (A1), we have that for P_X-almost all $x \in \mathcal{B}$,

$$\begin{aligned}|2\tilde{\eta}(x) - 1| &= |2\eta(x) - 1|\{1 - \rho_0(x) - \rho_1(x)\}\Big(1 - \frac{\rho_1(x) - \rho_0(x)}{\{2\eta(x) - 1\}\{1 - \rho_0(x) - \rho_1(x)\}}\Big) \\ &\geq |2\eta(x) - 1|(1 - 2\rho^*)(1 - a^*). \qquad \text{(A2)}\end{aligned}$$

In fact, the conclusion of (A2) remains true trivially when $x \in \mathcal{S}$. Thus, by Proposition 1,

$$\begin{aligned} R(C) - R(C^{\mathrm{Bayes}}) &\leq \inf_{\kappa>0}\Big\{\kappa\{\tilde{R}(C) - \tilde{R}(\tilde{C}^{\mathrm{Bayes}})\} + P_X(A_\kappa^c)\Big\} \\ &\leq \frac{\tilde{R}(C) - \tilde{R}(\tilde{C}^{\mathrm{Bayes}})}{(1-2\rho^*)(1-a^*)} + P_X(A^c_{(1-2\rho^*)^{-1}(1-a^*)^{-1}}) = \frac{\tilde{R}(C) - \tilde{R}(\tilde{C}^{\mathrm{Bayes}})}{(1-2\rho^*)(1-a^*)}, \end{aligned}$$

since $P_X(A^c_{(1-2\rho^*)^{-1}(1-a^*)^{-1}}) \leq P_X(A^c_{(1-2\rho^*)^{-1}(1-a^*)^{-1}} \cap \mathcal{B}) + P_X(\mathcal{B}^c) = 0$, by (A2). □

Proposition 1 is a special case of the following result.

PROPOSITION A1. *Let* $\mathcal{D} = \big\{x \in \mathcal{S}^c : \tilde{C}^{\mathrm{Bayes}}(x) = C^{\mathrm{Bayes}}(x)\big\}$*, and recall the definition of* A_κ *in Proposition* 1*. Then, for any classifier* C*,*

$$\begin{aligned} R(C) - R(C^{\mathrm{Bayes}}) \leq R(\tilde{C}^{\mathrm{Bayes}}) - R(C^{\mathrm{Bayes}}) + \min\Big[&\mathrm{pr}\big\{\{C(X) \neq \tilde{C}^{\mathrm{Bayes}}(X)\} \cap \{X \in \mathcal{D}\}\big\}, \\ &\inf_{\kappa>0}\big\{\kappa\{\tilde{R}(C) - \tilde{R}(\tilde{C}^{\mathrm{Bayes}})\} + E\big(|2\eta(X)-1|\mathbb{1}_{\{X\in\mathcal{D}\setminus A_\kappa\}}\big)\big\}\Big]. \end{aligned}$$

Remark: If (2) holds, i.e. $P_X(\mathcal{D}^c \cap \mathcal{S}^c) = 0$, then $R(\tilde{C}^{\mathrm{Bayes}}) = R(C^{\mathrm{Bayes}})$, and moreover we have that $E\big(|2\eta(X)-1|\mathbb{1}_{\{X\in\mathcal{D}\setminus A_\kappa\}}\big) \leq P_X\big(\mathcal{D}\setminus A_\kappa\big) \leq P_X\big(A_\kappa^c\big)$.

Proof of Proposition A1. First write

$$\begin{aligned} R(C) &= \int_{\mathcal{X}} \mathrm{pr}\{C(x) \neq Y \mid X = x\}\, dP_X(x) \\ &= \int_{\mathcal{X}} \big[\mathrm{pr}\{C(x)=0\}\mathrm{pr}(Y=1 \mid X=x) + \mathrm{pr}\{C(x)=1\}\mathrm{pr}(Y=0 \mid X=x)\big]\, dP_X(x) \\ &= \int_{\mathcal{X}} \big[\mathrm{pr}\{C(x)=0\}\{2\eta(x)-1\} + \{1-\eta(x)\}\big]\, dP_X(x). \end{aligned} \tag{A3}$$

Here we have implicitly assumed that the classifier C is random since it may depend on random training data. However, in the case that C is non-random, one should interpret $\mathrm{pr}\{C(x)=0\}$ as being equal to $\mathbb{1}_{\{C(x)=0\}}$, for $x \in \mathcal{X}$.

Now, for P_X-almost all $x \in \mathcal{D}$,

$$\begin{aligned} \big[\mathrm{pr}\{C(x)=0\} - \mathbb{1}_{\{\tilde{\eta}(x)<1/2\}}\big]\{2\eta(x)-1\} &= \big|\mathrm{pr}\{C(x)=0\} - \mathbb{1}_{\{\tilde{\eta}(x)<1/2\}}\big||2\eta(x)-1| \\ &\leq \big|\mathrm{pr}\{C(x)=0\} - \mathbb{1}_{\{\tilde{\eta}(x)<1/2\}}\big| \\ &= \mathrm{pr}\{C(x) \neq \tilde{C}^{\mathrm{Bayes}}(x)\}. \end{aligned}$$

Moreover, for P_X-almost all $x \in \mathcal{D}^c$, we have

$$\big[\mathrm{pr}\{C(x)=0\} - \mathbb{1}_{\{\tilde{\eta}(x)<1/2\}}\big]\{2\eta(x)-1\} \leq 0 \tag{A4}$$

It follows that

$$\begin{aligned} R(C) - R(\tilde{C}^{\mathrm{Bayes}}) &= \int_{\mathcal{X}} \big[\mathrm{pr}\{C(x)=0\} - \mathbb{1}_{\{\tilde{\eta}(x)<1/2\}}\big]\{2\eta(x)-1\}\, dP_X(x) \\ &= \int_{\mathcal{D}} \big[\mathrm{pr}\{C(x)=0\} - \mathbb{1}_{\{\tilde{\eta}(x)<1/2\}}\big]\{2\eta(x)-1\}\, dP_X(x) \\ &\qquad + \int_{\mathcal{D}^c} \big[\mathrm{pr}\{C(x)=0\} - \mathbb{1}_{\{\tilde{\eta}(x)<1/2\}}\big]\{2\eta(x)-1\}\, dP_X(x) \\ &\leq \mathrm{pr}\big(\{C(X) \neq \tilde{C}^{\mathrm{Bayes}}(X)\} \cap \{X \in \mathcal{D}\}\big). \end{aligned}$$

To see the right-hand bound, observe that by (A4), for $\kappa > 0$,

$$\begin{aligned} R(C) - R(\tilde{C}^{\text{Bayes}}) &= \int_{\mathcal{X}} \big[\text{pr}\{C(x) = 0\} - \mathbb{1}_{\{\tilde{\eta}(x)<1/2\}}\big]\{2\eta(x) - 1\}\, dP_X(x) \\ &\le \int_{\mathcal{D}} \big[\text{pr}\{C(x) = 0\} - \mathbb{1}_{\{\tilde{\eta}(x)<1/2\}}\big]\{2\eta(x) - 1\}\, dP_X(x) \\ &\le \kappa \int_{\mathcal{D}\cap A_\kappa} \big[\text{pr}\{C(x) = 0\} - \mathbb{1}_{\{\tilde{\eta}(x)<1/2\}}\big]\{2\tilde{\eta}(x) - 1\}\, dP_X(x) \\ &\qquad + E\big(|2\eta(X) - 1|\mathbb{1}_{\{X\in\mathcal{D}\setminus A_\kappa\}}\big) \\ &= \kappa\{\tilde{R}(C) - \tilde{R}(\tilde{C}^{\text{Bayes}})\} + E\big(|2\eta(X) - 1|\mathbb{1}_{\{X\in\mathcal{D}\setminus A_\kappa\}}\big), \end{aligned}$$

where the last step follows from (A3). □

Example A1. Suppose that $\mathcal{X} \subseteq \mathbb{R}^d$ and that the noise is ρ-homogeneous with $\rho \in (0, 1/2)$. Consider the 1-nearest neighbour classifier $\tilde{C}^{\text{1nn}}(x) = \tilde{Y}_{(1)}$, where $(X_{(1)}, \tilde{Y}_{(1)}) = (X_{(1)}(x), \tilde{Y}_{(1)}(x)) = (X_{i^*}, \tilde{Y}_{i^*})$ is the training data pair for which $i^* = \text{sargmin}_{i=1,\ldots,n} \|X_i - x\|$, where sargmin denotes the smallest index of the set of minimizers. We first study the first term in the minimum in (4). Noting that $\tilde{R}(\tilde{C}^{\text{Bayes}}) = E[\min\{\tilde{\eta}(X), 1 - \tilde{\eta}(X)\}]$, we have

$$\begin{aligned} &\big|\text{pr}\{\tilde{C}^{\text{1nn}}(X) \neq \tilde{C}^{\text{Bayes}}(X)\} - \tilde{R}(\tilde{C}^{\text{Bayes}})\big| \\ &\quad = \big|\text{pr}\{\tilde{Y}_{(1)}(X) \neq \tilde{C}^{\text{Bayes}}(X)\} - \tilde{R}(\tilde{C}^{\text{Bayes}})\big| \\ &\quad = \big|E[\mathbb{1}_{\{\tilde{\eta}(X)<1/2\}}\tilde{\eta}(X_{(1)}(X)) + \mathbb{1}_{\{\tilde{\eta}(X)\ge1/2\}}\{1 - \tilde{\eta}(X_{(1)}(X))\}] - \tilde{R}(\tilde{C}^{\text{Bayes}})\big| \\ &\quad = \big|E[\mathbb{1}_{\{\tilde{\eta}(X)<1/2\}}\{\tilde{\eta}(X_{(1)}(X)) - \tilde{\eta}(X)\} + \mathbb{1}_{\{\tilde{\eta}(X)\ge1/2\}}\{\tilde{\eta}(X) - \tilde{\eta}(X_{(1)}(X))\}]\big| \\ &\quad \le E\big|\tilde{\eta}(X_{(1)}(X)) - \tilde{\eta}(X)\big| \to 0, \end{aligned} \tag{A5}$$

where the final limit follows by Devroye et al. (1996, Lemma 5.4).

Now focusing on the second term in the minimum in (4), by Devroye et al. (1996, Theorem 5.1), we have

$$\tilde{R}(\tilde{C}^{\text{1nn}}) - \tilde{R}(\tilde{C}^{\text{Bayes}}) \to 2E[\tilde{\eta}(X)\{1 - \tilde{\eta}(X)\}] - \tilde{R}(\tilde{C}^{\text{Bayes}}).$$

Moreover, in this case, $P_X(A_\kappa^c) = 1$ for all $\kappa \le (1 - 2\rho)^{-1}$, and 0 otherwise. Therefore, if ρ is small enough that $\rho\tilde{R}(\tilde{C}^{\text{Bayes}}) < \tilde{R}(\tilde{C}^{\text{Bayes}}) - E[\tilde{\eta}(X)\{1 - \tilde{\eta}(X)\}]$, then

$$\begin{aligned} \lim_{n\to\infty} \inf_{\kappa>0} \Big\{\kappa\{\tilde{R}(\tilde{C}^{\text{1nn}}) - \tilde{R}(\tilde{C}^{\text{Bayes}})\} + P_X(A_\kappa^c)\Big\} &= \lim_{n\to\infty} \frac{\tilde{R}(\tilde{C}^{\text{1nn}}) - \tilde{R}(\tilde{C}^{\text{Bayes}})}{1 - 2\rho} \\ &= \frac{2E[\tilde{\eta}(X)\{1 - \tilde{\eta}(X)\}] - \tilde{R}(\tilde{C}^{\text{Bayes}})}{1 - 2\rho} \\ &< \tilde{R}(\tilde{C}^{\text{Bayes}}) = \lim_{n\to\infty} \text{pr}\{\tilde{C}^{\text{1nn}}(X) \neq \tilde{C}^{\text{Bayes}}(X)\}, \end{aligned} \tag{A6}$$

where the final equality is due to (A5). Thus, in this case, the second term in the minimum in (4) is smaller for sufficiently large n. However, if $\rho\tilde{R}(\tilde{C}^{\text{Bayes}}) > \tilde{R}(\tilde{C}^{\text{Bayes}}) - E[\tilde{\eta}(X)\{1 - \tilde{\eta}(X)\}]$, the asymptotically better bound is given by the first term in the minimum in the conclusion of Proposition 1, because then the inequality in (A6) is reversed. □

Proof of Corollary 1. Let $\epsilon_n = \max\big[\sup_{m\geq n}\{\tilde{R}(\tilde{C}_m) - \tilde{R}(\tilde{C}^{\mathrm{Bayes}})\}^{1/2}, n^{-1}\big]$. Then, by Proposition A1,

$$\begin{aligned} R(\tilde{C}_n) - R(\tilde{C}^{\mathrm{Bayes}}) &\leq \frac{1}{\epsilon_n}\{\tilde{R}(\tilde{C}_n) - \tilde{R}(\tilde{C}^{\mathrm{Bayes}})\} + E\big(|2\eta(X)-1|\mathbb{1}_{\{X\in\mathcal{D}\setminus A_{\epsilon_n^{-1}}\}}\big) \\ &\leq \{\tilde{R}(\tilde{C}_n) - \tilde{R}(\tilde{C}^{\mathrm{Bayes}})\}^{1/2} + P_X(\mathcal{D}\setminus A_{\epsilon_n^{-1}}). \end{aligned}$$

Since (ϵ_n) is decreasing, it follows that

$$\limsup_{n\to\infty} R(\tilde{C}_n) - R(C^{\mathrm{Bayes}}) \leq R(\tilde{C}^{\mathrm{Bayes}}) - R(C^{\mathrm{Bayes}}) + P_X(\tilde{\mathcal{S}}\cap\mathcal{D}).$$

In particular, if (2) holds, then

$$\limsup_{n\to\infty} R(\tilde{C}_n) - R(C^{\mathrm{Bayes}}) \leq P_X(\tilde{\mathcal{S}}\setminus\mathcal{S}),$$

as required. □

A·2. *Conditions and proof of Theorem* 2

A formal description of the conditions of Theorem 2 is given below:

Assumption A1. The probability measures P_0 and P_1 are absolutely continuous with respect to Lebesgue measure, with Radon–Nikodym derivatives f_0 and f_1, respectively. Moreover, the marginal density of X, given by $\bar{f} = \pi_0 f_0 + \pi_1 f_1$, is continuous and positive.

Assumption A2. The set $\mathcal{S}$ is non-empty and $\bar{f}$ is bounded on $\mathcal{S}$. There exists $\epsilon_0 > 0$ such that $\bar{f}$ is twice continuously differentiable on $\mathcal{S}^{\epsilon_0} = \mathcal{S} + B_{\epsilon_0}(0)$, and

$$F(\delta) = \sup_{x_0\in\mathcal{S}:\bar{f}(x_0)\geq\delta} \max\left\{\frac{\|\dot{\bar{f}}(x_0)\|}{\bar{f}(x_0)}, \frac{\sup_{u\in B_{\epsilon_0}(0)}\|\ddot{\bar{f}}(x_0+u)\|_{\mathrm{op}}}{\bar{f}(x_0)}\right\} = o(\delta^{-\tau}) \tag{A7}$$

as $\delta \searrow 0$, for every $\tau > 0$. Furthermore, recalling $a_d = \pi^{d/2}/\Gamma(1+d/2)$ and writing $p_\epsilon(x) = P_X(B_\epsilon(x))$, there exists $\mu_0 \in (0, a_d)$ such that for all $x \in \mathbb{R}^d$ and $\epsilon \in (0, \epsilon_0]$, we have

$$p_\epsilon(x) \geq \mu_0 \epsilon^d \bar{f}(x).$$

Assumption A3. We have $\inf_{x_0\in\mathcal{S}}\|\dot{\eta}(x_0)\| > 0$, so that $\mathcal{S}$ is a $(d-1)$-dimensional, orientable manifold. Moreover, $\sup_{x\in\mathcal{S}^{2\epsilon_0}}\|\dot{\eta}(x)\| < \infty$ and $\ddot{\eta}$ is uniformly continuous on $\mathcal{S}^{2\epsilon_0}$ with $\sup_{x\in\mathcal{S}^{2\epsilon_0}}\|\ddot{\eta}(x)\|_{\mathrm{op}} < \infty$. Finally, the function η is continuous, and

$$\inf_{x\in\mathbb{R}^d\setminus\mathcal{S}^{\epsilon_0}} |\eta(x) - 1/2| > 0.$$

Assumption A4(α). We have that $\int_{\mathbb{R}^d}\|x\|^\alpha\, dP_X(x) < \infty$ and $\int_{\mathcal{S}}\bar{f}(x_0)^{d/(\alpha+d)}\, d\mathrm{Vol}^{d-1}(x_0) < \infty$, where $d\mathrm{Vol}^{d-1}$ denotes the $(d-1)$-dimensional volume form on $\mathcal{S}$.

Proof of Theorem 2. Part 1: We show that the distribution $\tilde{P}$ of the pair $(X, \tilde{Y})$ satisfies suitably modified versions of Assumptions A1, A2, A3 and A4(α).

Assumption A1: For $r \in \{0,1\}$, let $\tilde{P}_r$ denote the conditional distribution of X given $\tilde{Y} = r$. For $x \in \mathbb{R}^d$, and $r = 0, 1$, define

$$\tilde{f}_r(x) = \frac{\pi_r\{1-\rho_r(x)\}f_r(x) + \pi_{1-r}\rho_{1-r}(x)f_{1-r}(x)}{\int_{\mathbb{R}^d}\pi_r\{1-\rho_r(z)\}f_{1-r}(z) + \pi_{1-r}\rho_{1-r}(z)f_{1-r}(z)\, dz}.$$

Now, for a Borel subset A of $\mathbb{R}^d$, we have that

$$\begin{aligned}\tilde{P}_1(A) = \mathrm{pr}(X \in A \mid \tilde{Y} = 1) &= \frac{\mathrm{pr}(X \in A, \tilde{Y} = 1)}{\mathrm{pr}(\tilde{Y} = 1)} \\ &= \frac{\pi_1 \mathrm{pr}(X \in A, \tilde{Y} = 1 \mid Y = 1) + \pi_0 \mathrm{pr}(X \in A, \tilde{Y} = 1 \mid Y = 1)}{\mathrm{pr}(\tilde{Y} = 1)} \\ &= \frac{\pi_1 \int_A \{1 - \rho_1(x)\} f_1(x)\, dx + \pi_0 \int_A \rho_0(x) f_0(x)\, dx}{\mathrm{pr}(\tilde{Y} = 1)} = \int_A \tilde{f}_1(x)\, dx.\end{aligned}$$

Similarly, $\tilde{P}_0(A) = \int_A \tilde{f}_0(x)\, dx$. Hence $\tilde{P}_0$ and $\tilde{P}_1$ are absolutely continuous with respect to Lebesgue measure, with Radon–Nikodym derivatives $\tilde{f}_0$ and $\tilde{f}_1$, respectively. Furthermore, $\tilde{f} = \mathrm{pr}(\tilde{Y} = 0)\tilde{f}_0 + \mathrm{pr}(\tilde{Y} = 1)\tilde{f}_1 = \bar{f}$ is continuous and positive.

Assumption A2: Since A2 refers mainly to the marginal distribution of X, which is unchanged under the addition of label noise, this assumption is trivially satisfied for $\tilde{f} = \bar{f}$, as long as $\tilde{\mathcal{S}} = \{x \in \mathbb{R}^d : \tilde{\eta}(x) = 1/2\} = \mathcal{S}$. To see this, let $\delta_0 > 0$ and note that for x satisfying $\eta(x) - 1/2 > \delta_0$, we have from (1) that

$$\begin{aligned}\tilde{\eta}(x) - 1/2 &= \{\eta(x) - 1/2\}\{1 - \rho_0(x) - \rho_1(x)\}\Big\{1 + \frac{\rho_0(x) - \rho_1(x)}{\{2\eta(x) - 1\}\{1 - \rho_0(x) - \rho_1(x)\}}\Big\} \\ &> \{\eta(x) - 1/2\}(1 - 2\rho^*)(1 - a^*) \geq \delta_0(1 - 2\rho^*)(1 - a^*). \qquad \text{(A8)}\end{aligned}$$

Similarly, if $1/2 - \eta(x) > \delta_0$, then we have that $1/2 - \tilde{\eta}(x) > \delta_0(1 - 2\rho^*)(1 - a^*)$. It follows that $\tilde{\mathcal{S}} \subseteq \mathcal{S}$. Now, for x such that $|\eta(x) - 1/2| < \delta$, we have

$$\tilde{\eta}(x) - 1/2 = \eta(x) - 1/2 + \{1 - \eta(x)\}g(\eta(x)) - \eta(x)g(1 - \eta(x)). \qquad \text{(A9)}$$

Thus $\mathcal{S} \subseteq \tilde{\mathcal{S}}$.

Assumption A3: Since g is twice continuously differentiable, we have that $\tilde{\eta}$ is twice continuously differentiable on the set $\{x \in \mathcal{S}^{2\epsilon_0} : |\eta(x) - 1/2| < \delta\}$. On this set, its gradient vector at x is

$$\dot{\tilde{\eta}}(x) = \dot{\eta}(x)\Big[1 - g(\eta(x)) - g(1 - \eta(x)) + \{1 - \eta(x)\}\dot{g}(\eta(x)) + \eta(x)\dot{g}(1 - \eta(x))\Big].$$

The corresponding Hessian matrix at x is

$$\begin{aligned}\ddot{\tilde{\eta}}(x) = \ddot{\eta}(x)&\Big[1 - g(\eta(x)) - g(1 - \eta(x)) + \{1 - \eta(x)\}\dot{g}(\eta(x)) + \eta(x)\dot{g}(1 - \eta(x))\Big] \\ &- \dot{\eta}(x)\Big[\dot{\eta}(x)^T\dot{g}(\eta(x)) - \dot{\eta}(x)^T\dot{g}(1 - \eta(x)) + \dot{\eta}(x)^T\dot{g}(\eta(x)) \\ &\qquad - \{1 - \eta(x)\}\dot{\eta}(x)^T\ddot{g}(\eta(x)) - \dot{\eta}(x)^T\dot{g}(1 - \eta(x)) + \eta(x)\dot{\eta}(x)^T\ddot{g}(1 - \eta(x))\Big].\end{aligned}$$

In particular, for $x_0 \in \mathcal{S}$ we have

$$\dot{\tilde{\eta}}(x_0) = \dot{\eta}(x_0)\{1 - 2g(1/2) + \dot{g}(1/2)\}; \quad \ddot{\tilde{\eta}}(x_0) = \ddot{\eta}(x_0)\{1 - 2g(1/2) + \dot{g}(1/2)\}. \qquad \text{(A10)}$$

Now define

$$\epsilon_1 = \sup\Big\{\epsilon > 0 : \sup_{x \in \mathcal{S}^{2\epsilon}} |\eta(x) - 1/2| < \delta\Big\} > 0,$$

where the fact that ϵ_1 is positive follows from Assumption A3. Set $\tilde{\epsilon}_0 = \min\{\epsilon_0, \epsilon_1\}/2$. Then, using the properties of g, we have that $\inf_{x_0 \in \mathcal{S}} \|\dot{\tilde{\eta}}(x_0)\| > 0$. Moreover, $\sup_{x \in \mathcal{S}^{2\tilde{\epsilon}_0}} \|\dot{\tilde{\eta}}(x)\| < \infty$ and $\ddot{\tilde{\eta}}$ is uniformly continuous on $\mathcal{S}^{2\tilde{\epsilon}_0}$ with $\sup_{x \in \mathcal{S}^{2\tilde{\epsilon}_0}} \|\ddot{\tilde{\eta}}(x)\|_{\mathrm{op}} < \infty$. Finally, the function $\tilde{\eta}$ is continuous since ρ_0, ρ_1 are continuous, and, by (A8),

$$\inf_{x \in \mathbb{R}^d \setminus \mathcal{S}^{\tilde{\epsilon}_0}} |\tilde{\eta}(x) - 1/2| > 0.$$

Assumption A4(α): This holds for $\tilde{P}$ because the marginal distribution of X is unaffected by the label noise and $\tilde{\mathcal{S}} = \mathcal{S}$.

Part 2: Recall the function F defined in (A7). Let $c_n = F(k/(n-1))$, and set $\epsilon_n = \{c_n \beta^{1/2} \log^{1/2}(n-1)\}^{-1}$, $\Delta_n = k(n-1)^{-1} c_n^d \log^d((n-1)/k)$, $\mathcal{R}_n = \{x \in \mathbb{R}^d : \bar{f}(x) > \Delta_n\}$ and $\mathcal{S}_n = \mathcal{S} \cap \mathcal{R}_n$. Then, by (A8) and the fact that $\inf_{x_0 \in \mathcal{S}} \|\dot{\tilde{\eta}}(x_0)\| > 0$, there exists $c_0 > 0$ such that for every $\epsilon \in (0, \tilde{\epsilon}_0]$,

$$\inf_{x \in \mathbb{R}^d \setminus \mathcal{S}^\epsilon} |\tilde{\eta}(x) - 1/2| > c_0 \epsilon.$$

Now let $\tilde{S}_n(x) = k^{-1} \sum_{i=1}^k \mathbb{1}_{\{\tilde{Y}_{(i)}=1\}}$, $X^n = (X_1, \ldots, X_n)$ and $\tilde{\mu}(x, X^n) = E\{\tilde{S}_n(x) \mid X^n\} = k^{-1} \sum_{i=1}^k \tilde{\eta}(X_{(i)})$. Define $A_k = \big\{\|X_{(k)}(x) - x\| \le \epsilon_n/2 \text{ for all } x \in \mathcal{R}_n\big\}$. Now suppose that $z_1, \ldots, z_N \in \mathcal{R}_n$ are such that $\|z_j - z_\ell\| > \epsilon_n/4$ for all $j \neq \ell$, but $\sup_{x \in \mathcal{R}_n} \min_{j=1,\ldots,N} \|x - z_j\| \le \epsilon_n/4$. Then by the final part of Assumption A2, for $n \ge 2$ large enough that $\epsilon_n/8 \le \epsilon_0$, we have

$$1 = P_X(\mathbb{R}^d) \ge \sum_{j=1}^N p_{\epsilon_n/8}(z_j) \ge \frac{N \mu_0 \beta^{d/2} \log^{d/2}(n-1)}{8^d (n-1)^{1-\beta}}.$$

Then by a standard binomial tail bound (Shorack & Wellner, 1986, Equation (6), p. 440), for such n and any $M > 0$,

$$\begin{aligned}
\mathrm{pr}(A_k^c) &= \mathrm{pr}\Big\{\sup_{x \in \mathcal{R}_n} \|X_{(k)}(x) - x\| > \epsilon_n/2\Big\} \le \mathrm{pr}\Big\{\max_{j=1,\ldots,N} \|X_{(k)}(z_j) - z_j\| > \epsilon_n/4\Big\} \\
&\le \sum_{j=1}^N \mathrm{pr}\big\{\|X_{(k)}(z_j) - z_j\| > \epsilon_n/4\big\} \le N \max_{j=1,\ldots,N} \exp\Big(-\frac{1}{2} n p_{\epsilon_n/4}(z_j) + k\Big) = O(n^{-M}),
\end{aligned}$$

uniformly for $k \in K_\beta$.

Now, on the event A_k, for $\epsilon_n < \tilde{\epsilon}_0$ and $x \in \mathcal{R}_n \setminus \mathcal{S}^{\epsilon_n}$, the k nearest neighbours of x are on the same side of $\mathcal{S}$, so

$$|\tilde{\mu}_n(x, X^n) - 1/2| = \Big|\frac{1}{k} \sum_{i=1}^k \tilde{\eta}(X_{(i)}) - \frac{1}{2}\Big| \ge \inf_{z \in B_{\epsilon_n/2}(x)} |\tilde{\eta}(z) - 1/2| \ge c_0 \frac{\epsilon_n}{2}.$$

Moreover, conditional on X^n, $\tilde{S}_n(x)$ is the sum of k independent terms. Therefore, by Hoeffding's inequality,

$$\begin{aligned}
&\sup_{k \in K_\beta} \sup_{x \in \mathcal{R}_n \setminus \mathcal{S}^{\epsilon_n}} \big|\mathrm{pr}\{\tilde{C}_n^{k\mathrm{nn}}(x) = 0\} - \mathbb{1}_{\{\tilde{\eta}(x) < 1/2\}}\big| \\
&\quad = \sup_{k \in K_\beta} \sup_{x \in \mathcal{R}_n \setminus \mathcal{S}^{\epsilon_n}} \big|\mathrm{pr}\{\tilde{S}_n(x) < 1/2\} - \mathbb{1}_{\{\tilde{\eta}(x) < 1/2\}}\big| \\
&\quad = \sup_{k \in K_\beta} \sup_{x \in \mathcal{R}_n \setminus \mathcal{S}^{\epsilon_n}} \big|E\{\mathrm{pr}\{\tilde{S}_n(x) < 1/2 \mid X^n\} - \mathbb{1}_{\{\tilde{\eta}(x) < 1/2\}}\}\big| \\
&\quad \le \sup_{k \in K_\beta} \sup_{x \in \mathcal{R}_n \setminus \mathcal{S}^{\epsilon_n}} E\big[\exp(-2k\{\tilde{\mu}_n(x, X^n) - 1/2\}^2) \mathbb{1}_{A_k}\big] + \sup_{k \in K_\beta} \mathrm{pr}(A_k^c) = O(n^{-M}) \quad \text{(A11)}
\end{aligned}$$

for every $M > 0$.

Next, for $x \in \mathcal{S}^{\epsilon_2}$, we have $|\eta(x) - 1/2| < \delta$, and therefore, letting $t = \eta(x) - 1/2$, from (A9) we can write

$$\begin{aligned} 2\eta(x) - 1 - \frac{2\tilde{\eta}(x) - 1}{1 - 2g(1/2) + \dot{g}(1/2)} & \\ = \{2\eta(x) - 1\}\Big\{1 - \frac{1 - g(\eta(x)) - g(1 - \eta(x))}{1 - 2g(1/2) + \dot{g}(1/2)}\Big\} - \frac{g(\eta(x)) - g(1 - \eta(x))}{1 - 2g(1/2) + \dot{g}(1/2)} & \\ = 2t\Big\{1 - \frac{1 - g(1/2 + t) - g(1/2 - t)}{1 - 2g(1/2) + \dot{g}(1/2)}\Big\} - \frac{g(1/2 + t) - g(1/2 - t)}{1 - 2g(1/2) + \dot{g}(1/2)} = G(t), & \end{aligned} \tag{170}$$

say. Observe that

$$\dot{G}(t) = 2\Big\{1 - \frac{1 - g(1/2 + t) - g(1/2 - t)}{1 - 2g(1/2) + \dot{g}(1/2)}\Big\} + \frac{(2t - 1)\dot{g}(1/2 + t) - (2t + 1)\dot{g}(1/2 - t)}{1 - 2g(1/2) + \dot{g}(1/2)};$$

and

$$\ddot{G}(t) = \frac{4\{\dot{g}(1/2 + t) - \dot{g}(1/2 - t)\}}{1 - 2g(1/2) + \dot{g}(1/2)} + \frac{(2t - 1)\ddot{g}(1/2 + t) + (2t + 1)\ddot{g}(1/2 - t)}{1 - 2g(1/2) + \dot{g}(1/2)}.$$

In particular, we have $G(0) = 0$, $\dot{G}(0) = 0$, $\ddot{G}(0) = 0$ and $\ddot{G}$ is bounded on $(-\delta, \delta)$. 175

Now there exists n_0 such that $\epsilon_n < \epsilon_2$, for all $n > n_0$ and $k \in K_\beta$. Therefore, writing $\mathcal{S}_n^{\epsilon_n} = \mathcal{S}^{\epsilon_n} \cap \mathcal{R}_n$, for $n > n_0$, we have that

$$\begin{aligned} & \Big| R(\tilde{C}^{k\mathrm{nn}}) - R(C^{\mathrm{Bayes}}) - \frac{\tilde{R}(\tilde{C}^{k\mathrm{nn}}) - \tilde{R}(\tilde{C}^{\mathrm{Bayes}})}{1 - 2g(1/2) + \dot{g}(1/2)} \Big| \\ & = \Big| \int_{\mathbb{R}^d} [\mathrm{pr}\{\tilde{C}^{k\mathrm{nn}}(x) = 0\} - \mathbb{1}_{\{\tilde{\eta}(x) < 1/2\}}]\Big\{2\eta(x) - 1 - \frac{2\tilde{\eta}(x) - 1}{1 - 2g(1/2) + \dot{g}(1/2)}\Big\}\, dP_X(x) \Big| \\ & \le \Big| \int_{\mathcal{S}_n^{\epsilon_n}} [\mathrm{pr}\{\tilde{C}^{k\mathrm{nn}}(x) = 0\} - \mathbb{1}_{\{\tilde{\eta}(x) < 1/2\}}]\Big\{2\eta(x) - 1 - \frac{2\tilde{\eta}(x) - 1}{1 - 2g(1/2) + \dot{g}(1/2)}\Big\}\, dP_X(x) \Big| \\ & \qquad + \Big(1 + \frac{1}{1 - 2g(1/2) + \dot{g}(1/2)}\Big) P_X(\mathcal{R}_n^c) + O(n^{-M}), \end{aligned} \tag{180}$$

uniformly for $k \in K_\beta$, where the final claim uses (A11). Then, by a Taylor expansion of G about $t = 0$, we have that

$$\begin{aligned} & \Big| \int_{\mathcal{S}_n^{\epsilon_n}} [\mathrm{pr}\{\tilde{C}^{k\mathrm{nn}}(x) = 0\} - \mathbb{1}_{\{\tilde{\eta}(x) < 1/2\}}]\Big\{2\eta(x) - 1 - \frac{2\tilde{\eta}(x) - 1}{1 - 2g(1/2) + \dot{g}(1/2)}\Big\}\, dP_X(x) \Big| \\ & \le \frac{1}{2} \sup_{t \in (-\delta, \delta)} |\ddot{G}(t)| \int_{\mathcal{S}_n^{\epsilon_n}} |\mathrm{pr}\{\tilde{C}^{k\mathrm{nn}}(x) = 0\} - \mathbb{1}_{\{\tilde{\eta}(x) < 1/2\}}| \{2\eta(x) - 1\}^2\, dP_X(x) \\ & \le \frac{1}{2} \sup_{t \in (-\delta, \delta)} |\ddot{G}(t)| \sup_{x \in \mathcal{S}_n^{\epsilon_n}} |2\eta(x) - 1| \int_{\mathcal{S}_n^{\epsilon_n}} \{\mathrm{pr}\{\tilde{C}^{k\mathrm{nn}}(x) = 0\} - \mathbb{1}_{\{\tilde{\eta}(x) < 1/2\}}\}\{2\eta(x) - 1\}\, dP_X(x) \\ & \le \frac{1}{2} \sup_{t \in (-\delta, \delta)} |\ddot{G}(t)| \sup_{x \in \mathcal{S}_n^{\epsilon_n}} |2\eta(x) - 1| \{R(\tilde{C}^{k\mathrm{nn}}) - R(C^{\mathrm{Bayes}})\} \\ & \le \frac{1}{2} \sup_{t \in (-\delta, \delta)} |\ddot{G}(t)| \sup_{x \in \mathcal{S}_n^{\epsilon_n}} |2\eta(x) - 1| \frac{\tilde{R}(\tilde{C}^{k\mathrm{nn}}) - \tilde{R}(\tilde{C}^{\mathrm{Bayes}})}{(1 - 2\rho^*)(1 - a^*)} = o\Big(\tilde{R}(\tilde{C}^{k\mathrm{nn}}) - \tilde{R}(\tilde{C}^{\mathrm{Bayes}})\Big), \end{aligned} \tag{185}$$

uniformly for $k \in K_\beta$.

Finally, to bound $P_X(\mathcal{R}_n^c)$, we have by the moment condition in Assumption A4(α) and Hölder's inequality, that for any $u \in (0,1)$, and $v > 0$,

$$P_X(\mathcal{R}_n^c) = \mathrm{pr}\{\bar{f}(X) \leq \Delta_n\} \leq (\Delta_n)^{\frac{\alpha(1-u)}{\alpha+d}} \int_{x:\bar{f}(x)\leq\Delta_n} \bar{f}(x)^{1-\frac{\alpha(1-u)}{\alpha+d}}\,dx$$
$$\leq (\Delta_n)^{\frac{\alpha(1-u)}{\alpha+d}} \Big\{\int_{\mathbb{R}^d} (1+\|x\|^\alpha)\bar{f}(x)\,dx\Big\}^{1-\frac{\alpha(1-u)}{\alpha+d}} \Big\{\int_{\mathbb{R}^d} \frac{1}{(1+\|x\|^\alpha)^{\frac{d+\alpha u}{\alpha(1-u)}}}\,dx\Big\}^{\frac{\alpha(1-u)}{\alpha+d}} = o\bigg(\Big(\frac{k}{n}\Big)^{\frac{\alpha(1-u)}{\alpha+d}-v}\bigg),$$

uniformly for $k \in K_\beta$.

Since $u \in (0,1)$ was arbitrary, we have shown that, that for any $v > 0$,

$$R(\tilde{C}^{k\mathrm{nn}}) - R(C^{\mathrm{Bayes}}) - \frac{\tilde{R}(\tilde{C}^{k\mathrm{nn}}) - \tilde{R}(\tilde{C}^{\mathrm{Bayes}})}{1-2g(1/2)+\dot{g}(1/2)} = o\bigg(\tilde{R}(\tilde{C}^{k\mathrm{nn}}) - \tilde{R}(\tilde{C}^{\mathrm{Bayes}}) + \Big(\frac{k}{n}\Big)^{\frac{\alpha}{\alpha+d}-v}\bigg),$$

uniformly for $k \in K_\beta$. Since Assumptions A1, A2, A3 and A4(α) hold for $\tilde{P}$, the proof is completed by an application of Cannings et al. (2018, Theorem 1), together with (A10). □

A·3. *Proofs from Section* 4·2

Before presenting the proofs from this section, we briefly discuss measurability issues for the SVM classifier. Since this is constructed by solving the minimization problem in (9), it is not immediately clear that it is measurable. It is convenient to let $\mathcal{C}_d$ denote the set of all measurable functions from $\mathbb{R}^d$ to $\{0,1\}$. By Steinwart & Christmann (2008, Definition 6.2, Lemma 6.3 and Lemma 6.23), we have that the function $\tilde{C}_n^{\mathrm{SVM}} : (\mathbb{R}^d \times \{0,1\})^n \to \mathcal{C}_d$ and the map from $(\mathbb{R}^d \times \{0,1\})^n \times \mathbb{R}^d$ to $\{0,1\}$ given by $\big((x_1,\tilde{y}_1),\ldots,(x_n,\tilde{y}_n),x\big) \mapsto \tilde{C}_n^{\mathrm{SVM}}(x)$ are measurable with respect to the universal completion of the product σ-algebras on $(\mathbb{R}^d \times \{0,1\})^n$ and $(\mathbb{R}^d \times \{0,1\})^n \times \mathbb{R}^d$, respectively. We can therefore avoid measurability issues by taking our underlying probability space $(\Omega, \mathcal{F}, \mathrm{pr})$ to be as follows: let $\Omega = (\mathbb{R}^d \times \{0,1\} \times \{0,1\})^{n+1}$, and $\mathcal{F}$ to be the universal completion of the product σ-algebra on Ω. Moreover, we let pr denote the canonical extension of the product measure on Ω. The triples $(X_1,Y_1,\tilde{Y}_1),\ldots,(X_n,Y_n,\tilde{Y}_n),(X,Y,\tilde{Y})$ can be taken to be the coordinate projections of the $(n+1)$ components of Ω.

Proof of Theorem 3. We first aim to show that $\tilde{P}$ satisfies the margin assumption with parameter γ_1, and has geometric noise exponent γ_2. For the first of these claims, by (A2), we have for all $t > 0$ that

$$P_X(\{x \in \mathbb{R}^d : 0 < |\tilde{\eta}(x) - 1/2| \leq t\}) \leq P_X\big(\{x : 0 < |\eta(x) - 1/2|(1-2\rho^*)(1-a^*) \leq t\}\big)$$
$$\leq \frac{\kappa_1}{(1-2\rho^*)^{\gamma_1}(1-a^*)^{\gamma_1}} t^{\gamma_1},$$

as required; see also the discussion in Section 3.9.1 of the 2015 Australian National University PhD thesis by M. van Rooyen (`https://openresearch-repository.anu.edu.au/handle/1885/99588`). The proof of the second claim is more involved, because we require a bound on $|2\tilde{\eta}(x) - 1|$ in terms of $|2\eta(x) - 1|$. We consider separately the cases where $|\eta(x) - 1/2|$ is small and large, and for $r > 0$, define $\mathcal{E}_r = \{x \in \mathbb{R}^d : |\eta(x) - 1/2| < r\}$. For $x \in \mathcal{E}_\delta \cap \mathcal{S}^c$, we can write $t_0 = \eta(x) - 1/2 \in (-\delta, \delta)$, so that by (A9) again,

$$2\tilde{\eta}(x) - 1 = \{2\eta(x)-1\}\Big\{1 - g(\eta(x)) - g(1-\eta(x)) + \frac{g(\eta(x)) - g(1-\eta(x))}{2\eta(x)-1}\Big\}$$
$$= \{2\eta(x)-1\}\Big\{1 - g(1/2+t_0) - g(1/2-t_0) + \frac{g(1/2+t_0) - g(1/2-t_0)}{2t_0}\Big\}. \quad \text{(A12)}$$

Now, by reducing $\delta > 0$ if necessary, and since $1 - 2g(1/2) + \dot{g}(1/2) > 0$ by hypothesis, we may assume that

$$\left|1 - g(1/2 + t_0) - g(1/2 - t_0) + \frac{g(1/2 + t_0) - g(1/2 - t_0)}{2t_0}\right| \leq 2\{1 - 2g(1/2) + \dot{g}(1/2)\} \quad \text{(A13)}$$

for all $t_0 \in [-\delta, \delta]$. Moreover, for $x \in \mathcal{E}_\delta^c$, we have

$$\begin{aligned}
&\Big|\{2\eta(x) - 1\}\{1 - \rho_0(x) - \rho_1(x)\} + \rho_0(x) - \rho_1(x)\Big| \\
&\qquad = |2\eta(x) - 1|\left|1 - \rho_0(x) - \rho_1(x) + \frac{\rho_0(x) - \rho_1(x)}{2\eta(x) - 1}\right| \\
&\qquad \leq |2\eta(x) - 1|\left\{1 + \frac{|\rho_0(x) - \rho_1(x)|}{2\delta}\right\} \leq |2\eta(x) - 1|\Big(1 + \frac{1}{2\delta_0}\Big). \qquad \text{(A14)}
\end{aligned}$$

Now that we have the required bounds on $|2\tilde{\eta}(x) - 1|$, we deduce from (A12), (A13) and (A14) that

$$\begin{aligned}
&\int_{\mathbb{R}^d} |2\tilde{\eta}(x) - 1| \exp\Big(-\frac{\tau_x^2}{t^2}\Big)\, dP_X(x) \\
&\qquad = \int_{\mathbb{R}^d} \Big|\{2\eta(x) - 1\}\{1 - \rho_0(x) - \rho_1(x)\} + \rho_0(x) - \rho_1(x)\Big| \exp\Big(-\frac{\tau_x^2}{t}\Big)\, dP_X(x) \\
&\qquad \leq \max\Big\{2 - 4g(1/2) + 2\dot{g}(1/2), 1 + \frac{1}{2\delta_0}\Big\} \int_{\mathbb{R}^d} |2\eta(x) - 1| \exp\Big(-\frac{\tau_x^2}{t}\Big)\, dP_X(x) \\
&\qquad \leq \max\Big\{2 - 4g(1/2) + 2\dot{g}(1/2), 1 + \frac{1}{2\delta_0}\Big\}\kappa_2 t^{\gamma_2 d},
\end{aligned}$$

so $\tilde{P}$ does indeed have geometric noise exponent γ_2.

Now, for an arbitrary classifier C, let $\tilde{L}(C) = \tilde{P}\big(\{(x, y) \in \mathbb{R}^d \times \{0, 1\} : C(x) \neq y\}\big)$ denote the test error. The quantity $\tilde{L}(\tilde{C}^{\mathrm{SVM}})$ is random because the classifier depends on the training data and the probability in the definition of $\tilde{L}(\cdot)$ is with respect to test data only. It follows by Steinwart & Scovel (2007, Theorem 2.8) that, for all $\epsilon > 0$, there exists $M > 0$ such that for all $n \in \mathbb{N}$ and all $\tau \geq 1$,

$$\mathrm{pr}\Big(\tilde{L}(\tilde{C}^{\mathrm{SVM}}) - \tilde{L}(\tilde{C}^{\mathrm{Bayes}}) > M\tau^2 n^{-\Gamma+\epsilon}\Big) \leq e^{-\tau}.$$

We conclude by Theorem 1(ii) that

$$\begin{aligned}
R(\tilde{C}^{\mathrm{SVM}}) - R(C^{\mathrm{Bayes}}) &\leq \frac{\tilde{R}(\tilde{C}^{\mathrm{SVM}}) - \tilde{R}(\tilde{C}^{\mathrm{Bayes}})}{(1 - 2\rho^*)(1 - a^*)} \\
&= \frac{1}{(1 - 2\rho^*)(1 - a^*)} \int_0^\infty \mathrm{pr}\Big(\tilde{L}(\tilde{C}^{\mathrm{SVM}}) - \tilde{L}(\tilde{C}^{\mathrm{Bayes}}) > u\Big)\, du \\
&= \frac{2Mn^{-\Gamma+\epsilon}}{(1 - 2\rho^*)(1 - a^*)} \int_0^\infty \tau \mathrm{pr}\Big(\tilde{L}(\tilde{C}^{\mathrm{SVM}}) - \tilde{L}(\tilde{C}^{\mathrm{Bayes}}) > M\tau^2 n^{-\Gamma+\epsilon}\Big)\, d\tau \\
&\leq \frac{2Mn^{-\Gamma+\epsilon}}{(1 - 2\rho^*)(1 - a^*)}\left\{\int_0^1 \tau\, d\tau + \int_1^\infty \tau \exp(-\tau)\, d\tau\right\} = \frac{Mn^{-\Gamma+\epsilon}}{(1 - 2\rho^*)(1 - a^*)}\Big(1 + \frac{4}{e}\Big),
\end{aligned}$$

as required. □

A·4. *Proofs from Section* 4·3

Proof of Lemma 1. Since, for homogeneous noise, the pair (X, Y) and the noise indicator Z are independent, we have $\mathrm{pr}\{C(X) \neq Y \mid Z = r\} = \mathrm{pr}\{C(X) \neq Y\}$, for $r = 0, 1$. It follows that

$$\begin{aligned}
\tilde{R}(C) = \mathrm{pr}\{C(X) \neq \tilde{Y}\} &= \mathrm{pr}(Z = 1)\mathrm{pr}\{C(X) \neq Y \mid Z = 1\} + \mathrm{pr}(Z = 0)\mathrm{pr}\{C(X) = Y \mid Z = 0\} \\
&= (1 - \rho)\mathrm{pr}\{C(X) \neq Y\} + \rho[1 - \mathrm{pr}\{C(X) \neq Y\}] \\
&= \rho + (1 - 2\rho)R(C).
\end{aligned}$$

Rearranging terms gives the first part of the lemma, and the second part follows immediately. □

Proof of Theorem 4. For $r \in \{0,1\}$, we have that $\hat{\pi}_r \xrightarrow{\text{a.s.}} (1-\rho)\pi_r + \rho\pi_{1-r} = (1-2\rho)\pi_r + \rho$. Now, writing

$$\hat{\mu}_r = \frac{n^{-1}\sum_{i=1}^n X_i \mathbb{1}_{\{\tilde{Y}_i = r\}}}{\hat{\pi}_r} = \frac{n^{-1}\sum_{i=1}^n X_i \mathbb{1}_{\{\tilde{Y}_i = r\}}(\mathbb{1}_{\{Y_i = r\}} + \mathbb{1}_{\{Y_i = 1-r\}})}{\hat{\pi}_r},$$

we see that

$$\hat{\mu}_r \xrightarrow{\text{a.s.}} \frac{(1-\rho)\pi_r\mu_r + \rho\pi_{1-r}\mu_{1-r}}{(1-\rho)\pi_r + \rho\pi_{1-r}}.$$

Hence

$$\begin{aligned}\hat{\mu}_1 + \hat{\mu}_0 &\xrightarrow{\text{a.s.}} \frac{(1-\rho)\pi_1\mu_1 + \rho\pi_0\mu_0}{(1-\rho)\pi_1 + \rho\pi_0} + \frac{(1-\rho)\pi_0\mu_0 + \rho\pi_1\mu_1}{(1-\rho)\pi_0 + \rho\pi_1} \\ &= \mu_1\left\{\frac{(1-2\rho)^2\pi_0\pi_1 + 2\rho(1-\rho)\pi_1}{(1-2\rho)^2\pi_0\pi_1 + \rho(1-\rho)}\right\} + \mu_0\left\{\frac{(1-2\rho)^2\pi_0\pi_1 + 2\rho(1-\rho)\pi_0}{(1-2\rho)^2\pi_0\pi_1 + \rho(1-\rho)}\right\}.\end{aligned}$$

Moreover

$$\begin{aligned}\hat{\mu}_1 - \hat{\mu}_0 &\xrightarrow{\text{a.s.}} \frac{(1-\rho)\pi_1\mu_1 + \rho\pi_0\mu_0}{(1-\rho)\pi_1 + \rho\pi_0} - \frac{(1-\rho)\pi_0\mu_0 + \rho\pi_1\mu_1}{(1-\rho)\pi_0 + \rho\pi_1} \\ &= \left\{\frac{(1-2\rho)\pi_0\pi_1}{(1-2\rho)^2\pi_0\pi_1 + \rho(1-\rho)}\right\}(\mu_1 - \mu_0).\end{aligned}$$

Observe further that

$$\begin{aligned}\hat{\Sigma} &\xrightarrow{\text{a.s.}} \operatorname{cov}\big((X_1 - \tilde{\mu}_1)(X_1 - \tilde{\mu}_1)^T \mathbb{1}_{\{\tilde{Y}_1 = 1\}} + (X_1 - \tilde{\mu}_0)(X_1 - \tilde{\mu}_0)^T \mathbb{1}_{\{\tilde{Y}_1 = 0\}}\big) \\ &= \{(1-2\rho)\pi_1 + \rho\}\tilde{\Sigma}_1 + \{(1-2\rho)\pi_0 + \rho\}\tilde{\Sigma}_0,\end{aligned}$$

where $\tilde{\Sigma}_r = \operatorname{cov}(X \mid \tilde{Y} = r)$, and we now seek to express $\tilde{\Sigma}_0$ and $\tilde{\Sigma}_1$ in terms of ρ, π_0, π_1, μ_0, μ_1 and Σ. To that end, we have that

$$\tilde{\Sigma}_r = E\{\operatorname{cov}(X \mid Y, \tilde{Y} = r) \mid \tilde{Y} = r\} + \operatorname{cov}\{E(X \mid Y, \tilde{Y} = r) \mid \tilde{Y} = r\} = \Sigma + \operatorname{cov}\{\mu_Y \mid \tilde{Y} = r\}.$$

Note that

$$\operatorname{pr}(Y = 1 \mid \tilde{Y} = 1) = \frac{\operatorname{pr}(Y = 1, \tilde{Y} = 1)}{\operatorname{pr}(\tilde{Y} = 1)} = \frac{\pi_1(1-\rho)}{\pi_1(1-\rho) + \pi_0\rho} = \frac{\pi_1(1-\rho)}{\pi_1(1-2\rho) + \rho}.$$

Hence

$$E(\mu_Y \mid \tilde{Y} = 1) = \mu_1 \operatorname{pr}(Y = 1 \mid \tilde{Y} = 1) + \mu_0 \operatorname{pr}(Y = 0 \mid \tilde{Y} = 1) = \frac{\pi_1\mu_1(1-\rho) + \pi_0\mu_0\rho}{\pi_1(1-2\rho) + \rho}.$$

It follows that

$$\begin{aligned}
\tilde{\Sigma}_1 &= \frac{\pi_1(1-\rho)}{\pi_1(1-2\rho)+\rho}\Big(\mu_1 - \frac{\pi_1\mu_1(1-\rho)+\pi_0\mu_0\rho}{\pi_1(1-2\rho)+\rho}\Big)\Big(\mu_1 - \frac{\pi_1\mu_1(1-\rho)+\pi_0\mu_0\rho}{\pi_1(1-2\rho)+\rho}\Big)^T \\
&\quad + \frac{\pi_0\rho}{\pi_1(1-2\rho)+\rho}\Big(\mu_0 - \frac{\pi_1\mu_1(1-\rho)+\pi_0\mu_0\rho}{\pi_1(1-2\rho)+\rho}\Big)\Big(\mu_0 - \frac{\pi_1\mu_1(1-\rho)+\pi_0\mu_0\rho}{\pi_1(1-2\rho)+\rho}\Big)^T \\
&= \frac{\pi_1(1-\rho)}{\pi_1(1-2\rho)+\rho}\Big(\frac{\pi_0\rho(\mu_1-\mu_0)}{\pi_1(1-2\rho)+\rho}\Big)\Big(\frac{\pi_0\rho(\mu_1-\mu_0)}{\pi_1(1-2\rho)+\rho}\Big)^T \\
&\quad + \frac{\pi_0\rho}{\pi_1(1-2\rho)+\rho}\Big(\frac{\pi_1(1-\rho)(\mu_0-\mu_1)}{\pi_1(1-2\rho)+\rho}\Big)\Big(\frac{\pi_1(1-\rho)(\mu_0-\mu_1)}{\pi_1(1-2\rho)+\rho}\Big)^T \\
&= \frac{\pi_0\pi_1\rho(1-\rho)}{(\pi_1(1-\rho)+\pi_0\rho)^2}(\mu_1-\mu_0)(\mu_1-\mu_0)^T.
\end{aligned}$$

Similarly

$$\tilde{\Sigma}_0 = \frac{\pi_0\pi_1\rho(1-\rho)}{(\pi_0(1-\rho)+\pi_1\rho)^2}(\mu_1-\mu_0)(\mu_1-\mu_0)^T.$$

We deduce that

$$\tilde{\Sigma} \stackrel{\text{a.s.}}{\to} \Sigma + \frac{\pi_0\pi_1\rho(1-\rho)}{\pi_1\pi_0(1-2\rho)^2+\rho(1-\rho)}(\mu_1-\mu_0)(\mu_1-\mu_0)^T = \Sigma + \alpha(\mu_1-\mu_0)(\mu_1-\mu_0)^T,$$

where $\alpha = \pi_0\pi_1\rho(1-\rho)/\{\pi_0\pi_1(1-2\rho)^2+\rho(1-\rho)\}$. Now

$$\big(\Sigma + \alpha(\mu_1-\mu_0)(\mu_1-\mu_0)^T\big)^{-1} = \Sigma^{-1} - \frac{\alpha\Sigma^{-1}(\mu_1-\mu_0)(\mu_1-\mu_0)^T\Sigma^{-1}}{1+\alpha\Delta^2},$$

where $\Delta^2 = (\mu_1-\mu_0)^T\Sigma^{-1}(\mu_1-\mu_0)$. It follows that there exists an event Ω_0 with $\mathrm{pr}(\Omega_0)=1$ such that on this event, for every $x \in \mathbb{R}^d$,

$$\begin{aligned}
&\Big(x - \frac{\hat{\mu}_1+\hat{\mu}_0}{2}\Big)^T\hat{\Sigma}^{-1}(\hat{\mu}_1-\hat{\mu}_0) \\
&\quad\to \left[x - \frac{\mu_1}{2}\Big\{\frac{(1-2\rho)^2\pi_0\pi_1+2\rho(1-\rho)\pi_1}{(1-2\rho)^2\pi_0\pi_1+\rho(1-\rho)}\Big\} + \frac{\mu_0}{2}\Big\{\frac{(1-2\rho)^2\pi_0\pi_1+2\rho(1-\rho)\pi_0}{(1-2\rho)^2\pi_0\pi_1+\rho(1-\rho)}\Big\}\right]^T \\
&\qquad\Big(\Sigma^{-1} - \frac{\alpha\Sigma^{-1}(\mu_1-\mu_0)(\mu_1-\mu_0)^T\Sigma^{-1}}{1+\alpha\Delta^2}\Big)\Big\{\frac{(1-2\rho)\pi_0\pi_1}{(1-2\rho)^2\pi_0\pi_1+\rho(1-\rho)}\Big\}(\mu_1-\mu_0) \\
&\quad= \left[x - \frac{\mu_1}{2}\Big\{\frac{(1-2\rho)^2\pi_0\pi_1+2\rho(1-\rho)\pi_1}{(1-2\rho)^2\pi_0\pi_1+\rho(1-\rho)}\Big\} + \frac{\mu_0}{2}\Big\{\frac{(1-2\rho)^2\pi_0\pi_1+2\rho(1-\rho)\pi_0}{(1-2\rho)^2\pi_0\pi_1+\rho(1-\rho)}\Big\}\right]^T \\
&\qquad\Big(\frac{1}{1+\alpha\Delta^2}\Big)\Big\{\frac{(1-2\rho)\pi_0\pi_1}{(1-2\rho)^2\pi_0\pi_1+\rho(1-\rho)}\Big\}\Sigma^{-1}(\mu_1-\mu_0) \\
&\quad= \Big(x - \frac{\mu_1+\mu_0}{2}\Big)^T\Big(\frac{1}{1+\alpha\Delta^2}\Big)\Big\{\frac{(1-2\rho)\pi_0\pi_1}{(1-2\rho)^2\pi_0\pi_1+\rho(1-\rho)}\Big\}\Sigma^{-1}(\mu_1-\mu_0) \\
&\qquad - \left[\frac{\mu_1}{2}\Big\{\frac{(2\pi_1-1)\rho(1-\rho)}{(1-2\rho)^2\pi_0\pi_1+\rho(1-\rho)}\Big\} + \frac{\mu_0}{2}\Big\{\frac{(2\pi_0-1)\rho(1-\rho)}{(1-2\rho)^2\pi_0\pi_1+\rho(1-\rho)}\Big\}\right]^T \\
&\qquad\Big(\frac{1}{1+\alpha\Delta^2}\Big)\Big\{\frac{(1-2\rho)\pi_0\pi_1}{(1-2\rho)^2\pi_0\pi_1+\rho(1-\rho)}\Big\}\Sigma^{-1}(\mu_1-\mu_0).
\end{aligned}$$

Hence, on Ω_0,

$$\lim_{n\to\infty}\tilde{C}^{\mathrm{LDA}}(x) = \begin{cases} 1 \text{ if } c_0 + \big(x - \frac{\mu_1+\mu_0}{2}\big)^T\Sigma^{-1}(\mu_1-\mu_0) > 0 \\ 0 \text{ if } c_0 + \big(x - \frac{\mu_1+\mu_0}{2}\big)^T\Sigma^{-1}(\mu_1-\mu_0) < 0, \end{cases}$$

where

$$c_0 = \frac{(1+\alpha\Delta^2)\rho(1-\rho)}{\alpha(1-2\rho)}\log\biggl(\frac{(1-2\rho)\pi_1+\rho}{(1-2\rho)\pi_0+\rho}\biggr) - \frac{(\pi_1-\pi_0)\alpha\Delta^2}{2\pi_0\pi_1}.$$

This proves the first claim of the theorem. It follows that

$$\lim_{n\to\infty} R(\tilde{C}^{\mathrm{LDA}}) = \pi_0\Phi\biggl(\frac{c_0}{\Delta}-\frac{\Delta}{2}\biggr) + \pi_1\Phi\biggl(-\frac{c_0}{\Delta}-\frac{\Delta}{2}\biggr),$$

which proves the second claim. Now consider the function

$$\psi(c_0) = \pi_0\Phi\biggl(\frac{c_0}{\Delta}-\frac{\Delta}{2}\biggr) + \pi_1\Phi\biggl(-\frac{c_0}{\Delta}-\frac{\Delta}{2}\biggr).$$

We have

$$\dot{\psi}(c_0) = \frac{\pi_0}{\Delta}\phi\biggl(\frac{c_0}{\Delta}-\frac{\Delta}{2}\biggr) - \frac{\pi_1}{\Delta}\phi\biggl(-\frac{c_0}{\Delta}-\frac{\Delta}{2}\biggr) = \frac{\pi_0}{\Delta}\phi\biggl(\frac{c_0}{\Delta}-\frac{\Delta}{2}\biggr)\Bigl\{1-\frac{\pi_1}{\pi_0}\exp(-c_0)\Bigr\},$$

where ϕ denotes the standard normal density function. Since $\mathrm{sgn}\bigl(\dot{\psi}(c_0)\bigr) = \mathrm{sgn}\bigl(c_0 - \log(\pi_1/\pi_0)\bigr)$, we deduce that

$$\pi_0\Phi\biggl(\frac{c_0}{\Delta}-\frac{\Delta}{2}\biggr) + \pi_1\Phi\biggl(-\frac{c_0}{\Delta}-\frac{\Delta}{2}\biggr) \geq R(C^{\mathrm{Bayes}}),$$

and it remains to show that if $\rho \in (0,1/2)$ and $\pi_1 \neq \pi_0$, then there is a unique $\Delta > 0$ with $c_0 = \log(\pi_1/\pi_0)$. To that end, suppose without loss of generality that $\pi_1 > \pi_0$ and note that

$$\begin{aligned}\frac{(\pi_1-\pi_0)(1-2\rho)}{(1-2\rho)^2\pi_0\pi_1+\rho(1-\rho)} &= \frac{\pi_1(1-2\rho)+\rho}{(1-2\rho)^2\pi_1\pi_0+\rho(1-\rho)} - \frac{\pi_0(1-2\rho)+\rho}{(1-2\rho)^2\pi_1\pi_0+\rho(1-\rho)}\\ &= \frac{1}{(1-2\rho)\pi_0+\rho} - \frac{1}{(1-2\rho)\pi_1+\rho}.\end{aligned}$$

Hence, writing $t = (1-2\rho)\pi_1+\rho > 1/2$, we have

$$\log\Bigl(\frac{(1-2\rho)\pi_1+\rho}{(1-2\rho)\pi_0+\rho}\Bigr) - \frac{(\pi_1-\pi_0)(1-2\rho)}{2\{(1-2\rho)^2\pi_1\pi_0+\rho(1-\rho)\}} = \log\Bigl(\frac{t}{1-t}\Bigr) + \frac{1}{2t} - \frac{1}{2(1-t)} < 0.$$

Next, let

$$\begin{aligned}\chi(\pi_1) &= \log\Bigl(\frac{\pi_1}{\pi_0}\Bigr) - \frac{\rho(1-\rho)}{\alpha(1-2\rho)}\log\Bigl(\frac{(1-2\rho)\pi_1+\rho}{(1-2\rho)\pi_0+\rho}\Bigr)\\ &= \log\Bigl(\frac{\pi_1}{1-\pi_1}\Bigr) - \frac{(1-2\rho)^2\pi_1(1-\pi_1)+\rho(1-\rho)}{(1-2\rho)\pi_1(1-\pi_1)}\log\Bigl(\frac{(1-2\rho)\pi_1+\rho}{(1-2\rho)(1-\pi_1)+\rho}\Bigr).\end{aligned}$$

Then

$$\dot{\chi}(\pi_1) = \frac{\rho(1-\rho)(1-2\pi_1)}{(1-2\rho)\pi_1^2(1-\pi_1)^2}\log\Bigl(\frac{(1-2\rho)\pi_1+\rho}{(1-2\rho)(1-\pi_1)+\rho}\Bigr) < 0,$$

for all $\pi_1 \in (0,1)$. Since $\chi(1/2) = 0$, we conclude that $\chi(\pi_1) < 0$ for all $\pi_1 > \pi_0$. But

$$c_0 - \log\Bigl(\frac{\pi_1}{\pi_0}\Bigr) = \frac{\Delta^2\rho(1-\rho)}{1-2\rho}\Bigl\{\log\Bigl(\frac{(1-2\rho)\pi_1+\rho}{(1-2\rho)\pi_0+\rho}\Bigr) - \frac{(\pi_1-\pi_0)(1-2\rho)}{2\{(1-2\rho)^2\pi_1\pi_0+\rho(1-\rho)\}}\Bigr\} - \chi(\pi_1),$$

so the final claim follows. □

REFERENCES

Cannings, T. I., Berrett, T. B. & Samworth, R. J. (2018) Local nearest neighbour classification with applications to semi-supervised learning. *ArXiv e-prints*, 1704.00642.

Devroye, L., Györfi, L. & Lugosi, G. (1996) *A Probabilistic Theory of Pattern Recognition.* Springer, New York.
Shorack, G. R. & Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics*. Wiley, New York.
Steinwart, I. (2005) Consistency of support vector machines and other regularized kernel classifiers. *IEEE Trans. Inf. Th.*, **51**, 128–142.
Steinwart, I. & Christmann, A. (2008) *Support Vector Machines.* Springer, New York.
Steinwart, I. & Scovel, C. (2007) Fast rates for support vector machines using Gaussian kernels. *Ann. Statist.*, **35**, 575–607.