A Derivation of the Generalized Bayes Classifier

In this Appendix, we prove Theorem 3, in which we characterize a stochastic generalization of the Bayes classifier to arbitrary CMMs. We reiterate the theorem for the reader here:

Theorem 3. *If* $\operatorname{argmax}_{\widehat{Y}\in\mathcal{SC}} M(C_{\widehat{Y}}) \neq \varnothing$*, then there exists a regression-thresholding classifier*

$$\widehat{Y}_{p,t,\eta} \in \operatorname*{argmax}_{\widehat{Y}\in\mathcal{SC}} M(C_{\widehat{Y}}).$$

As described in the main paper, unlike prior results [Koyejo et al., 2014, Yan et al., 2018, Wang et al., 2019b], we do not assume that the distribution of $\eta(X)$ is absolutely continuous. This makes proving Theorem considerably more complicated than these previous results. We prove Theorem 3 in a sequence of steps, constructing optimal classifiers in forms progressively closer to that of the generalized Bayes classifier described in Theorem 3. Specifically, we first show, in Lemma 18, that there exists an optimal classifier that is a (stochastic) function of the true regression function η^*. We then construct an optimal classifier in which this function of η^* is non-decreasing. Finally, we construct an optimal classifier in which this function of η^* is a threshold function, as in Theorem 3.

Lemma 18. *For any stochastic classifier* $\widehat{Y} : \mathcal{X} \to \mathcal{B}$*, there is a stochastic classifier* $\widehat{Y}' : \mathcal{X} \to \mathcal{B}$ *of the form*

$$\widehat{Y}'(x) \sim \textit{Bernoulli}(f(\eta^*(x))), \tag{6}$$

for some $f : [0,1] \to [0,1]$*, such that* $C_{\widehat{Y}'} = C_{\widehat{Y}}$.

Proof. We start by defining the regression function and discussing formal probability notation. Let $\mathcal{F} = \sigma(\eta^*(Z))$ be the σ-field generated by the true regression function $\eta^*(Z)$. Define $\widehat{Y}' : \mathcal{X} \to \mathcal{B}$ by

$$\widehat{Y}'(x) \sim \text{Bernoulli}\left(\mathop{\mathbb{E}}_{Z\sim P_X}\left[\widehat{Y}(Z)\middle|\mathcal{F}\right](\eta^*(x))\right),$$

where by the $\eta^*(x)$ we are explicitly indicating that $\eta^*(x)$ is the input to the conditional expectation, as the conditional expectation is a measurable function of $\eta^*(Z)$. In the sequel, following standard conventions, we omit such notation. Note that $\widehat{Y}'$ has the desired form and is defined almost surely.

Now, we get our final notes in place before proceeding with calculations. Let $\mathcal{G} = \sigma(X)$ be the σ-field generated by X.

A key step in the proof is showing the equality

$$\mathbb{E}\left[Y\widehat{Y}'(X)\middle|\mathcal{F}\right] = \eta^*(X)\,\mathbb{E}\left[\widehat{Y}(X)\middle|\mathcal{F}\right] = \mathbb{E}\left[Y\widehat{Y}(X)\middle|\mathcal{F}\right]. \tag{7}$$

We first provide an intuitive summary. Observe that conditioned on $\eta^*(X)$, Y is independent of $\widehat{Y}$ and $\widehat{Y}'$ separately. Thus, we can integrate over Y and get the $\eta^*(X)$. By the definition of $\widehat{Y}'$ we obtain the conditional expectation with respect to Y, and we have the resulting equality.

For the time being, taking Equation (7) as fact, we complete the rest of the proof. Combining Equation (7) and the tower property of conditional expectation, we have

$$\text{TP}_{\widehat{Y}'} = \mathbb{E}\left[Y\widehat{Y}'(X)\right] = \mathbb{E}\left[\mathbb{E}\left[Y\widehat{Y}'(X)\middle|\mathcal{F}\right]\right] = \mathbb{E}\left[\mathbb{E}\left[Y\widehat{Y}(X)\middle|\mathcal{F}\right]\right] = \mathbb{E}\left[Y\widehat{Y}(X)\right] = \text{TP}_{\widehat{Y}}.$$

Similarly, one can check that $C_{\widehat{Y}'} = C_{\widehat{Y}}$. □

Proof of Equation (7). We now present the formal details to establish Equation (7), which may be skipped if one is uninterested in the measure-theoretic details. We first to set up the random variables needed to formalize the above intuition. Let S, T, U, V, and W be random variables. Let $(S,T) = (\eta^*(X), X)$. We also require U and V to be uniform on $[0,1]$. We need not specify the precise distribution W. Further, let (S,T), U, V, and W all be independent. For notation purposes, it is convenient for the underlying probability space $(\Omega, \mathcal{G}, \mathbb{P})$ to be the product probability space of some probability spaces to the extent possible. In particular, we require the set $\Omega = \times_{i=S}^{W} \Omega_i$ and a set B in $\mathcal{G}$ to have a product structure $B = \times_{i=S}^{W} B_i$ for ease of exposition. Note that the measure $\mathbb{P}$

is not a product measure, as S and T are dependent, although it is a product measure with respect to the laws of (S,T), U, V, W.

Now, we define Y, $\widehat{Y}'$, and $\widehat{Y}$ in terms of U, V, W, and X. Define

$$\begin{aligned} Y(\eta^*(X)) &= \mathbf{1}\{U \leqslant \eta^*(X)\} \\ \widehat{Y}'(\eta^*(X)) &= \mathbf{1}\{V \leqslant \mathbb{E}[\widehat{Y}|\mathcal{F}]\} \\ \widehat{Y}(X) &= r(S,T,W), \end{aligned}$$

where here U serves as the noise in Y, V serves as the internal randomization of $\widehat{Y}'$, and W serves as the possible internal randomization of $\widehat{Y}$. Here, $\mathbf{1}$ is the indicator taking the value 1 if the event in curly braces occurs and 0 otherwise, and r is some function. Note that this is still of the desired form.

To prove the first part of Equation (7), i.e., to show that the middle term $M = \eta^*(X)\,\mathbb{E}[\widehat{Y}(X)|\mathcal{F}] = S\,\mathbb{E}[\widehat{Y}|\mathcal{F}]$ is indeed the conditional expectation of the left of Equation (7), we must verify two conditions: (i) M is $\mathcal{F}$-measurable and (ii) the integrals of $Y\widehat{Y}'$ and M are identical on any event A in $\mathcal{F}$ [Durrett, 2010, page 221]. Condition (i) is immediate, since M is a function of $\eta^*(X)$ and nothing more.

For condition (ii), we do a bit of computation. Let $A = A_S \times A_T \times A_U \times A_V \times A_W$ be an event in $\mathcal{F}$. Note that because A is $\mathcal{F}$-measurable, the sets A_U and A_V must be all of Ω_U and Ω_V or the empty set. Suppose for the moment that both are non-empty. Then, we have

$$\begin{aligned} \int_A Y\widehat{Y}'d\mathbb{P} &= \int_{A_S}\int_{A_V}\int_{A_U} \mathbf{1}\{U \leqslant S\}\,\mathbf{1}\{V \leqslant \mathbb{E}[\widehat{Y}|\mathcal{F}]\}d\mathbb{P}_U d\mathbb{P}_V d\mathbb{P}_S \\ &= \int_{A_S}\int_{A_V} S\mathbf{1}\{V \leqslant \mathbb{E}[\widehat{Y}|\mathcal{F}]\}d\mathbb{P}_V d\mathbb{P}_S \\ &= \int_{A_S} S\,\mathbb{E}[\widehat{Y}|\mathcal{F}]d\mathbb{P}_S \\ &= \int_A S\,\mathbb{E}[\widehat{Y}|\mathcal{F}]d\mathbb{P}. \end{aligned}$$

Note that the crucial first and last equalities in which we break and reform the integral over the entire event in terms of its coordinates is possible due to Fubini's theorem for integrable functions. This proves the desired equality when $A_U = \Omega_U$ and $A_V = \Omega_V$, and from the preceding calculation we can see that the integrals are both 0 when either $A_U = \varnothing$ or $A_V = \varnothing$, and so the desired equality holds. This establishes the left equality of Equation (7).

Establishing the right hand side of Equation (7) also takes a bit of calculation. First, we have already verified the measurability condition (i), and so all that remains is to check the integral equality. Again, let A be an event in $\mathcal{F}$, and assume that A_T, A_U, and A_W are non-empty. We have

$$\begin{aligned} \int_A Y\widehat{Y}d\mathbb{P} &= \int_{A_S}\int_{A_T}\int_{A_W}\int_{A_U} \mathbf{1}\{U \leqslant S\}\,r(S,T,W)d\mathbb{P}_U d\mathbb{P}_W d\mathbb{P}_T d\mathbb{P}_S \\ &= \int_{A_S}\int_{A_T}\int_{A_W} S\,r(S,T,W)d\mathbb{P}_W d\mathbb{P}_T d\mathbb{P}_S \\ &= \int_{A_S} S\left(\int_{A_T}\int_{A_W} r(S,T,W)d\mathbb{P}_W d\mathbb{P}_T\right)d\mathbb{P}_S. \end{aligned}$$

Note that Fubini's theorem is again used in the first step. In the event that any of A_T, A_U, or A_W is empty, then the above equation is 0 and equality holds. Now, we have to show that the random variable $M' = \int_{A_T}\int_{A_W} r(S,T,W)d\mathbb{P}_W d\mathbb{P}_{T|S}$ is the conditional expectation of $\widehat{Y}$ with respect to $\mathcal{F}$. Measurability is readily apparent, as M' is a function of S and no more. Now, we check the condition that $\widehat{Y}$ and M' have the same integral on an event A in $\mathcal{F}$. Again assume that A_U and A_W are non-empty, observing in the calculation to follow that equality holds with the value 0 if either is

empty. Using Fubini's theorem and some direct computation, we have

$$\begin{aligned}\int_A \widehat{Y} d\mathbb{P} &= \int_A r(S,T,W) d\mathbb{P} \\ &= \int_{A_S}\int_{A_T}\int_{A_W} r(S,T,W) d\mathbb{P}_W d\mathbb{P}_T d\mathbb{P}_S \\ &= \int_A \left(\int_{A_T}\int_{A_W} r(S,T,W) d\mathbb{P}_W d\mathbb{P}_T\right) d\mathbb{P} \\ &= \int_A M' d\mathbb{P}.\end{aligned}$$

This completes the proof that the conditional expectation of $\widehat{Y}$ is M', and so it proves that the conditional expectation of $Y\widehat{Y}$ is $\eta^*(X)\,\mathbb{E}[\widehat{Y}|\mathcal{F}]$. This is the right equality of Equation (7), thus completing the proof. □

It follows from Lemma 18 that, if $M(C_{\widehat{Y}})$ is maximized by any stochastic classifier, then it is maximized by a classifier $\widehat{Y}$ of the form in Eq. (6). It remains to show that f in Eq. (6) can be of the form $z \mapsto p1\{z = t\} + 1\{z > t\}$ for some threshold $(p, t) \in [0, 1]^2$. Before proving this, we give a simplifying lemma showing that the problem of maximizing a CMM can be equivalently framed as a particular functional optimization problem. This will allow us to to significantly simplify the notation in the subsequent proofs.

Lemma 19. *Let M be a CMM, and suppose that $M(C_{\widehat{Y}})$ is maximized (over $\mathcal{SC}$) by a classifier $\widehat{Y}$ of the form*

$$\widehat{Y}(x) \sim \text{Bernoulli}(f(\eta(x))),$$

for some $f^* : [0,1] \to [0,1]$. Let f be a solution to the optimization problem

$$\max_{f:[0,1]\to[0,1]} \mathbb{E}[\eta(X)f(\eta(X))] \;\; \textit{s.t.} \;\; \mathbb{E}[(1-\eta(X))f(\eta(X))] \leqslant \mathbb{E}[(1-\eta(X))f^*(\eta(X))]. \tag{8}$$

Then, the classifier

$$\widehat{Y}'(x) \sim \text{Bernoulli}(f(\eta(x))),$$

also maximizes $M(C_{\widehat{Y}'})$ (over $\mathcal{SC}$).

Proof. This result follows from the definition (Definition 1) of a CMM. Specifically, by construction of $\widehat{Y}'$,

$$\text{TP}_{\widehat{Y}'} = \mathbb{E}[\eta(X)f(\eta(X))] \geqslant \mathbb{E}[\eta(X)f^*(\eta(X))] = \text{TP}_{\widehat{Y}}$$

and

$$\text{FP}_{\widehat{Y}'} = \mathbb{E}[(1-\eta(X))f(\eta(X))] \leqslant \mathbb{E}[(1-\eta(X))f^*(\eta(X))] = \text{FP}_{\widehat{Y}}.$$

Moreover, since the proportions of positive and negative true labels are independent of the chosen classifier (i.e., $\text{TP}_{\widehat{Y}'} + \text{FN}_{\widehat{Y}'} = \text{TP}_{\widehat{Y}} + \text{FN}_{\widehat{Y}}$ and $\text{FP}_{\widehat{Y}'} + \text{TN}_{\widehat{Y}'} = \text{FP}_{\widehat{Y}} + \text{TN}_{\widehat{Y}}$), we have

$$C_{\widehat{Y}'} = \begin{bmatrix} \text{TN}_{\widehat{Y}} + \epsilon_1 & \text{FP}_{\widehat{Y}} - \epsilon_1 \\ \text{FN}_{\widehat{Y}} - \epsilon_2 & \text{TP}_{\widehat{Y}} + \epsilon_2, \end{bmatrix},$$

where $\epsilon_1 := \text{FP}_{\widehat{Y}} - \text{FP}_{\widehat{Y}'} \in [0, \text{FP}_{\widehat{Y}}]$ and $\epsilon_2 := \text{TP}_{\widehat{Y}'} - \text{TP}_{\widehat{Y}} \in [0, \text{FN}]$. Thus, by the definition (Definition 1) of a CMM, $M(C_{\widehat{Y}'}) \geqslant M(C_{\widehat{Y}})$. □

Lemma 19 essentially shows that maximizing any CMM M is equivalent to performing Neyman-Pearson classification, at some particular false positive level α depending on M (through f^*) and on the distribution of $\eta(X)$. For our purposes, this simplifies the remaining steps in proving Theorem 3 by allowing us to ignore the details of the particular CMM M and regression function η and focus on characterizing solutions to an optimization problem of the form (8) (see, specifically, (9) below).

To characterize solutions to this optimization problem, we will utilize the following two measure-theoretic technical lemmas:

Lemma 20. *Let μ be a measure on $[0,1]$ with $\mu([0,1]) > 0$. Then, there exists $z \in \mathbb{R}$ such that, for all $\epsilon > 0$, $\mu([0,1] \cap (z-\epsilon, z)) > 0$.*

Proof. We prove the contrapositive. Suppose that, for every $z \in [0,1]$, there exists $\epsilon_z > 0$ such that $\mu([0,1] \cap (z - \epsilon_z, z)) = 0$. The family $\mathcal{S} := \{[0,1] \cap (z - \epsilon_z, z) : z \in \mathbb{R}\}$ is an open cover of $[0,1]$. Since $[0,1]$ is compact, there exists a finite sub-cover $\mathcal{S}' \subseteq \mathcal{S}$ of $[0,1]$. Thus, by countable subaddivity of measures,

$$\mu([0,1]) \leqslant \sum_{S \in \mathcal{S}'} \mu(\mathcal{S}) = 0.$$

□

Lemma 21. *Let $(\mathcal{X}, \Sigma, \mu)$ be a measure space, let $E, F \in \Sigma$ be measurable sets, and let $f : \mathcal{X} \to \mathbb{R}$ be a Σ-measurable function. If*

$$\operatorname*{ess\,sup}_{\mu} f(E) > \operatorname*{ess\,inf}_{\mu} f(F),$$

then there exist measurable sets $A \subseteq E$ and $B \subseteq F$ with $\mu(A), \mu(B) > 0$, and

$$\inf_{x \in A} f(x) > \sup_{x \in B} f(x).$$

Proof. If $\operatorname{ess\,sup}_\mu f(E) > \operatorname{ess\,inf}_\mu f(F)$, then there exist $a > b$ such that $\operatorname{ess\,sup}_\mu f(E) > a > b > \operatorname{ess\,inf}_\mu f(F)$. Since $\operatorname{ess\,sup}_\mu f(E) > a$, it follows that $P_Z(E \cap \{z : f(z) \geqslant a\}) > 0$. Similarly, since $\operatorname{ess\,inf}_\mu f(F) < b$, it follows that $P_Z(F \cap \{z : f(z) \leqslant b\}) > 0$. Hence, letting $A = E \cap \{z : f(z) \geqslant a\}$ and $B = F \cap \{z : f(z) \leqslant b\}$, we have

$$\inf_{z \in A} f(z) \geqslant a > b \geqslant \sup_{z \in B} f(z).$$

□

We are now ready for the main remaining step in the proof of Theorem 3, namely characterizing solutions of (a generalization of) the optimization problem (8):

Lemma 22. *Let Z be a $[0,1]$-valued random variable, and let $c \in [0,1]$. Suppose that the optimization problem*

$$\max_{f:[0,1]\to[0,1] \text{ measurable}} \mathbb{E}[Zf(Z)] \quad \text{subject to} \quad \mathbb{E}[(1-Z)f(Z)] \leqslant c \tag{9}$$

has a solution. Then, there is a solution to (9) *that is a stochastic threshold function.*

Proof. Suppose that there exists a solution f to (9). We will construct a stochastic threshold function that solves (9) in two main steps. First, we will construct a monotone solution to (9). Second, we will show that this monotone solution is equal to a stochastic threshold function except perhaps on a set of probability 0 with respect to Z. This stochastic threshold function is therefore a solution to (9).

Construction of Monotone Solution to (9)**:** Define

$$g(z) := \operatorname*{ess\,sup}_{P_Z} f([0,z]) \quad \text{and} \quad h(z) := \operatorname*{ess\,inf}_{P_Z} f((z,1]),$$

where the essential supremum and infimum are taken with respect to the measure P_Z of Z, with the conventions $g(z) = 0$ whenever $P_Z([0,z]) = 0$ and $h(z) = 1$ whenever $P_Z((z,1]) = 0$. We first show that, for all $z \in [0,1]$, $g(z) \leqslant h(z)$. We will then use this to show that $g = f$ except on a set of P_Z measure 0 (i.e., $P_Z(\{z \in [0,1] : g(z) \neq f(z)\}) = 0$). Therefore, both $\mathbb{E}[Zg(Z)] = \mathbb{E}[Zf(Z)]$ and $\mathbb{E}[(1-Z)g(Z)] = \mathbb{E}[(1-Z)f(Z)]$. Since $g : [0,1] \to [0,1]$ is clearly monotone non-decreasing, the result follows.

Suppose, for sake of contradiction, that, for some $z \in [0,1]$, $g(z) > h(z)$. By Lemma 21, there exist $A \subseteq [0,z]$ and $B \subseteq (z,1]$ such that $\inf_{z\in A} f(z) > \sup_{z\in B} f(z)$ and $P_Z(A), P_Z(B) > 0$. Define $z_A := \mathbb{E}[Z|Z \in A]$ and $z_B := \mathbb{E}[Z|Z \in B]$, and note that, since $A \subseteq [0,z]$ and $B \subseteq (z,1]$, $z_A < z_B$. Define,

$$\epsilon := \min\left\{ \frac{P_Z(A)(1-z_A)}{P_Z(B)(1-z_B)} \inf_{z\in A} f(z), \quad 1 - \sup_{z\in B} f(z) \right\} > 0$$

and define $\phi : [0,1] \to [0,1]$ by

$$\phi(z) := \begin{cases} f(z) - \epsilon \frac{P_Z(B)(1-z_B)}{P_Z(A)(1-z_A)} & \text{if } z \in A \\ f(z) + \epsilon & \text{if } z \in B \\ f(z) & \text{otherwise,} \end{cases}$$

noting that, by construction of ϵ, $\phi(z) \in [0,1]$ for all $z \in [0,1]$. Then, by construction of ϕ,

$$\begin{aligned} \mathbb{E}[(1-Z)\phi(Z)] - \mathbb{E}[(1-Z)f(Z)] &= -(1-z_A)\epsilon \frac{P_Z(B)(1-z_B)}{P_Z(A)(1-z_A)} P_Z(A) + (1-z_B)\epsilon P_Z(B) \\ &= 0 \cdot \epsilon = 0, \end{aligned}$$

while

$$\begin{aligned} \mathbb{E}[Z\phi(Z)] - \mathbb{E}[Zf(Z)] &= -z_A \epsilon \frac{P_Z(B)(1-z_B)}{P_Z(A)(1-z_A)} P_Z(A) + z_B \epsilon P_Z(B) \\ &= \left(-\frac{z_A}{1-z_A}(1-z_B) + z_B \right) \epsilon P_Z(B) > 0, \end{aligned}$$

since the function $z \mapsto \frac{z}{1-z}$ is strictly increasing. This contradicts the assumption that f optimizes (9), implying $g \leqslant h$.

We now show that $g = f$ except on a set of P_Z measure 0. First, note that, if $g(z) \neq f(z)$, then $g(z) = \text{ess sup}\, f([0,z]) = \text{ess sup}\, f([0,z))$, and so g is left-continuous at z.

For any $\delta > 0$, define

$$A_\delta := \{z \in [0,1] : g(z) < f(z) - \delta\} \quad \text{and} \quad B_\delta := \{z \in [0,1] : g(z) > f(z) + \delta\}.$$

Since

$$\{z \in [0,1] : g(z) < f(z)\} = \bigcup_{j=1}^{\infty} \left\{ z \in [0,1] : g(z) < f(z) - \frac{1}{j} \right\}$$

and

$$\{z \in [0,1] : g(z) > f(z)\} = \bigcup_{j=1}^{\infty} \left\{ z \in [0,1] : g(z) > f(z) + \frac{1}{j} \right\},$$

by countable subadditivity, it suffices to show that $P_Z(A_\delta) = P_Z(B_\delta) = 0$ for all $\delta > 0$.

Suppose, for sake of contradiction, that $P_Z(A_\delta) > 0$. Applying Lemma 20 to the measure $E \mapsto P_Z(A_\delta \cap E)$, there exists $z \in \mathbb{R}$ such that, for any $\epsilon > 0$, $P_Z(A_\delta \cap (z-\epsilon, z)) > 0$. Since g is continuous at z, there exists $\epsilon > 0$ such that $g(z-\epsilon) \geqslant g(z) - \delta$, so that, for all $z \in A_\delta \cap (z-\epsilon, z)$, $f(z) > g(z) + \delta$. Then, since $P_Z(A_\delta \cap (z-\epsilon, z)) > 0$, we have the contradiction

$$g(z) \geqslant \text{ess sup}\, f(A_\delta \cap (z-\epsilon, z)) > g(z).$$

On the other hand, suppose, for sake of contradiction, that $P_Z(B_\delta) > 0$. Applying Lemma 20 to the measure $E \mapsto P_Z(B_\delta \cap E)$, there exists $z \in \mathbb{R}$ such that, for any $\epsilon > 0$, $P_Z(B_\delta \cap (z-\epsilon, z)) > 0$. Since g is continuous at z, there exists $\epsilon > 0$ such that $g(z-\epsilon) \geqslant g(z) - \delta$. At the same time, since g is non-decreasing, for $t \in B_\delta \cap (z-\epsilon, z)$, $f(t) < g(t) - \delta \leqslant g(z) - \delta$. Thus, since $P_Z(B_\delta \cap (z-\epsilon, z)) > 0$, we have $h(z-\epsilon) < g(z) - \delta < g(z-\epsilon)$, contradicting the previously shown fact that $g \leqslant h$.

To conclude, we have shown that $P_Z(\{z \in [0,1] : g(z) \neq f(z)\}) = 0$.

Construction of a Stochastic Threshold Solution: We now construct a solution to (9) that is equal to a stochastic threshold function (i.e., a function that has the form $p1\{z = t\} + 1\{z > t\}$) except on a set of P_Z-measure 0. To show this, it suffices to construct a function $f : [0,1] \to [0,1]$ such that (a) f is monotone non-decreasing and (b) the set $f^{-1}((0,1))$ is the union of the singleton $\{t\}$ and a set of P_Z-measure 0.

From the previous step of this proof, we may assume that we have a solution f to (9) that is monotone non-decreasing. It suffices therefore to show that $A := f^{-1}((0,1))$ is the union of a singleton and a set of P_Z-measure 0. Define

$$t_0 := \inf\{z \in [0,1] : P_Z(A \cap [0,z]) > 0\} \quad \text{and} \quad t_1 := \sup\{z \in [0,1] : P_Z(A \cap [z,1]) > 0\}.$$

Then, for all $\epsilon > 0$, $P_Z(A \cap [0, t_0 - \epsilon]) = P_Z(A \cap [t_1 + \epsilon, 1]) = 0$. Hence, if $t_0 = t_1$, then, since

$$A \backslash \{t_0\} = \bigcup_{j=1}^{\infty} A \cap ([0, t_0 - 1/j] \cup [t_1 + 1/j, 1])$$

by countable subadditivity, $P_Z(A \backslash \{t_0\}) = 0$, which implies that $A = \{t_0\} \cup (A \backslash \{t_0\})$ is the union of a singleton and a set of measure 0.

It suffices therefore to prove that $t_0 = t_1$. It is easy to see, from the definitions of t_0 and t_1, that $t_0 \leqslant t_1$. Suppose, for sake of contradiction, that $t_0 < t_1$. Then, there exists $t \in (t_0, t_1)$, and, by definition of t_0 and t_1, both $P_Z(A \cap [0, t)) > 0$ and $P_Z(A \cap (t, 1]) > 0$. For any $\delta \geqslant 0$, define

$$B_\delta := \{z \in [0, t) : \delta < f(z) < 1 - \delta\} \quad \text{and} \quad C_\delta := \{z \in (t, 1] : \delta < f(z) < 1 - \delta\},$$

so that $P_Z(B_0) > 0$ and $P_Z(C_0) > 0$. By countable subadditivity, there exists $\delta > 0$ such that $P_Z(B_\delta) > 0$ and $P_Z(C_\delta) > 0$.

Define $\epsilon := \delta \cdot \min\{P_Z(B_\delta), P_Z(C_\delta)\} > 0$. Define $g : [0, 1] \to \mathbb{R}$ for all $z \in [0, 1]$ by

$$g(z) = \begin{cases} f(z) - \frac{\epsilon}{P_Z(B_\delta)} & \text{if } z \in B_\delta \\ f(z) + \frac{\epsilon}{P_Z(C_\delta)} & \text{if } z \in C_\delta \\ f(z) & \text{otherwise.} \end{cases},$$

and note that, by definition of ϵ, B_δ, and C_δ, $g : [0, 1] \to [0, 1]$. Then,

$$\mathbb{E}[g(Z)] - \mathbb{E}[f(Z)] = -\frac{\epsilon}{P_Z(B_\delta)} P_Z(B_\delta) + \frac{\epsilon}{P_Z(C_\delta)} P_Z(C_\delta) = 0,$$

while

$$\begin{aligned} \mathbb{E}[Zg(Z)] - \mathbb{E}[Zf(Z)] &= -\mathbb{E}[Z|Z \in B_\delta] \frac{\epsilon}{P_Z(B_\delta)} P_Z(B_\delta) + \mathbb{E}[Z|Z \in C_\delta] \frac{\epsilon}{P_Z(C_\delta)} P_Z(C_\delta) \\ &= \epsilon \left(\mathbb{E}[Z|Z \in C_\delta] - \mathbb{E}[Z|Z \in B_\delta]\right). \end{aligned}$$

Since $B_\delta \subseteq [0, t)$ and $C_\delta \subseteq (t, 1]$, this difference is strictly positive, contradicting the assumption that f optimizes (9). □

Combining Lemma 22 with Lemma 19 completes the proof of our main result, Theorem 3.

A.1 Extension to AUROC

For any regression function η, the receiver operating characteristic (ROC) function $\mathrm{ROC}_\eta : [0, 1] \to [0, 1]$ is

$$\mathrm{ROC}_\eta(x) := \sup_{(p,t) \in [0,1]^2} \mathrm{TP}_{\widehat{Y}_{p,t,\eta}} \cdot 1\{\mathrm{FP}_{\widehat{Y}_{p,t,\eta}} \leqslant x\} \quad \text{for all} \quad x \in [0, 1], \tag{10}$$

i.e., $\mathrm{ROC}(x)$ is the maximum true positive probability (over all regression-thresholding classifiers with regression function η) achievable while keeping the false positive probability below x. The area under the ROC curve (AUROC) is then given by

$$\mathrm{AUROC}_\eta = \int_0^1 \mathrm{ROC}_\eta(x)\, dx. \tag{11}$$

While AUROC is not a CMM (as it depends on the entire family of confusion matrices computed at all possible thresholds $(p, t) \in [0, 1]^2$), AUROC is widely used to measure performance of classifiers across the classification thresholds. Here, we show that our Theorem 3 extends naturally from CMMs to AUROC.

We begin by noting that, for any $x \in [0, 1]$, the performance measure $\mathrm{TP} \cdot 1\{\mathrm{FP} \leqslant x\}$ is a CMM. Therefore, letting η^* denote the true regression function, by Theorem 3, there exists a threshold $(p, t) \in [0, 1]^2$ such that the regression-thresholding classifier $\widehat{Y}_{p,t,\eta^*} \in \mathrm{argmax}_{\widehat{Y} \in \mathcal{SC}} \mathrm{TP} \cdot 1\{\mathrm{FP} \leqslant x\}$; i.e., $\widehat{Y}_{p,t,\eta^*}$ maximizes $\mathrm{TP} \cdot 1\{\mathrm{FP} \leqslant x\}$ over all stochastic classifiers. By definition of ROC (Eq. (10)), it follows that, for any $x \in [0, 1]$,

$$\eta^* \in \operatorname*{argmax}_{\eta : \mathcal{X} \to [0,1]} \mathrm{ROC}_\eta(x),$$

and, by definition AUROC (Eq. (11)), it then follows that

$$\eta^* \in \underset{\eta:\mathcal{X}\to[0,1]}{\operatorname{argmax}} \operatorname{AUROC}_\eta .$$

To conclude, we have shown that thresholding the true regression function is optimal not only under any CMM but also under AUROC. A identical argument can be made for other performance measures, such as the area under the precision-recall curve (AUPRC), that aggregate CMMs across multiple classification thresholds.

B Relative Performance Guarantees in terms of the Generalized Bayes Classifier

In this Appendix, we prove Lemmas 5 and 6, as well as their consequence, Corollary 8. Also, in Section B.1, we demonstrate, in a few key examples, how to compute the Lipschitz constant used in Corollary 8.

We begin with the proof of Lemma 5, which, at a given threshold (p, t), bounds the difference between the confusion matrices of the true regression function η and an estimate η' of η. We restate the result for the reader's convenience:

Lemma 5. *Let $p, t \in [0, 1]$ and let $\eta, \eta' : \mathcal{X} \to [0, 1]$. Then,*

$$\left\| C_{\widehat{Y}_{p,t,\eta}} - C_{\widehat{Y}_{p,t,\eta'}} \right\|_\infty \leqslant \mathbb{P}\left[|\eta(X) - t| \leqslant \|\eta - \eta'\|_\infty \right]. \tag{12}$$

Proof. For the true negative probability, we have

$$\begin{aligned} \left| \mathrm{TN}_{\widehat{Y}_{p,t,\eta}} - \mathrm{TN}_{\widehat{Y}_{p,t,\eta'}} \right| &= \left| \mathbb{P}\left[Y = 0, \eta'(X) \leqslant t < \eta(X) \right] - \mathbb{P}[Y = 0, \eta(X) \leqslant t < \eta'(X)] \right| \\ &\leqslant \mathbb{P}\left[|\eta(X) - t| \leqslant \|\eta - \eta'\|_\infty \right]. \end{aligned}$$

This type of inequality is standard and follows from the fact that, if t lies between η and η', then the difference of η and t is necessarily less than η and η'. Repeating this calculation for the true positive, false positive, and false negative probabilities gives (12). □

Note that, in the presence of degree r Uniform Class Imbalance (see Section 5), one can obtain a tighter error bound $r\mathbb{P}\left[|\eta(X) - t| \leqslant \|\eta - \eta'\|_\infty \right]$ for the true positive and false negative probabilities because, for all $x \in \mathcal{X}$, $\mathbb{P}[Y = 1|X = x] \leqslant r$. However, the weaker bound (12) simplifies the exposition.

We now turn to proving Lemma 6, which we use to bound the maximum difference between the empirical and true confusion matrices of a regression-thresholding classifier over thresholds (p, t). Specifically, we will use this result to bound the difference in confusion matrices between the optimal threshold (p^*, t^*) and the threshold $(\widehat{p}, \widehat{t})$ selected by maximizing the empirical CMM. We actually prove a more general version of Lemma 6, for arbitrary classifiers, based on the following definition:

Definition 23 (Stochastic Growth Function)**.** *Let $\mathcal{F}$ be a family of $[0, 1]$-valued functions on $\mathcal{X}$. The* stochastic growth function $\Pi_\mathcal{F} : \mathbb{N} \to \mathbb{N}$*, defined by*

$$\Pi_\mathcal{F}(n) := \max_{\substack{x_1,\ldots,x_n\in\mathcal{X},\\ z_1,\ldots,z_n\in[0,1]}} \left| \{ (1\{f(x_i) > z_i\})_{i=1}^n : f \in \mathcal{F} \} \right| \quad \textit{for all} \quad n \in \mathbb{N},$$

is the maximum number of distinct classifications of n points $x_1, ..., x_n$ by a stochastic classifier $\widehat{Y}$ with $(x \mapsto \mathbb{E}[\widehat{Y}(x)]) \in \mathcal{F}$ and randomness given by $z_1, ..., z_n$.

Definition 23 generalizes the growth function [Mohri et al., 2018], a classical measure of the complexity of a hypothesis class originally due to Vapnik and Chervonenkis [2015], to non-deterministic classifiers. Importantly for our purposes, one can easily bound the stochastic growth function of regression-thresholding classifiers:

Example 24 (Stochastic Growth Function of Regression-Thresholding Classifiers)**.** Suppose

$$\mathcal{F} = \{f : \mathcal{X} \to [0,1] | \text{ for some } p, t \in [0,1], f(x) = p \cdot 1\{\eta(x) = t\} + 1\{\eta(x) > t\} \text{ for all } x \in \mathcal{X}\},$$

so that $\{\widehat{Y}_{f,\eta} : f \in \mathcal{F}\}$ is the class of regression-thresholding classifiers. Any set of points $(x_1, z_1), ..., (x_n, z_n)$, can be sorted in increasing order by $\eta(x)$'s, breaking ties in decreasing order by z's. Having sorted the points in this way, $\{f(x) > z\} = 0$ for the first j points and $\{f(x) > z\} = 1$ for the remaining $n - j$ points, for some $j \in [n] \cup \{0\}$. Thus, $\Pi_{\mathcal{F}}(n) = n + 1$.

We will now prove the following result, from which, together with Example 24, Lemma 6 follows immediately:

Lemma 6 (Generalized Version)**.** *Let $\mathcal{F}$ be a family of $[0,1]$-valued functions on $\mathcal{X}$. Then, with probability at least $1 - \delta$,*

$$\sup_{f \in \mathcal{F}} \left\| \widehat{C}_{\widehat{Y}_f} - C_{\widehat{Y}_f} \right\|_\infty \leqslant \sqrt{\frac{8}{n} \log \frac{32 \Pi_{\mathcal{F}}(2n)}{\delta}}.$$

Before proving Lemma 6, we note a standard symmetrization lemma, which allows us to replace the expectation of $\widehat{\text{TN}}_{\widehat{Y}_{p,t,\eta}}$ with its value on an independent, identically distributed "ghost sample".

Lemma 25 (Symmetrization; Lemma 2 of Bousquet et al. [2003])**.** *Let X and X' be independent realizations of a random variable with respect to which $\mathcal{F}$ is a family of integrable functions. Then, for any $\epsilon > 0$,*

$$\mathbb{P}\left[\sup_{f \in \mathcal{F}} f(X) - \mathbb{E}\, f(X) > \epsilon\right] \leqslant 2\mathbb{P}\left[\sup_{f \in \mathcal{F}} f(X) - f(X') > \frac{\epsilon}{2}\right].$$

We now use this lemma to prove Lemma 6.

Proof. To facilitate analyzing the stochastic aspect of the classifier $\widehat{Y}_{f,\eta}$, let $Z_1, ..., Z_n \overset{IID}{\sim} \text{Uniform}([0,1])$, such that $\widehat{Y}_{f,\eta}(X_i) = 1\{Z_i < f(\eta((X_i))\}$.

Now suppose that we have a ghost sample $(X_1', Y_1', Z_1'), ..., (X_n', Y_n', Z_n')$. Let $\widehat{\text{TN}}'_{\widehat{Y}_{f,\eta}}$ denote the empirical true negative probability computed on this ghost sample, and let $\widehat{\text{TN}}^{(i)}_{\widehat{Y}_{f,\eta}}$ denote the empirical true negative probability computed on

$$(X_1, Y_1, Z_1), ..., (X_{i-1}, Y_{i-1} Z_{i-1}), (X_i', Y_i', Z_i'), (X_{i+1}, Y_{i+1}, Z_{i+1}), ...(X_n, Y_n, Z_n)$$

(i.e., replacing only the i^{th} sample with its ghost). By the Symmetrization Lemma,

$$\begin{aligned}
\mathbb{P}\left[\sup_{f \in \mathcal{F}} \widehat{\text{TN}}_{\widehat{Y}_{f,\eta}} - \mathbb{E}\, \widehat{\text{TN}}_{\widehat{Y}_{f,\eta}} > \epsilon\right] &\leqslant 2\mathbb{P}\left[\sup_{f \in \mathcal{F}} \widehat{\text{TN}}_{\widehat{Y}_{f,\eta}} - \widehat{\text{TN}}'_{\widehat{Y}_{f,\eta}} > \epsilon/2\right] \\
&\leqslant 2\Pi_{\mathcal{F}}(2n) \sup_{f \in \mathcal{F}} \mathbb{P}\left[\widehat{\text{TN}}_{\widehat{Y}_{f,\eta}} - \widehat{\text{TN}}'_{\widehat{Y}_{f,\eta}} > \epsilon/2\right] \\
&\leqslant 4\Pi_{\mathcal{F}}(2n) \sup_{f \in \mathcal{F}} \mathbb{P}\left[\widehat{\text{TN}}_{\widehat{Y}_{f,\eta}} - \mathbb{E}\, \widehat{\text{TN}}_{\widehat{Y}_{f,\eta}} > \epsilon/4\right], \qquad (13)
\end{aligned}$$

where the second inequality is a union bound over the $\Pi_{\mathcal{F}}(2n)$ distinct classifications of $2n$ points that can be assigned by $\widehat{Y}_{f,\eta}$ with $f \in \mathcal{F}$, and the last inequality is from the fact that $\widehat{\text{TN}}_{\widehat{Y}_{f,\eta}}$ and $\widehat{\text{TN}}'_{\widehat{Y}_{f,\eta}}$ are identically distributed and the algebraic fact that, if $a - b > \epsilon$, then either $a - c > \epsilon/2$ or $b - c > \epsilon/2$.

For any particular $f \in \mathcal{F}$, by McDiarmid's inequality [McDiarmid, 1998],

$$\mathbb{P}\left[\widehat{\text{TN}}_{\widehat{Y}_{f,\eta}} - \mathbb{E}\, \widehat{\text{TN}}_{\widehat{Y}_{f,\eta}} > \epsilon/4\right] \leqslant e^{-n\epsilon^2/8}, \qquad (14)$$

since, for any $i \in [n]$,

$$\left| \widehat{\mathrm{TN}}_{\widehat{Y}_{f,\eta}} - \widehat{\mathrm{TN}}^{(i)}_{\widehat{Y}_{f,\eta}} \right| = \frac{1}{n} \left| 1\left\{ Y_i = \widehat{Y}_{f,\eta}(X_i) = 0 \right\} - 1\left\{ Y_i' = \widehat{Y}_{f,\eta}(X_i') = 0 \right\} \right| \leqslant \frac{1}{n}.$$

Plugging Inequality (14) into Inequality (13) gives

$$\mathbb{P}\left[\sup_{f \in \mathcal{F}} \widehat{\mathrm{TN}}_{\widehat{Y}_{f,\eta}} - \mathbb{E}\, \widehat{\mathrm{TN}}_{\widehat{Y}_{f,\eta}} > \epsilon \right] \leqslant 4\Pi_{\mathcal{F}}(2n) e^{-n\epsilon^2/8}.$$

Repeating this argument with $-\widehat{\mathrm{TN}}$ instead of $\widehat{\mathrm{TN}}$, as well as with $\widehat{\mathrm{TP}}$, $\widehat{\mathrm{FN}}$, $\widehat{\mathrm{FP}}$ and their negatives, and taking a union bound over these 8 cases, gives the desired result. □

Finally, we will use these two lemmas, together with the margin and Lipschitz assumptions, to prove Corollary 8, which bounds the sub-optimality of the trained classifier, relative to the generalized Bayes classifier, in terms of the desired CMM.

Corollary 8. *Let $\eta : \mathcal{X} \to [0,1]$ denote the true regression function, and let $\widehat{\eta} : \mathcal{X} \to [0,1]$ denote any empirical regressor. Let*

$$(\widehat{p}, \widehat{t}) := \operatorname*{argmax}_{(p,t) \in [0,1]^2} M\left(\widehat{C}_{\widehat{Y}_{p,t,\widehat{\eta}}} \right) \quad \text{and} \quad (p^*, t^*) := \operatorname*{argmax}_{(p,t) \in [0,1]^2} M\left(C_{\widehat{Y}_{p,t,\eta}} \right)$$

denote the empirically selected and true optimal thresholds, respectively. Suppose that M is Lipschitz continuous with constant L_M with respect to the uniform ($\mathcal{L}_\infty$) metric on $\mathcal{C}$. Finally, suppose that P_X and η satisfies a (C, β)-margin condition around t^*. Then, with probability at least $1 - \delta$,

$$M\left(C_{\widehat{Y}_{p,t,\eta}}(p^*, t^*) \right) - M\left(C_{\widehat{Y}_{p,t,\widehat{\eta}}}(\widehat{p}, \widehat{t}) \right) \leqslant L_M \left(C \left\| \eta - \widehat{\eta} \right\|_\infty^\beta + 2\sqrt{\frac{8}{n} \log \frac{32(2n+1)}{\delta}} \right).$$

Proof. First, note that

$$\begin{aligned} M\left(C_{\widehat{Y}_{p^*,t^*,\eta}} \right) - M\left(C_{\widehat{Y}_{\widehat{p},\widehat{t},\widehat{\eta}}} \right) &\leqslant M\left(C_{\widehat{Y}_{p^*,t^*,\eta}} \right) - M\left(C_{\widehat{Y}_{p^*,t^*,\widehat{\eta}}} \right) \\ &\quad + M\left(C_{\widehat{Y}_{p^*,t^*,\widehat{\eta}}} \right) - M\left(\widehat{C}_{\widehat{Y}_{p^*,t^*,\widehat{\eta}}} \right) \\ &\quad + M\left(\widehat{C}_{\widehat{Y}_{\widehat{p},\widehat{t},\widehat{\eta}}} \right) - M\left(C_{\widehat{Y}_{\widehat{p},\widehat{t},\widehat{\eta}}} \right), \end{aligned}$$

since, by definition of $(\widehat{p}, \widehat{t})$,

$$M\left(\widehat{C}_{\widehat{Y}_{p^*,t^*,\widehat{\eta}}} \right) - M\left(\widehat{C}_{\widehat{Y}_{\widehat{p},\widehat{t},\widehat{\eta}}} \right) \leqslant 0;$$

this term sits between the second and third lines above. By the Lipschitz assumption,

$$\begin{aligned} &M\left(C_{\widehat{Y}_{p^*,t^*,\eta}} \right) - M\left(C_{\widehat{Y}_{\widehat{p},\widehat{t},\widehat{\eta}}} \right) \\ &\leqslant L_M \Bigg(\left\| C_{\widehat{Y}_{p^*,t^*,\eta}} - C_{\widehat{Y}_{p^*,t^*,\widehat{\eta}}} \right\|_\infty && (15) \\ &\quad + \left\| C_{\widehat{Y}_{p^*,t^*,\widehat{\eta}}} - \widehat{C}_{\widehat{Y}_{p^*,t^*,\widehat{\eta}}} \right\|_\infty && (16) \\ &\quad + \left\| \widehat{C}_{\widehat{Y}_{\widehat{p},\widehat{t},\widehat{\eta}}} - C_{\widehat{Y}_{\widehat{p},\widehat{t},\widehat{\eta}}} \right\|_\infty \Bigg). && (17) \end{aligned}$$

Corollary 8 follows by applying Lemma 5 and the (C, β)-margin condition to (15) and applying Lemma 6 to both terms (16) and (17). □

B.1 Lipschitz constants for some common CMMs

Corollary 8 assumed that the CMM M was Lipschitz continuous with respect to the $\sup$-norm on confusion matrices. In this section, we show how to compute appropriate Lipschitz constants for several simple example CMMs. We begin with a simple example:

Example 26 (Weighted Accuracy)**.** For a fixed $w \in (0,1)$, the w-weighted accuracy is given by $M(C) = (1-w)\text{TP} + w\text{TN}$. In this case, M clearly has Lipschitz constant $L_M = \max\{w, 1-w\}$.

For the remainder of this section (only), we will use $P := \mathbb{E}[Y]$ to denote the positive probability of the true labels and $\widehat{P} := \frac{1}{n}\sum_{i=1}^{n} Y_i$ to denote the empirical positive probability of the true labels. Many CMMs of interest, such as Recall and F_β scores, are not Lipschitz continuous over all of $\mathcal{C}$. Fortunately, inspecting the proof of Corollary 8, it suffices for the CMM M to be Lipschitz continuous on the line segments between three specific pairs of confusion matrices, given in Eqs. (15), (16), and (17). Deriving the appropriate Lipschitz constants is a bit more complex, and we demonstrate here how to derive them for the specific CMMs of Recall and F_β scores.

Of the six confusion matrices in Eqs. (15), (16), and (17), four are true confusion matrices, while the other two are empirical. The four true confusion matrices have the same positive probability $\text{TP} + \text{FN} = P$, which is a function of the true distribution of labels. The two empirical confusion matrices have the positive probability $\widehat{\text{TP}} + \widehat{\text{FN}} = \widehat{P}$, which is a function of the data. By a multiplicative Chernoff bound, with probability at least $1 - e^{-nP/8}$, $\widehat{P} \geqslant P/2$. Thus, with high probability, it suffices for the CMM M to be Lipschitz continuous over confusion matrices with positive probability at least $P/2$. For Recall and F_β scores, this gives the following Lipschitz constants:

Example 27 (Recall)**.** Recall is given by $M(C) = \frac{\text{TP}}{\text{TP}+\text{FN}} = \frac{\text{TP}}{P}$. Thus, M is Lipschitz continuous with constant $L_M = \frac{2}{P}$ over the confusion matrices in Eqs. (15), (16), and (17).

Example 28 (F_β Score)**.** For $\beta \in (0,\infty)$, the F_β score is given by

$$M(C) = \frac{(1+\beta^2)\text{TP}}{(1+\beta^2)\text{TP} + \text{FP} + \beta^2\text{FN}} = \frac{(1+\beta^2)\text{TP}}{\text{TP} + \text{FP} + \beta^2 P}.$$

Hence,

$$\left|\frac{\partial}{\partial \text{TP}} M(C)\right| = (1+\beta^2)\frac{\text{FP} + \beta^2 P}{(\text{TP} + \text{FP} + \beta^2 P)^2} \leqslant \frac{1+\beta^2}{\beta^2 P},$$

while, since $\text{TP} \leqslant P$,

$$\left|\frac{\partial}{\partial \text{FP}} M(C)\right| = (1+\beta^2)\frac{\text{TP}}{(\text{TP} + \text{FP} + \beta^2 P)^2} \leqslant \frac{1+\beta^2}{\beta^4 P}.$$

Hence, M is Lipschitz continuous with constant $\frac{2(1+\beta^2)}{P}\max\left\{\beta^{-2}, \beta^{-4}\right\}$ over the confusion matrices in Eqs. (15), (16), and (17).

As Examples 27 and 28 demonstrate, the Lipschitz constants of some CMMs can become large when the proportion P is positive samples is small. In particular, when $P \in O\left(\sqrt{\frac{\log n}{n}}\right)$, the $\asymp L_M\sqrt{\frac{\log(n/\delta)}{n}}$ term of Corollary 8 fails to vanish as $n \to \infty$. We believe that some loss of convergence rate is inevitable if $P \to 0$ as $n \to \infty$, due to the inherent instability of such metrics, but further work is needed to understand if the rates given by Corollary 8 are optimal under these metrics. See also Dembczyński et al. [2017] for detailed discussion of Lipschitz constants of many common CMMs.

C Bounds on Uniform Error of the Nearest Neighbor Regressor

In this appendix, we prove our upper bound on the uniform risk of the kNN regressor (Theorem 15), as well as the corresponding minimax lower bound (Theorem 17).

C.1 Upper Bounds

Here, we prove Theorem 15, our upper bound on the uniform error of the k-NN regressor, restated below:

Theorem 15. *Under Assumptions 13 and 14, whenever $k/n \leqslant p_*(\epsilon^*)^d/2$, for any $\delta > 0$, with probability at least $1 - N\left((2k/(p_*n))^{1/d}\right)e^{-k/4} - \delta$, we have the uniform error bound*

$$\|\eta - \widehat{\eta}\|_\infty \leqslant 2^\alpha Lr\left(\frac{2k}{p_*n}\right)^{\alpha/d} + \frac{2}{3k}\log\frac{2S(n)}{\delta} + \sqrt{\frac{2r}{k}\log\frac{2S(n)}{\delta}}. \tag{18}$$

Proof. For any $x \in \mathcal{X}$, let

$$\widetilde{\eta}_k(x) := \frac{1}{k}\sum_{j=1}^{k}\eta(X_{\sigma_j(x)})$$

denote the mean of the true regression function over the k nearest neighbors of x. By the triangle inequality,

$$\|\eta - \widehat{\eta}\|_\infty \leqslant \|\eta - \widetilde{\eta}_k\|_\infty + \|\widetilde{\eta}_k - \widehat{\eta}\|_\infty,$$

wherein $\|\eta - \widetilde{\eta}_k\|_\infty$ captures bias due to smoothing and $\|\widetilde{\eta}_k - \widehat{\eta}\|_\infty$ captures variance due to label noise. We separately show that, with probability at least $1 - N\left(\left(\frac{2k}{p_*n}\right)^{1/d}\right)e^{-k/4}$,

$$\|\eta - \widetilde{\eta}_k\|_\infty \leqslant 2^\alpha Lr\left(\frac{2k}{p_*n}\right)^{\alpha/d},$$

and that, with probability at least $1 - \delta$,

$$\|\widetilde{\eta}_k - \widehat{\eta}\|_\infty \leqslant \frac{2}{3k}\log\frac{2S(n)}{\delta} + \sqrt{\frac{2r}{k}\log\frac{2S(n)}{\delta}}.$$

Bounding the smoothing bias Fix some $r > 0$ to be determined, and let $\{B_r(z_1), ..., B_r(z_{N(r)})\}$ be a covering of $(\mathcal{X}, \rho)$ by $N(r)$ balls of radius r, with centers $z_1, ..., z_{N(r)} \in \mathcal{X}$.

By the lower bound assumption on P_X, each $P_X(B_r(z_j)) \geqslant p_* r^d$. Therefore, by a multiplicative Chernoff bound, with probability at least $1 - N(r)e^{-p_* n r^d/8}$, each $B_r(z_j)$ contains at least $p_* n r^d/2$ samples. In particular, if $r \geqslant \left(\frac{2k}{p_*n}\right)^{1/d}$, then each B_k contains at least k samples, and it follows that, for every $x \in \mathcal{X}$, $\rho(x, X_{\sigma_k(x)}) \leqslant 2r$. Thus, by Hölder continuity of η,

$$|\eta(x) - \widetilde{\eta}_k(x)| = \left|\eta(x) - \frac{1}{k}\sum_{j=1}^{k}\eta(X_{\sigma_j(x)})\right| \leqslant \frac{1}{k}\sum_{j=1}^{k}\left|\eta(x) - \eta(X_{\sigma_j(x)})\right| \leqslant L(2r)^\alpha.$$

Finally, if $\frac{k}{n} \leqslant \frac{p_*}{2}(r^*)^d$, then we can let $r = \left(\frac{2k}{p_*n}\right)^{1/d}$.

Bounding variance due to label noise Let $\Sigma := \{\sigma(x) \in [n]^k : x \in \mathcal{X}\}$ denote the set of possible k-nearest neighbor index sets. One can check from the definition of the shattering coefficient that $|\Sigma| \leqslant S(n)$.

For any $\sigma \in [n]^k$, let $Z_\sigma := \sum_{j=1}^{k} Y_{\sigma_j}$ and let $\mu_\sigma := \mathbb{E}[Z_\sigma]$. Note that the conditional random variables $Y_{\sigma_j}|X_1, ..., X_n$ have conditionally independent Bernoulli distributions with means $\mathbb{E}[Y_{\sigma_j}|X_1, ..., X_n] = \eta(X_{\sigma_j})$ and variances $\mathbb{E}\left[\left(Y_{\sigma_j} - \eta(X_{\sigma_j})\right)^2|X_1, ..., X_n\right] = \eta(X_{\sigma_j})(1 - \eta(X_{\sigma_j})) \leqslant r$. Therefore, by Bernstein's inequality (Eq. (2.10) of Boucheron et al. [2013]), for any $\epsilon > 0$,

$$\mathbb{P}\left[|Z_\sigma/k - \mu_\sigma| \geqslant \epsilon\right] \leqslant 2\exp\left(-\frac{k\epsilon^2}{2(r + \epsilon/3)}\right). \tag{19}$$

Moreover, for any $x \in \mathcal{X}$, $\mu_{\sigma(x)} = \widetilde{\eta}_k(x)$ and $Z_{\sigma(x)}/k = \widehat{\eta}(x)$. Hence, by a union bound over σ in Σ,

$$\begin{aligned}
\mathbb{P}\left(\sup_{x\in\mathcal{X}} |\widetilde{\eta}_k(x) - \widehat{\eta}(x)| > \epsilon \middle| X_1, ..., X_n\right) &= \mathbb{P}\left(\sup_{x\in\mathcal{X}} \left|\mu_{\sigma(x)} - Z_{\sigma(x)}/k\right| > \epsilon \middle| X_1, ..., X_n\right) \\
&\leqslant \mathbb{P}\left(\sup_{\sigma\in\Sigma} |\mu_\sigma - Z_\sigma/k| > \epsilon \middle| X_1, ..., X_n\right) \\
&\leqslant |\Sigma| \sup_{\sigma\in\Sigma} \mathbb{P}\left(|\mu_\sigma - Z_\sigma/k| > \epsilon | X_1, ..., X_n\right) \\
&\leqslant 2S(n) \exp\left(-\frac{k\epsilon^2}{2(r+\epsilon/3)}\right).
\end{aligned}$$

Since the right-hand side is independent of $X_1, ..., X_n$, the unconditional bound

$$\mathbb{P}\left(\sup_{x\in\mathcal{X}} \|\widetilde{\eta}_k(x) - \widehat{\eta}(x)\|_\infty > \epsilon\right) \leqslant 2S(n) \exp\left(-\frac{k\epsilon^2}{2(r+\epsilon/3)}\right)$$

follows. Plugging in

$$\epsilon = \frac{1}{3k}\log\frac{2S(n)}{\delta} + \sqrt{\left(\frac{1}{3k}\log\frac{2S(n)}{\delta}\right)^2 + \frac{2r}{k}\log\frac{2S(n)}{\delta}} \leqslant \frac{2}{3k}\log\frac{2S(n)}{\delta} + \sqrt{\frac{2r}{k}\log\frac{2S(n)}{\delta}}$$

and simplifying gives the final result. □

Recall that there is a small (polylogarithmic in r) gap between our upper and lower bounds. We believe that the upper bound may be slightly loose, and that this might be tightened by using a stronger concentration inequality, such as Bennett's inequality [Bennett, 1962], instead of Bernstein's inequality in Inequality (19).

Naively applying Theorem 15 results in very slow convergence rates in high dimensions. For this reason, we close this section with a corollary of Theorem 15, illustrating that the convergence rates provided by Theorem 15 improve if the covariates are assumed to lie on an (unknown) lower dimensional manifold:

Corollary 29 (Implicit Manifold Case). *Suppose Z is a $[0,1]^d$-valued random variable with a density lower bounded away from 0, and suppose that, for some Lipschitz map $T : [0,1]^d \to \mathbb{R}^D$, $X = T(Z)$. Then, $N(\epsilon) \leqslant (2/\epsilon)^d$, and $S(n) \leqslant 2n^{D+1} + 2$, and so, by Theorem 15, $k \asymp n^{\frac{2\alpha}{2\alpha+d}} (\log n)^{\frac{d}{2\alpha+d}} r^{-\frac{d}{2\alpha+d}}$,*

$$\|\eta - \widehat{\eta}\|_\infty \in O_P\left(\left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+d}} r^{\frac{\alpha+d}{2\alpha+d}}\right).$$

This shows that, if the D covariates lie implicitly on a d-dimensional manifold, convergence rates depend on d, which may be much smaller than D.

C.2 Lower Bounds

In this section, we prove Theorem 17, our lower bound on the minimax uniform error of estimating a Hölder continuous regression function. We use a standard approach based on the following version of Fano's lemma:

Lemma 30. *(Fano's Lemma; Simplified Form of Theorem 2.5 of Tsybakov 2009) Fix a family $\mathcal{P}$ of distributions over a sample space $\mathcal{X}$ and fix a pseudo-metric $\rho : \mathcal{P} \times \mathcal{P} \to [0, \infty]$ over $\mathcal{P}$. Suppose there exist $P_0 \in \mathcal{P}$ and a set $T \subseteq \mathcal{P}$ such that*

$$\sup_{P\in T} D_{KL}(P, P_0) \leqslant \frac{\log|T|}{16},$$

where $D_{KL} : \mathcal{P} \times \mathcal{P} \to [0, \infty]$ denotes Kullback-Leibler divergence. Then,

$$\inf_{\widehat{P}} \sup_{P\in\mathcal{P}} \mathbb{P}\left(\rho(P, \widehat{P}) \geqslant \frac{1}{2} \inf_{P\in T} \rho(P, P_0)\right) \geqslant 1/8,$$

where the first inf *is taken over all estimators $\widehat{P}$.*

Proof. We now proceed to construct an appropriate $P_0 \in \mathcal{P}$ and $T \subseteq \mathcal{P}$. Let $g : [-1,1]^d \to [0,1]$ defined by

$$g(x) = \begin{cases} \exp\left(1 - \frac{1}{1-\|x\|_2^2}\right) & \text{if } \|x\|_2 < 1 \\ 0 & \text{else} \end{cases}$$

denote the standard bump function supported on $[-1,1]^d$, scaled to have $\|g\|_{\mathcal{X},\infty} = 1$. Since g is infinitely differentiable and compactly supported, it has a finite α-Hölder semi-norm:

$$\|g\|_{\Sigma^\alpha} := \sup_{\ell \in \mathbb{N}^d : \|\ell\|_1 \leqslant \alpha} \sup_{x \neq y \in \mathcal{X}} \frac{|g^\ell(x) - g^\ell(y)|}{\|x-y\|^{\alpha - \|\ell\|_1}} < \infty, \tag{20}$$

where ℓ is any $\lfloor\beta\rfloor$-order multi-index and g^ℓ is the corresponding mixed derivative of g. Define $M := \left(\frac{64(2\alpha+d)nr}{d\log(nr)}\right)^{\frac{1}{2\alpha+d}} \geqslant 1$, since $r \geqslant 1/n$. For each $m \in [M]^d$, define $g_m : \mathcal{X} \to [0,1]$ by

$$g_m(x) := g\left(Mx - \frac{2m - 1_d}{2}\right),$$

so that $\{g_m : m \in [M]^d\}$ is a grid of M^d bump functions with disjoint supports. Let $\zeta_0 \equiv \frac{1}{4}$ denote the constant-$\frac{1}{4}$ function on $\mathcal{X}$. Finally, for each $m \in [M]^d$, define $\zeta_m : \mathcal{X} \to [0,1]$ by

$$\zeta_m := \zeta_0 + \min\left\{\frac{1}{2}, \frac{L}{\|g\|_{\Sigma^\alpha}}\right\} M^{-\alpha} g_m. \tag{21}$$

Note that, for any $m \in [M]^d$,

$$\|\zeta_m\|_{\Sigma^\alpha} \leqslant L M^{-\alpha} \frac{\|g_m\|_{\Sigma^\alpha}}{\|g\|_{\Sigma^\alpha}} = L,$$

so that ζ_m satisfies the Hölder smoothness condition. For any particular η, let P_η denote the joint distribution of (X,Y). Note that $P_\zeta(x,1) = \zeta(x) \geqslant 1/4$. Moreover, one can check that, for all $x \geqslant -2/3$, $-\log(1+x) \leqslant x^2 - x$. Hence, for any $x \in \mathcal{X}$,

$$\begin{aligned}
P_{\eta_m}(x,1) \log \frac{P_{\eta_m}(x,1)}{P_\eta(x,1)} &= r P_{\zeta_m}(x,1) \log \frac{P_{\zeta_m}(x,1)}{P_\zeta(x,1)} \\
&= r\zeta_m(x) \log \frac{\zeta_m(x)}{\zeta(x)} \\
&= -r\zeta_m(x) \log\left(1 + \frac{\zeta(x) - \zeta_m(x)}{\zeta_m(x)}\right) \\
&\leqslant r\zeta_m(x)\left(\left(\frac{\zeta(x) - \zeta_m(x)}{\zeta_m(x)}\right)^2 - \frac{\zeta(x) - \zeta_m(x)}{\zeta_m(x)}\right) \\
&= r\left(\frac{(\zeta(x) - \zeta_m(x))^2}{\zeta_m(x)} - \zeta(x) + \zeta_m(x)\right) \\
&\leqslant r\left(4(\zeta(x) - \zeta_m(x))^2 - \zeta(x) + \zeta_m(x)\right),
\end{aligned}$$

and, similarly, since $P_\zeta(x,0) = 1 - \zeta(x) \geqslant 1/4$,

$$P_{r\eta_m}(x,0) \log \frac{P_{r\eta_m}(x,0)}{P_{r\eta}(x,0)} \leqslant r\left(4(\zeta(x) - \zeta_m(x))^2 + \zeta(x) - \zeta_m(x)\right).$$

Adding these two terms gives

$$\begin{aligned}
D_{\mathrm{KL}}\left(P_{r\eta}^n, P_{r\eta_m}^n\right) &= n\left(\int_{\mathcal{X}} P_{r\eta_m}(x,0)\log\frac{P_{r\eta}(x,0)}{P_{r\eta_m}(x,0)}\,dx + \int_{\mathcal{X}} P_{r\eta_m}(x,1)\log\frac{P_{r\eta}(x,1)}{P_{r\eta_m}(x,1)}\,dx\right)\\
&\leqslant 8nr\int_{\mathcal{X}}\left(\zeta(x)-\zeta_m(x)\right)^2\\
&= 8nr\|\zeta-\zeta_m\|_2^2\\
&\leqslant 2nrM^{-2\alpha}\|g_m\|_2^2\\
&= 2nrM^{-(2\alpha+d)}\|g\|_2^2\\
&= 2nr\left(\left(\frac{64(2\alpha+d)nr}{d\log(nr)}\right)^{\frac{1}{2\alpha+d}}\right)^{-(2\alpha+d)}\|g\|_2^2\\
&= \frac{1}{32}\frac{d}{2\alpha+d}\|g\|_2^2\log(nr)\\
&\leqslant \frac{1}{16}\frac{d}{2\alpha+d}\left(\log(nr)-\log\log(nr)+\log\frac{64(2\alpha+d)}{d}\right) = \frac{\log|[M]^d|}{16},
\end{aligned}$$

where the second inequality comes from the definition of ζ_m (Eq. 21) and the third inequality comes from the facts that $\|g\|_2^2 \leqslant 1$ and $\log\log x \leqslant \frac{1}{2}\log x$ for all $x > 1$. Fano's lemma therefore implies the lower bound

$$\inf_{\widehat{\eta}}\sup_{r\in(0,1],\zeta\in\Sigma^\alpha(L)}\mathbb{P}_{\{(X_i,Y_i)\}_{i=1}^n\sim P_\eta^n}\left(\left\|r\zeta - r\widehat{\zeta}\right\|_\infty \geqslant C\left(\frac{\log(nr)}{n}\right)^{\frac{\alpha}{2\alpha+d}} r^{\frac{\alpha+d}{2\alpha+d}}\right) \geqslant \frac{1}{8},$$

where

$$C = \frac{1}{2}\min\left\{\frac{1}{2}, \frac{L}{\|g\|_{\Sigma^\alpha}}\right\}\left(\frac{d}{64(2\alpha+d)}\right)^{\frac{\alpha}{2\alpha+d}}.$$

□

D Efficient Computation of the Optimal Stochastic Threshold

Although the focus of this paper is on *statistical* properties of regression-thresholding classifiers, we note that, given an estimate $\widehat{\eta}$ of the regression function, the empirically optimal stochastic threshold $(\widehat{p}, \widehat{t})$, i.e., that which maximizes $M(\widehat{C}_{\widehat{Y}_{\widehat{\eta},\widehat{p},\widehat{t}}})$, can be efficiently computed. In this appendix, we describe a simple algorithm for doing so. The key insight is that, because $(\widehat{p}, \widehat{t})$ is used to threshold the observed empirical class probabilities $\widehat{\eta}(X_1), ..., \widehat{\eta}(X_n)$ before computing M, $M(\widehat{C}_{\widehat{Y}_{\widehat{\eta},p,t}})$ only needs to be computed at the n values of $\widehat{\eta}$ actually observed in the data.

We also note that, while, by Corollary 8, one can safely use the original training dataset to compute $(\widehat{p}, \widehat{t})$, one can also safely use a much smaller subset of the data, since the rate of convergence in Lemma 6 is quite fast in n.

For large n, the runtime of Algorithm 1 is dominated by Line 3, which involves lexicographically sorting n pairs. This can be done in $O(n\log n)$ time using standard comparison-based sorting algorithms. Hence, the overall runtime of Algorithm 1 is $O(n\log n)$.

E Further Experimental Details

Experiments were run using the `numpy` and `scikit-learn` packages in Python 3.9, on a machine running Ubuntu 20.04 with an Intel Core i5-9600 CPU and 64 gigabytes of memory. Each experiment took about 10 minutes to run. Python code and instructions for reproducing Figures 2b and 2a are available at `https://gitlab.tuebingen.mpg.de/shashank/imbalanced-binary-classification-experiments`.

Algorithm 1: Efficient threshold-optimization algorithm.

Input: Estimated regression function $\widehat{\eta}$, training covariate samples $X_1, ..., X_n$, CMM M.

Output: Estimated optimal stochastic threshold $(\widehat{p}, \widehat{t})$

Sample $Z_1, ..., Z_n \overset{IID}{\sim} \text{Uniform}([0,1])$
$e_1, ..., e_n \leftarrow \widehat{\eta}(X_1), ..., \widehat{\eta}(X_n)$
$(e_1, Z_1), ..., (e_n, Z_n) \leftarrow \mathsf{LexicographicSort}((e_1, Z_n), ..., (e_n, Z_n))$
$\text{TP} \leftarrow \frac{1}{n}\sum_{i=1}^{n} Y_i$
$\text{FP} \leftarrow 1 - \text{TP}$
$\text{TN}, \text{FN} \leftarrow 0$
$(\widehat{p}, \widehat{t}) \leftarrow (0, 0)$
$M_{\text{max}} \leftarrow M\left(\begin{bmatrix} \text{TN} & \text{FP} \\ \text{FN} & \text{TP} \end{bmatrix}\right)$
for $i = 1; i <= n; i++$ **do**
    $\text{TP} \leftarrow \text{TP} - Y_i/n$
    $\text{FP} \leftarrow \text{TP} - (1 - Y_i)/n$
    $\text{TN} \leftarrow \text{TN} + (1 - Y_i)/n$
    $\text{FN} \leftarrow \text{FN} + Y_i/n$
    $M_{\text{new}} \leftarrow M\left(\begin{bmatrix} \text{TN} & \text{FP} \\ \text{FN} & \text{TP} \end{bmatrix}\right)$
    if $M_{new} > M_{max}$ **then**
        $(\widehat{p}, \widehat{t}) \leftarrow (e_i, Z_i)$
        $M_{\text{max}} \leftarrow M_{\text{new}}$
end
return $(\widehat{p}, \widehat{t})$

F Experiments with Real Data: Case Study in Credit Card Fraud Detection

In this section, we explore theoretical predictions from the main paper in a real dataset, the Kaggle Credit Card Fraud Detection dataset (available at `https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud` under an Open Database License (ODbL)), a widely used benchmark dataset for imbalanced classification. This dataset contains 29 continuous features (computed via PCA from an underlying set of features) for each of 284,807 credit card transactions, of which 492 (0.172%) are labeled as fraudulent, and the remaining are assumed to be non-fraudulent. The supervised learning task is to predict whether a credit card transaction is fraudulent, given its 29 PCA features. Due to computational limitations, we down-sampled the negative set (non-fraudulent transactions) by a factor of 0 before conducting our experiments; however, we expect our main observations to hold on the full dataset as well. We also Z-scored each feature (to have mean 0 and variance 1).

The main question we sought to investigate here was whether the theoretical finding, in Theorem 3, that stochastic classification is sometimes necessary in order to obtain optimal prediction performance under general performance metrics, would be visible in real data. To investigate this, we partitioned the dataset randomly into a training subset (60% of samples), a validation subset (20% of samples), and a test subset (20% of samples). We fit a k-nearest neighbor regressor (with Euclidean distance as the underlying metric) to the training subset, used the validation subset to select optimal deterministic and generalization thresholds, and then used the test subset to evaluate performance. We evaluated performance in terms of F_1 score, since it is perhaps the CMM most widely used with imbalanced datasets. We then repeated this experiment with 100 random train/validation/test splits and report aggregate results over these independent trials.

We generally found that, as predicted by our theoretical results, stochastic thresholding generally outperforms deterministic thresholding by a small but consistent margin. Figure 3 shows that, for fixed nearest neighbor hyperparameter $k = 4$, this effect is robust across differing degrees of class imbalance, for imbalance ratios ranging from $1:1$ (perfect balance) to $57:1$ (the full dataset), where class imbalance here was manipulated by down-sampling the negative class. Similarly,

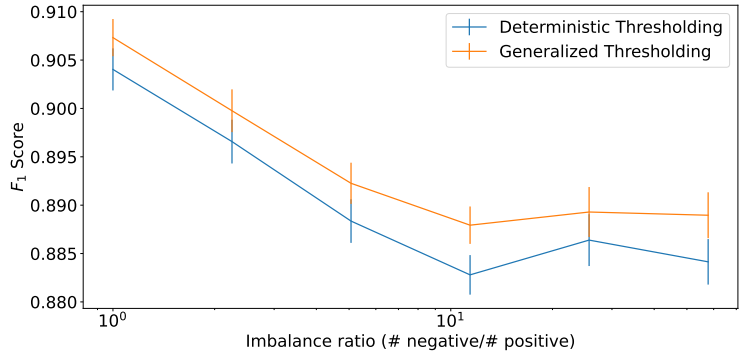

Figure 3: Mean F_1 scores (over 100 random training/validation/test splits) of optimal deterministic and stochastic thresholding nearest neighbor classifiers, on the credit card fraud dataset, at various degrees of class imbalance. Error bars denote standard errors, computed over the 100 random training/validation/test splits.

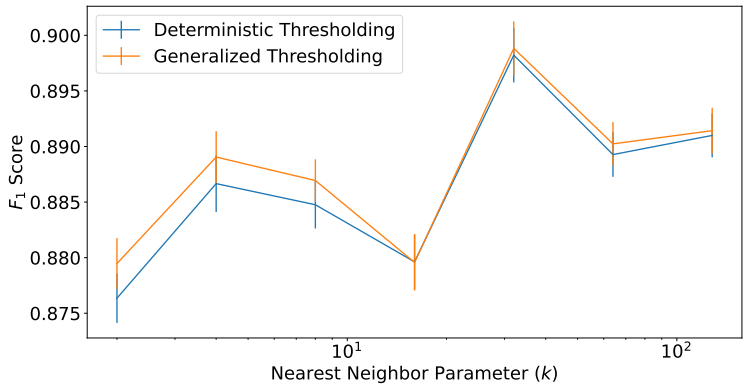

Figure 4: Mean F_1 scores (over 100 random training/validation/test splits) of optimal deterministic and stochastic thresholding nearest neighbor classifiers, on the credit card fraud dataset, for various values of the nearest neighbor hyperparameter k. Error bars denote standard errors, computed over the 100 random training/validation/test splits.

Figure 4 shows that this effect is robust over different values of the nearest neighbor hyperparameter $k \in \{2, 4, 8, 16, 32, 64, 128\}$.