Near-Exponential Convergence Rates for kNN Classification based on Boltzmann Margin

Luyuan Yang[1] **Shayan Shafaei**[1] **Chao Lan**[1]

[1]School of Computer Science, University of Oklahoma, Norman, Oklahoma, USA

Abstract

Convergence-rate analysis for classifiers is often conducted under either Tsybakov margin or Massart margin. The former is a relatively weak condition that typically yields polynomial rates, while the latter is substantially stronger but can guarantee exponential rates. In this paper, we introduce a new condition, called *Boltzmann margin*, that bridges the gap between these two regimes. It is weaker than Massart margin, generally stronger than Tsybakov margin, and can imply many of their properties under suitable conditions. We apply Boltzmann margin to the analysis of kNN classifiers and establish the first *near-exponential* convergence rates for kNN classification. We also present extensions of the main results and provide numerical evidence supporting the main theoretical implications.

1 INTRODUCTION

In machine learning, a fundamental question is how fast the classification error converges to the Bayes error as training data size increases. The convergence rate has been widely studied under two margin conditions: *Tsybakov margin* [Tsybakov, 2004] and *Massart margin* [Massart and Nédélec, 2006]. Tsybakov margin is a relatively weak condition that assumes data decay polynomially fast towards the Bayes decision boundary and typically yields polynomial rates [Chaudhuri and Dasgupta, 2014, Döring et al., 2018]. Massart margin is substantially stronger, as it assumes no data exist near the boundary, but can guarantee exponential rates [Cabannnes and Vigogna, 2023, Vigogna et al., 2022].

There remains a substantial gap between the two margins. To establish any super-polynomial convergence rate, existing analyses typically require Massart margin, which may be unnecessarily strong for many problems. *Can there be a more balanced margin condition that yields super-polynomial convergence rates without assuming an absence of data near the boundary?* This question motivates our study.

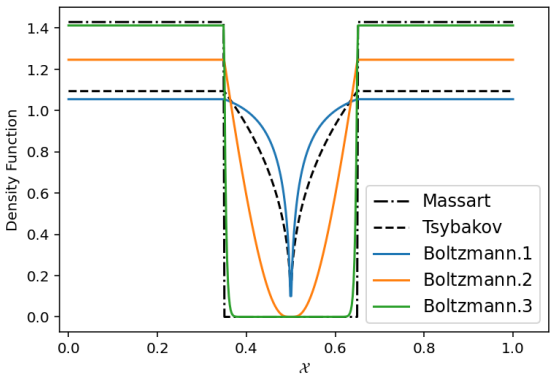

Figure 1: Example data densities on $[0, 1]$ that satisfy different margins respectively. Bayes decision boundary is 0.5.

In this paper, we introduce *Boltzmann margin*, a more balanced condition that assumes data decay exponentially fast towards the Bayes decision boundary. It is weaker than Massart margin, generally stronger than Tsybakov margin, and can imply many of their properties under suitable conditions. Figure 1 illustrates the relations among the three margins. In general, the decay behavior under Boltzmann margin lies between those under Massart and Tsybakov margins. Moreover, through suitable choices of its parameters, Boltzmann margin can resemble the decay behavior of either margin in certain regions, making it a flexible framework for convergence rate analysis.

We apply Boltzmann margin to analyze k-nearest-neighbor (kNN) classifiers. kNN is popular due to its flexibility and interpretability [Chen et al., 2018], and is increasingly used to enhance the explainability of complex models [Khandelwal et al., 2020, Rajani et al., 2020, Dai and Wang, 2021, Li et al., 2024]. These developments motivate a deeper understanding of the fundamental properties of kNN.

Convergence rates for kNN and ensemble kNN (ekNN)

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:7818–7837.

Model	Work	Posterior	Margin	Density	Error Order
kNN	[Kulkarni and Posner, 2002]	Hölder	x	x	$n^{-\frac{2\upsilon}{D}}$
	[Gyorfi, 2006]	Hölder*	x	x	$n^{-\frac{\upsilon}{2\upsilon+D}}$
	[Kohler and Krzyzak, 2007]	Hölder	Tsybakov*	strong	$n^{-\frac{\upsilon(1+\beta)}{2\upsilon+D}}$
	[Samworth, 2012b]	Hölder	Tsybakov	min.mass	$n^{-\frac{2(1+\beta)}{4+D}}$
	[Chaudhuri and Dasgupta, 2014]	Hölder*	Tsybakov	x	$n^{-\frac{\upsilon(1+\beta)}{2\upsilon+D}}$
	[Chaudhuri and Dasgupta, 2014]	Hölder*	Massart	x	$e^{-c'n}$
	[Gadat et al., 2016]	Lipschitz	Tsybakov	min.mass	$n^{-\frac{1+\beta}{2+D}}$
	[Döring et al., 2018]	Lipschitz*	Tsybakov	x	$n^{-\frac{1+\beta}{2+D}}$
	[Györfi and Weiss, 2021]	Hölder	Tsybakov	x	$n^{-\frac{\upsilon(1+\beta)}{2\upsilon+D}}$
	[Yang and Zhang, 2025]	Hölder	Tsybakov	x	$n^{-\frac{3}{4}}$
	Ours	Hölder	Boltzmann	x	$e^{-cn^{\frac{\beta}{\beta+C_\upsilon}}}$
	Ours	Hölder	Massart	x	e^{-cn}
ekNN	[Biau et al., 2010]	Lipschitz	x	x	$n^{-\frac{2}{2+D}}$
	[Samworth, 2012a]	continuous	Tsybakov	min.mass	$n^{-\frac{4}{4+D}}$
	[Xue and Kpotufe, 2018]	Hölder	Tsybakov	min.mass	$n^{-\frac{\upsilon(1+\beta)}{2\upsilon+d}}$
	[Qiao et al., 2019]	Hölder*	Tsybakov	x	$n^{-\frac{\upsilon(1+\beta)}{2\upsilon+D}}$
	Ours	Hölder	Boltzmann	x	$e^{-cn^{\frac{g\beta}{\beta+C_\upsilon}}}$

Table 1: Convergence Rates for kNN Classifiers. Notations are loosely defined to cover both basic and generalized (marked by '*') definitions. n is training data size; β is the main margin parameter; υ is the main Hölder parameter (for Hölder smooth posterior function η); D/d is input/intrinsic data dimension; $C_\upsilon = 2 + d/\upsilon$ and c', c, g are positive constants.

classifiers have long been studied. Table 1 summarizes the known results. As shown in the table, almost all studies assume a Tsybakov margin condition and establish polynomial rates. It is also shown in [Chaudhuri and Dasgupta, 2014, Appendix C.4] that Massart margin can imply an exponential convergence rate for kNN. In this paper, we apply Boltzmann margin to the analysis of kNN classifiers and establish their first *near-exponential* convergence rates.

In the remainder of the paper, we review related work in Section 2, introduce and analyze Boltzmann margin in Section 3, present new results for kNN and ekNN in Sections 4 and 5, and provide numerical results in Section 6.

2 RELATED WORK

To facilitate comparison of existing results, we adopt the notation used in Table 1 in this section. Formal notation for the present study is introduced in Section 3.

2.1 KNN CLASSIFICATION

Early studies on kNN convergence rates are based on relatively mild assumptions, notably Hölder smoothness of the posterior function η. [Kulkarni and Posner, 2002, Corollary 1] presents a $O(n^{-\frac{2\upsilon}{D}})$ convergence rate of 1NN error (to twice the Bayes error) when both η and the conditional label variance are Hölder. [Gyorfi, 2006] presents an asymptotic rate of $O(n^{-\frac{\upsilon}{2\upsilon+D}})$ for kNN when η is generalized Hölder (which, informally, requires the average posterior within any open ball to approach the posterior at the center of the ball as the ball radius shrinks, with the convergence rate controlled by a Hölder-type parameter υ) and $k = n^{\frac{2\upsilon}{2\upsilon+D}}$.

Later studies establish faster rates for kNN under Tsybakov margin and additional density assumptions. [Kohler and Krzyzak, 2007, Theorem 6] presents a $\tilde{O}(n^{-\frac{\upsilon(1+\beta)}{2\upsilon+D}})$ rate when η is Hölder, the data domain is bounded, the data density is bounded away from zero (i.e., strong density), and $k = \tilde{O}(n^{\frac{2\upsilon}{2\upsilon+D}})$. [Samworth, 2012b, Theorem 1] presents a $O(n^{-\frac{2(1+\beta)}{4+D}})$ rate when η is Hölder, the data density has minimal mass (which, informally, requires any open ball centered at any point to have a mass that scales with its radius), and $k = O(n^{\frac{4}{4+D}})$. [Gadat et al., 2016, Theorem 3.1] presents a $O(n^{-\frac{1+\beta}{2+D}})$ rate when η is Lipschitz continuous, the data density has minimal mass, and $k = O(n^{\frac{2}{2+D}})$.

Recent studies remain rooted in the Tsybakov margin and focus on relaxing conditions. [Chaudhuri and Dasgupta, 2014, Theorem 4] presents a $O(n^{-\frac{\upsilon(1+\beta)}{2\upsilon+D}})$ rate when η is generalized Hölder and $k = O(n^{\frac{2\upsilon}{2\upsilon+D}})$. [Döring et al., 2018, Theorem 1] presents a $O(n^{-\frac{1+\beta}{2+D}})$ rate when η is modified Lips-

chitz (which, informally, requires the posterior gap between two points to decrease as the radius of an open ball that contains them decreases) and $k = O(n^{\frac{2}{2+D}})$. [Györfi and Weiss, 2021, Theorem 6] presents a $O(n^{-\frac{\upsilon(1+\beta)}{2\upsilon+D}})$ rate for multi-class kNN in a metric space when η is (weighted) Hölder, the distribution function is continuous, and $k = O(n^{\frac{2\upsilon}{2\upsilon+D}})$. Recently, [Yang and Zhang, 2025, Theorem 4] shows that the convergence rate can vary across different phases and reach $O(Dn^{-3/4})$ when $n \leq D^2$, provided that η is Hölder, $D > 10$, and $k = O(n^{3/4})$. An exception appears in the appendix of [Chaudhuri and Dasgupta, 2014], where the established polynomial rate is extended to an asymptotic exponential rate under Massart margin.

Convergence rates have also been studied for several kNN variants. Under Tsybakov margin, [Reeve and Kabán, 2019, Theorem 4.3] presents an asymptotic rate of $\tilde{O}(n^{-\frac{\upsilon(\beta+1)}{2\upsilon+1}})$ for robustified kNN when η is modified Lipschitz and $k = \tilde{O}(n^{\frac{2\upsilon}{2\upsilon+1}})$; [Zhang and Li, 2023, Theorem 5] presents a $O(n^{-\frac{1+\beta}{1+d}})$ rate for compressed kNN when η is modified Lipschitz, the compressed data dimension is sufficiently large, and $k = O(n^{\frac{2}{2+d}})$. Without a margin assumption, [Rimanic et al., 2020, Theorem 4.8] presents an asymptotic rate of $O(n^{-\frac{1}{2+d}})$ for kNN trained in a bounded transformed domain when η is probabilistic Lipschitz and $k = O(n^{\frac{2}{2+D}})$.

2.2 ENSEMBLE KNN CLASSIFICATION

Ensemble learning is a popular strategy for improving classification performance by combining weak classifiers, typically learned from bootstrap samples, into a strong classifier [Dietterich, 2000, Zhou, 2025]. Let $\tilde{n}$ be the bootstrap sample size and m be the number of weak classifiers.

Early studies on ensemble kNN assume $m = \infty$. [Biau et al., 2010, Corollary 10] shows a bagged 1NN regression estimator has a $O(n^{-\frac{2}{2+D}})$ error rate when η is Lipschitz, the label variance is bounded, $D \geq 3$, $\tilde{n} \propto n^{\frac{D}{2+D}}$, and $k = O(n^{\frac{2}{2+D}})$. (The same rate applies to the induced bagged 1NN classifier.) [Samworth, 2012a, Corollary 4] shows that under an implied Tsybakov margin, a bagged 1NN classifier has a $O(n^{-\frac{4}{4+D}})$ rate when η is continuous, the data density has minimal mass, and $n^r \leq \tilde{n} \leq n^{1-r}$ for any $r \in (0, 1/2)$.

Recent studies remain rooted in Tsybakov margin but do not require $m = \infty$. [Xue and Kpotufe, 2018] considers a special bagged 1NN classifier, where each bootstrap sample is relabeled using the training set and the kNN classification rule with $k = \tilde{O}(n^{\frac{2\upsilon}{2\upsilon+d}})$. It shows that the classifier has an asymptotic rate of $\tilde{O}(n^{-\frac{\upsilon}{2\upsilon+d}}) + \tilde{O}(\tilde{n}^{-\frac{\upsilon}{d}})$ when η is Hölder and the data density has minimal mass, and that the first term dominates when $\tilde{n}/n = \Omega(n^{-\frac{\upsilon}{2\upsilon+d}})$. [Qiao et al., 2019] presents a $O(n^{-\frac{\upsilon(1+\beta)}{2\upsilon+D}})$ rate for bagged kNN when η is modified Hölder and $k = O(\tilde{n}^{\frac{2\upsilon}{2\upsilon+D}} m^{-\frac{D}{2\upsilon+D}}) \to \infty$ as $n \to \infty$. In addition to standard ensemble kNN, [Hang et al., 2022, Theorem 3] designs an under-bagging kNN classifier for imbalanced classes; it shows that the classifier has an asymptotic rate of $\tilde{O}(n^{-\frac{\upsilon(\beta+1)}{2\upsilon+D}})$ when η is Hölder, $k = \tilde{O}(m\tilde{n}^{-\frac{D}{2\upsilon+D}})$, and m scales at a rate of $\tilde{O}(n^{\frac{2\upsilon}{2\upsilon+D}})$.

2.3 CONSISTENCY FOR EKNN CLASSIFIER

Bayes consistency is a desirable property of classifiers. Informally, a classifier is *weakly consistent* if its expected error converges to the Bayes error, and *strongly consistent* if its error converges to the Bayes error with probability one (over the random choice of training data). Strong consistency implies weak consistency [Devroye et al., 2013].

Weak consistency is often easier to establish. In fact, most studies reviewed in Section 2.2 already imply weak consistency, since their convergence rates imply that the expected classification error converges to the Bayes error. A few other studies also establish weak consistency for eknn classifiers. [Hall and Samworth, 2005] studies a bag of infinitely many 1NN classifiers, each trained on $\tilde{n}_+$ positive points sampled from n_+ candidates and $\tilde{n}_-$ negative points sampled from n_- candidates, and shows that the classifier is weakly consistent if $\tilde{n}_+, \tilde{n}_-$ diverge and $\frac{\tilde{n}_+}{n_+}, \frac{\tilde{n}_-}{n_-}, \frac{\tilde{n}_+}{n_+} - \frac{\tilde{n}_-}{n_-}$ all converge to zero. [Biau and Devroye, 2010] shows that a bag of m 1NN classifiers is weakly consistent if $m, \tilde{n}$ diverge and $\frac{\tilde{n}}{n} \to 0$. [Biau et al., 2008] further shows that a probabilistic version of this classifier, where $\tilde{n}$ depends on a sampling probability q_n for each candidate point, is weakly consistent if $\mathbb{E}\tilde{n} = nq_n$ diverges and $q_n \to 0$.

Strong consistency often requires stronger conditions. kNN is shown to be strongly consistent if $k \to \infty$ and $k/n \to 0$ [Devroye et al., 2013, Chapter 11]. 1NN is not strongly consistent [Cover and Hart, 1967], but its variants based on compressed training data are strongly consistent [Kontorovich and Weiss, 2015, Hanneke et al., 2020]. To the best of our knowledge, strong consistency for ekNN has not been established. This paper presents the first such guarantee.

3 PRELIMINARIES

In this section, we introduce notation and concepts. When clarity is needed, a probability (or expectation) is denoted by $\Pr_Z$ (or $\mathbb{E}_Z$) when taken with respect to the randomness of Z. As in prior studies, we focus on binary classification.

Let $\mathcal{X}$ be a data space equipped with norm $||\cdot||$ and $\mathcal{Y} = \{0, 1\}$ be a label set. Let $B(x, r) = \{x' \in \mathcal{X}; ||x - x'|| \leq r\}$ denote a closed ball centered at x with radius r. For $\mathcal{X}$, let D be its input dimension and d be its *intrinsic dimension* introduced in [Xue and Kpotufe, 2018], i.e., an integer for which there exists a constant $C_d > 0$ such that $\Pr_z\{z \in B(x, r)\} \geq \min(C_d r^d, 1)$ for all $x \in \mathcal{X}$ and $r > 0$. While

most known rates depend on D, our rates depend on d.

Let $X \in \mathcal{X}$ be a random data point and $Y \in \mathcal{Y}$ be its label. The posterior function η is defined by

$$\eta(x) = \mathbb{E}_Y[Y \mid X = x].$$

The following assumption is made throughout the analysis.

Assumption (Hölder Smooth)**.** *The posterior function η is (λ, υ)-Hölder smooth for constants $\lambda, \upsilon > 0$, i.e., for any $x, x' \in \mathcal{X}$, we have $|\eta(x) - \eta(x')| \leq \lambda \cdot ||x - x'||^{\upsilon}$.*

Let $\hat{y}$ be a classifier and $er(\hat{y}) = \Pr_{X,Y}\{\hat{y}(X) \neq Y\}$ be its error. Let $\hat{y}_*$ be a *Bayes classifier*, i.e., for all $x \in \mathcal{X}$:

$$\hat{y}_*(x) = \mathbb{I}\{\eta(x) \geq 1/2\},$$

where $\mathbb{I}$ is the indicator function.

For a classifier $\hat{y}$ trained on a random sample $Z \subseteq \mathcal{X} \times \mathcal{Y}$, its *expected excess error* is defined as

$$||\hat{y} - \hat{y}_*||_e := \mathbb{E}_Z[er(\hat{y})] - er(\hat{y}_*). \tag{1}$$

Unless otherwise specified, all convergence rates discussed in this paper are w.r.t. the expected excess error in (1).

3.1 BOLTZMANN MARGIN

We first revisit two popular margin conditions. Let α, β, γ, t_* be positive constants. A distribution on $\mathcal{X} \times \mathcal{Y}$ is said to have a t_* Massart margin [Massart and Nédélec, 2006] if

$$\Pr_X\{|1/2 - \eta(X)| \leq t_*\} = 0,$$

and a (α, β) *Tsybakov margin* [Tsybakov, 2004] if

$$\forall t > 0, \ \Pr_X\{|1/2 - \eta(X)| \leq t\} \leq \alpha t^{\beta}.$$

It is clear that Massart margin is much stronger than Tsybakov margin. In the following, we introduce a new margin condition that interpolates between them. It is named after the Boltzmann factor in physics [Lei and Lan, 2020] due to the similarity of their forms.

Definition (Boltzmann Margin)**.** *A distribution on $\mathcal{X} \times \mathcal{Y}$ is said to have a (α, β, γ) Boltzmann margin if*

$$\forall t > 0, \ \Pr_X\{|1/2 - \eta(X)| \leq t\} \leq \alpha \exp(-\gamma t^{-\beta}). \tag{2}$$

Figure 1 illustrates the relations among the margins. The three Boltzmann curves are generated with $\beta = 0.1, 0.4, 0.8$, respectively, while fixing α, γ. As β increases, the Boltzmann curve approaches the Massart curve. As β decreases, it approaches the Tsybakov curve and even becomes weaker in certain regions. These relations illustrate the flexibility of the Boltzmann margin in convergence analysis.

We now formalize the relations among the margins. The first lemma shows that Massart margin implies Boltzmann margin, and the reverse holds under suitable parameter choices.

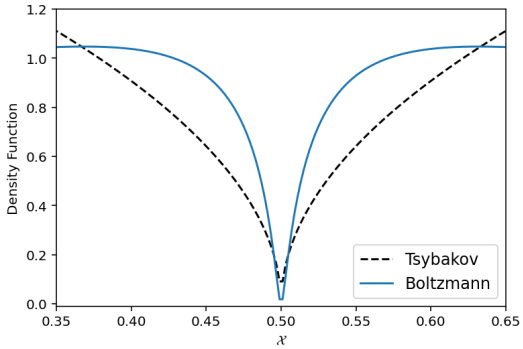

Figure 2: Example Data Densities

Lemma 3.1. *A t_* Massart margin implies a (α, β, γ) Boltzmann margin for $\alpha, \beta, \gamma > 0$ satisfying $\alpha \exp(-\frac{\gamma}{t_*^{\beta}}) = 1$. Reversely, for any $t_* > 0$, a $(1, \beta, t_*^{\beta})$ Boltzmann margin approaches the t_* Massart margin condition as $\beta \to \infty$, provided that $|1/2 - \eta(X)| \neq t_*$ almost everywhere.*

The next lemma shows Tsybakov margin implies Boltzmann margin and the reverse holds far from the decision boundary.

Lemma 3.2. *A (α, β, γ) Boltzmann margin implies a $(\frac{\gamma}{\alpha}, \beta)$ Tsybakov margin. Reversely, a (α, β) Tsybakov margin satisfies the $(\alpha', \beta', \gamma')$ Boltzmann margin property (2) for $t \in [(\frac{\gamma'}{\ln(\alpha'/\alpha)})^{1/\beta'}, 0.5]$, if $\alpha', \beta' > 0$ and $\gamma' \leq \ln(\alpha'/\alpha)$.*

Figure 2 illustrates the reversed relation. The decay is much slower under Boltzmann margin except near the decision boundary. Although this may seem less significant from the margin perspective, where the focus is often on near-boundary properties, it is highly relevant to machine learning. The reason is that far-boundary properties often govern non-asymptotic model behavior, which is particularly important in practice because training data are finite.

To better understand the above implication, we first establish a monotonicity property of the Boltzmann margin.

Lemma 3.3. *A (α, β, γ) Boltzmann margin implies a $(\alpha', \beta', \gamma')$ Boltzmann margin whenever (i) $\alpha' \geq \alpha$, (ii) $\beta' \geq \beta$ or (iii) $\gamma' \leq \gamma$, with the remaining parameters fixed.*

Now consider a model trained on n data points. Suppose data distribution has a (α, β) Tsybakov margin, and set $t = 1/n$. Then Lemma 3.2 implies the distribution also satisfies the $(\alpha', 1, \gamma')$ Boltzmann margin property whenever

$$2 \leq n \leq \gamma'^{-1} \ln(\alpha'/\alpha).$$

By Lemma 3.3, one may pick $\gamma' \leq 0.001$ and $\alpha' \geq \alpha^2$, yielding the range $n \in [2, 2000]$. This suggests the model may exhibit Boltzmann-margin behavior for moderate n.

Remark 3.4. *It is worth noting that in Lemma 3.2, for a (α, β) Tsybakov margin to imply a Boltzmann margin, the*

constraint on t does not depend on β. This is because we set $t \leq 0.5$ so that $t^\beta \leq 1$ for any β. This setting is often sufficient; for example, when $t = 1/n$, it covers all $n \geq 2$.

4 RESULTS FOR KNN CLASSIFIER

Let S be an i.i.d. sample of n points from $\mathcal{X} \times \mathcal{Y}$. The kNN regression estimator for $\eta(x)$ based on S is

$$\hat{\eta}(x; S) = \frac{1}{k} \sum_{(x',y') \in S} \mathbb{I}\{x' \in N_k(x; S) \wedge y' = 1\},$$

where $N_k(x; S)$ is the set of k nearest neighbors of x in S. The induced kNN classifier is defined by

$$\hat{y}(x) = \mathbb{I}\{\hat{\eta}(x; S) \geq 1/2\}.$$

As in prior studies, we assume ties occur with probability zero e.g., ties in the distance to the kth nearest neighbor or in the decision rule $\eta(x) = 1/2$.

Our main result is stated as follows.

Theorem 4.1. *Suppose data distribution has a (α, β, γ) Boltzmann margin. Let $C_\upsilon = 2 + d/\upsilon$. If $k = O(n^{\frac{\beta+2}{\beta+C_\upsilon}})$, then for $n \geq (C' \ln n)^{1+C_\upsilon/\beta}$:*

$$||\hat{y} - \hat{y}_*||_e \leq C_\alpha \exp\left(-C_\gamma n^{\frac{\beta}{\beta+C_\upsilon}}\right),$$

where $C_\alpha = 2 + \alpha$, $C_\gamma = \min(\gamma, 1/C')$ and $C' > 0$ is a constant depending on λ, υ and the structure of $\mathcal{X}$.

Theorem 4.1 establishes the first near-exponential convergence rate for kNN classification, and the exponent depends on β, d and υ. In the following, we discuss its implications.

Larger β makes the rate closer to exponential. This makes sense because, by Lemma 3.1, increasing β causes Boltzmann margin to approach Massart margin, and the latter is known to warrant exponential rates [Cabannnes and Vigogna, 2023, Vigogna et al., 2022]. (However, this does not trivially imply that kNN has an exponential rate under Massart margin. We establish such a result later.)

Smaller d/υ makes the rate closer to exponential. This is consistent with prior results, which show smaller D/υ implies faster rates (Table 1). A minor difference is that our rate depends on the intrinsic dimension d instead of D.

The theorem also yields several other observations. First, larger $\upsilon\beta/d$ permits a slower growth rate for k but also leads to slower convergence – a similar tradeoff appears in prior kNN convergence guarantees. Second, the assumed scaling rate for k is slightly faster than in prior results, as we assume $k = O(n^{\frac{2+\beta}{2+\beta+d/\upsilon}})$ while they assume $k = O(n^{\frac{2}{2+d/\upsilon}})$ or $O(n^{\frac{2}{2+D/\upsilon}})$. Third, our rate holds for large n, which is a common constraint in existing rates that are either asymptotic [Gyorfi, 2006, Reeve and Kabán, 2019, Rimanic et al., 2020, Xue and Kpotufe, 2018] or multi-phase Yang and Zhang [2025]. The next remark shows that, under suitable conditions, the large-n condition in Theorem 4.1 can be removed without affecting the convergence rate.

Remark 4.2. *Theorem 4.1 implies that, for $n \geq 1$:*

(i) If $k = O(n^{\frac{\beta+2}{\beta+C_\upsilon}})$, then there exists a constant $\tilde{C}_\alpha > 0$ such that $||\hat{y} - \hat{y}_*||_e \leq \tilde{C}_\alpha \exp\left(-C_\gamma n^{\frac{\beta}{\beta+C_\upsilon}}\right)$.

(ii) Take $\beta \to \infty$. Then $||\hat{y} - \hat{y}_*||_e \leq C'_\alpha \exp(-C_\gamma n)$ for $k = O(n)$, where $C'_\alpha = \lim_{\beta\to\infty} \tilde{C}_\alpha < \infty$.

Claim (ii) suggests that Boltzmann margin may warrant an exponential rate under sufficiently fast data decay. This is not a trivial consequence of (i), since one must verify that $\tilde{C}_\alpha$ remains bounded as $\beta \to \infty$. Combined with Lemma 3.1, this implies an exponential and non-asymptotic convergence rate for kNN classification under Massart margin, which is formalized in the next subsection. Hereafter, unless otherwise specified, the constants C_γ, C', C_υ are as defined in Theorem 4.1, while $\tilde{C}_\alpha$, C'_α are as defined in Remark 4.2.

4.1 EXTENDED RESULTS

In this section, we present three corollaries of Theorem 4.1. The first one provides a probabilistic convergence rate.

Corollary 4.3 (Probabilistic)**.** *Suppose data distribution has a (α, β, γ) Boltzmann margin. If $k = O(n^{\frac{\beta+2}{\beta+C_\upsilon}})$, then for any $\delta > 0$ and $n \geq 1$:*

$$er(\hat{y}) - er(\hat{y}_*) \leq \tilde{C}_\alpha \delta^{-1} \exp\left(-C_\gamma n^{\frac{\beta}{\beta+C_\upsilon}}\right),$$

with probability at least $1 - \delta$ over the random choice of S.

This probabilistic rate is near-exponential, in contrast to the polynomial probabilistic rates in [Reeve and Kabán, 2019, Rimanic et al., 2020]. Moreover, the dependence on δ is multiplicative in our bound, whereas it is additive in the prior bounds, e.g., $\tilde{O}(n^{-\frac{\upsilon(1+\beta)}{2\upsilon+1}} + \delta)$. With the common choice $\delta = O(1/n)$, our bound remains near-exponential.

Theorem 4.1 also yields convergence rates under other margins. The first is under Tsybakov margin.

Corollary 4.4 (Tsybakov)**.** *Suppose data distribution has a (α', β') Tsybakov margin and let $\alpha, \beta, \gamma > 0$ be constants satisfying $\gamma \leq \ln(\alpha/\alpha')$. If $k = O(n^{\frac{2+\beta}{C_\upsilon+\beta}})$, then*

$$||\hat{y} - \hat{y}_*||_e \leq \tilde{C}_\alpha \exp(-C_\gamma n^{\frac{\beta}{\beta+C_\upsilon}}),$$

for $2 \leq n \leq \left(\frac{1}{\gamma}\ln(\alpha/\alpha')\right)^{1+C_\upsilon/\beta}$.

This can be viewed as a multi-phase convergence rate, where the rate is near-exponential within the above range and becomes polynomial beyond it. Multi-phase rates have recently

been studied in [Yang and Zhang, 2025]. The authors show that the rate can be improved to $O(n^{-3/4})$ when $n \leq D^2$, but all rates remain polynomial. By contrast, we show the rate can be near-exponential over a finite range of n, and this range can be $O(D^2)$ when α is large or γ is small.

It is worth noting that β' does not appear in the error bound or the range of n, for the reason explained in Remark 3.4. Moreover, under suitable conditions, the bound may be expressed in terms of the Tsybakov-margin parameters and the admissible range only. For example, when α is large, γ is small, and $\beta \to \infty$, it follows that the range becomes nearly $n \leq \frac{1}{\gamma}\ln(\alpha/\alpha') := B$ and the error bound becomes $\frac{n}{B}\exp(-\ln(\alpha/\alpha')+\ln(\alpha+2))$; optimizing it over α yields a bound independent of the Boltzmann margin parameters.

The last corollary specializes the rate under Massart margin.

Corollary 4.5 (Massart)**.** *Let $k = O(n)$. If data distribution has a t_* Massart margin, then*

$$||\hat{y} - \hat{y}_*||_e \leq C\exp(-C_* n), \tag{3}$$

for all $n \geq C_*^{-1}\ln n$, where C and $C_* \propto t_*^{C_\upsilon}$ are positive constants. Moreover, for every $t_* > 0$, there exists a data distribution satisfying the t_* Massart margin for which (3) holds for all $n \geq 1$ (possibly with different constants).

The corollary establishes an exponential convergence rate for kNN classification under Massart margin. We can compare it with the rate in [Chaudhuri and Dasgupta, 2014, Appendix C.4]. First, the two rates are developed using fundamentally different techniques (ours via Boltzmann margin). Second, both rates are $O(\exp(-Cn))$, where $C \propto t_*^p$ for some constant p. The prior rate has $p = 2 + \frac{1}{\alpha'}$, where $\alpha' = \upsilon/D$ is the generalized Hölder parameter, while we have $p = C_\upsilon = 2 + \frac{d}{\upsilon} = 2 + \frac{d}{D}\frac{1}{\alpha'}$. Since $d/D < 1$, our p is smaller, implying a larger t_*^p for $t_* \in (0,1)$. Consequently, the exponential constant in our bound increases slightly faster with t_*. Finally, both rates require $k = O(n)$.

The corollary contains two distinct claims: one holds under arbitrary Massart margins but only yields an asymptotic rate, while the other only holds under certain Massart margins but yields a non-asymptotic rate. Moreover, condition $k = O(n)$ is stronger than $O(n^{\frac{\beta+2}{\beta+C_\upsilon}})$ used in near-exponential rates.

5 RESULTS FOR EKNN CLASSIFIER

In this section, we present new convergence rates for ensemble kNN (ekNN) classification under Boltzmann margin.

For ensemble classifiers, the excess error definition (1) requires a slight adjustment to account for the additional randomness introduced by bootstrapping. There are two common settings. One assumes a training set is sampled from the population and bootstrap samples are drawn from it. In this case, one can make claims that account for the additional bootstrapping randomness, e.g., [Xue and Kpotufe, 2018] establishes error rates with high probability over the random choices of both the training set and bootstrap samples. The other assumes bootstrap samples are drawn directly from the population and treats their collection as the training set, e.g., [Biau and Devroye, 2010]; in this case, (1) incorporates all randomness, so no special treatment is needed. For convenience, we adopt the second setting, but our results also apply to the first setting (Remark 5.2).

Let $\tilde{S} := \{S_1, \ldots, S_m\}$ be a collection of m samples, where each S_j contains $\tilde{n}$ points i.i.d. sampled from $\mathcal{X} \times \mathcal{Y}$, and all samples are drawn i.i.d. Hence $n = m\tilde{n}$. Consider the following bagged kNN regression estimator [Sutton, 2005, Hastie et al., 2009, Hang et al., 2022] based on $\tilde{S}$:

$$\hat{\eta}_e(x;\tilde{S}) = \frac{1}{m}\sum_{j=1}^{m}\hat{\eta}(x;S_j).$$

The induced ekNN classifier is

$$\hat{y}_e(x) = \mathbb{I}\{\hat{\eta}_e(x;\tilde{S}) \geq 1/2\}.$$

Our main result is stated as follows.

Theorem 5.1. *Suppose data distribution has a (α,β,γ) Boltzmann margin. If $k = O(\tilde{n}^{\frac{\beta+2}{\beta+C_\upsilon}})$, then*

$$||\hat{y}_e - \hat{y}_*||_e \leq C_\alpha\exp(-C_\gamma\tilde{n}^{\frac{\beta}{\beta+C_\upsilon}}), \tag{4}$$

for $\tilde{n} \geq (2\max\{\ln\tilde{n}, 2\ln m\}/C')^{1+C_\upsilon/\beta}$.

Moreover, if $m = O(\sqrt{\tilde{n}})$, then (4) holds for all $\tilde{n} \geq 1$, with C_α replaced by a possibly different constant $\tilde{C}_\alpha > 0$.

We have some interesting observations from the theorem. First, recall that the known convergence rates for ekNN classifiers are polynomial and typically assume $\tilde{n} = \Theta(n^g)$ for some constant g, e.g., $\frac{D}{2+D}$ or $\frac{\upsilon+d}{2\upsilon+d}$ (Section 2.2). Under the same choice of $\tilde{n}$, our theorem yields

$$||\hat{y}_e - \hat{y}_*||_e = O(e^{-C_\gamma n^{\frac{g\beta}{\beta+C_\upsilon}}}).$$

This rate is slightly weaker than the one we established for kNN classification, but can be close when g is near 1, e.g., under the choice $g = \frac{D}{2+D}$ with large D. As a result, it is generally faster than the known polynomial rates.

Second, unlike [Biau et al., 2010, Samworth, 2012a], our rate does not require $m = \infty$. In fact, if $m = \infty$, our rate only holds for $\tilde{n} = \infty$, which limits its practical relevance. That said, Theorem 5.1 does allow m to diverge no faster than $\tilde{O}(e^{\tilde{n}})$. A related condition appears in [Qiao et al., 2019], which requires m to diverge no faster than $O(\tilde{n}^{2\mathcal{V}/D})$ to ensure $k = O(\tilde{n}^{\frac{2\upsilon}{2\upsilon+D}}m^{-\frac{D}{2\upsilon+D}}) \to \infty$. Our condition is weaker in two respects: (i) it allows but does not require the divergence of m; (ii) $e^{\tilde{n}}$ grows exponentially faster than $\tilde{n}^{2\mathcal{V}/D}$, allowing m to diverge substantially faster.

Finally, Theorem 5.1 also applies to the alternative bootstrap setting discussed earlier in this section. In the following remark, conditions (ii) and (iii) are standard in ekNN analysis, e.g., [Xue and Kpotufe, 2018, Qiao et al., 2019].

Remark 5.2. *Theorem 5.1 also applies when (i) a training set S is sampled i.i.d. from the population, (ii) subsets $S_1, \ldots, S_m$ are drawn i.i.d. from S, and (iii) in each subset, points are sampled without replacement.*

5.1 OTHER RESULTS

Theorem 5.1 also implies a probabilistic convergence rate.

Corollary 5.3 (Probabilistic)**.** *Suppose data distribution has (α, β, γ) Boltzmann margin. If $k = O(\tilde{n}^{\frac{\beta+2}{\beta+C_\upsilon}})$ and $m = O(\sqrt{\tilde{n}})$, then for any $\delta > 0$ and $\tilde{n} \geq 1$,*

$$er(\hat{y}_e) - er(\hat{y}_*) \leq \tilde{C}_\alpha \delta^{-1} \exp(-C_\gamma \tilde{n}^{\frac{\beta}{\beta+C_\upsilon}}),$$

with probability at least $1 - \delta$ over the random choice of $\tilde{S}$, where $\tilde{C}_\alpha$ is as defined in Theorem 5.1.

We can compare the above with the asymptotic rate for the bagged 1NN classifier in [Xue and Kpotufe, 2018], which is $\tilde{O}\left(n^{-1}\ln(1/\delta)\right)^{\frac{1}{2+d/\upsilon}}$ if $\frac{\tilde{n}}{n} = \Omega(n^{-\frac{\upsilon}{2\upsilon+d}})$. Under the same choice of $\tilde{n}/n$, our bound becomes $O(e^{-C_\gamma n^{\frac{1+d}{C_\upsilon}\cdot\frac{\beta}{\beta+C_\upsilon}}})$. This rate is weaker than the asymptotic rate we established for kNN, but can be close when d/υ is small or β is large – in either case, our bound remains near-exponential and can be faster than the prior polynomial rate.

Our next result is a strong consistency guarantee for ekNN. Recall that a classifier $\hat{y}$ trained on a random sample Z is strongly Bayes consistent if $\Pr_Z\{\lim_{|Z|\to\infty} er(\hat{y}) = er(\hat{y}_*)\} = 1$ [Devroye et al., 2013].

Theorem 5.4 (Strong Consistency)**.** *Suppose data distribution has a (α, β, γ) Boltzmann margin. Then $\hat{y}_e$ is strongly Bayes consistent if $k = O(\tilde{n}^{\frac{\beta+2}{\beta+C_\upsilon}})$ and $m = \tilde{O}(e^{\tilde{n}})$.*

To the best of our knowledge, this is the first result that formally establishes strong consistency for ekNN classification. We can compare its condition on k with those for weak consistency, such as $O(n^{\frac{4}{4+D}})$ or $O(n^{\frac{\upsilon(1+\beta)}{2\upsilon+D}})$ (Section 2.2). Our $O(\tilde{n}^{\frac{\beta+2}{\beta+2+d/\upsilon}})$ has similar implications: k needs to scale with $\tilde{n}$, and a larger data dimension allows it to scale more slowly. A major difference is in the impact of β: in prior conditions, an arbitrarily large β can require k to scale arbitrarily fast alongside n, e.g., $k = O(n^5)$; in our condition, $k = O(\tilde{n})$ even when $\beta \to \infty$.

6 NUMERICAL RESULTS

We first construct a distribution that satisfies Boltzmann margin and sample data from it for empirical evaluation.

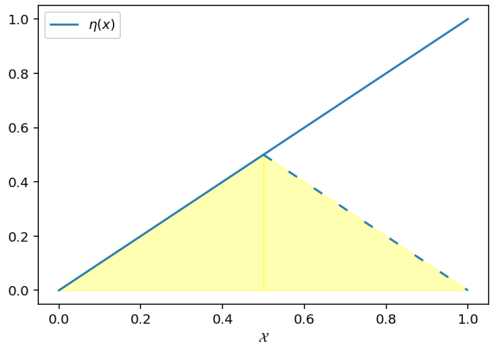

Figure 3: Posterior Function

Consider $\mathcal{X} = [0, 1]$ with posterior $\eta(X) = X$. The Bayes decision boundary is at $X = 0.5$. Figure 3 illustrates the posterior function and the corresponding Bayes error under a uniform distribution (i.e., the area of the highlighted region).

To evaluate the impact of Boltzmann margin, we construct a family of distributions whose density decays near the Bayes decision boundary and remains constant elsewhere. Given constants $\alpha, \beta, \gamma > 0$, define the density function

$$f_b(x) = \begin{cases} \frac{\alpha\beta\gamma}{2N_T} \cdot \frac{e^{-\frac{\gamma}{|x-0.5|^\beta}}}{|x-0.5|^{\beta+1}}, & \text{if } |x-0.5| \leq T, \\ \frac{\alpha\beta\gamma}{2N_T} \cdot \frac{e^{-\gamma T^{-\beta}}}{T^{\beta+1}}, & \text{if } |x-0.5| > T. \end{cases} \tag{5}$$

It can be shown that the distribution induced by f_b satisfies (α, β, γ) Boltzmann margin. The detailed construction and verification are presented in Appendix D.1.

Remark. *In Figure 1, the Boltzmann-margin density is generated based on (5) with $(T, \gamma) = (0.15, 10)$ and $\beta = 0.1, 0.2, 0.8$ for Boltzmann curves 1, 2, 3 respectively.*

Other densities are constructed in similar ways. For Massart margin, it is $f_t(x) = \begin{cases} 0, & |x-0.5| \leq T \\ \frac{1}{1-2T}, & |x-0.5| > T, \end{cases}$. For (α', β') Tsybakov margin, it is

$$f_t(x) = \begin{cases} \frac{\alpha'\beta'}{2N'_T} \cdot |x-0.5|^{\beta'-1}, & \text{if } |x-0.5| \leq T, \\ \frac{\alpha'\beta'}{2N'_T} \cdot T^{\beta'-1}, & \text{if } |x-0.5| > T, \end{cases}$$

where $N'_T = \alpha' T^{\beta'} + 2C'_T(0.5 - T)$ and $C'_T = \frac{\alpha'\beta'}{2} T^{\beta'-1}$. In the figure, $(\alpha', \beta') = (1, 1.5)$ and $T = 0.15$ for both.

After constructing Boltzmann-margin distribution, we generate data points from it by applying the rejection sampling technique Bishop and Nasrabadi [2006]. The detailed sampling procedure is described in Appendix D.2.

6.1 CLASSIFICATION PERFORMANCE

For empirical evaluation, 1000 points are generated i.i.d. with $T = 0.15$ and $\alpha = 1$. We randomly shuffle them,

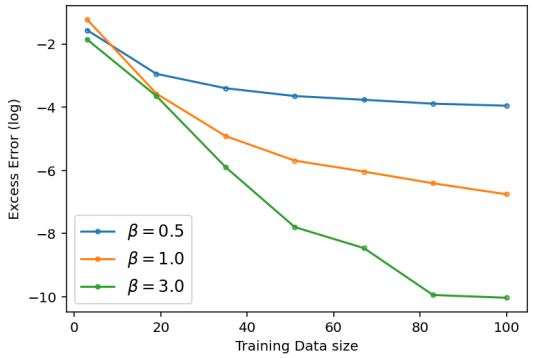

Figure 4: kNN Classification Performance

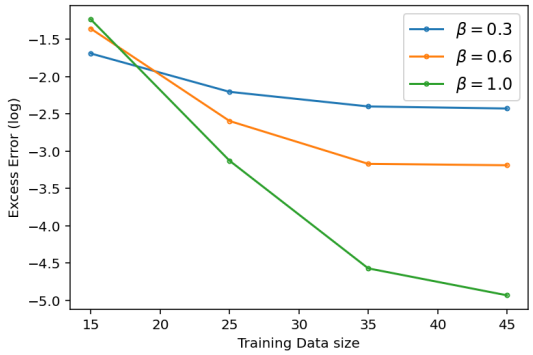

Figure 5: ekNN Classification Performance

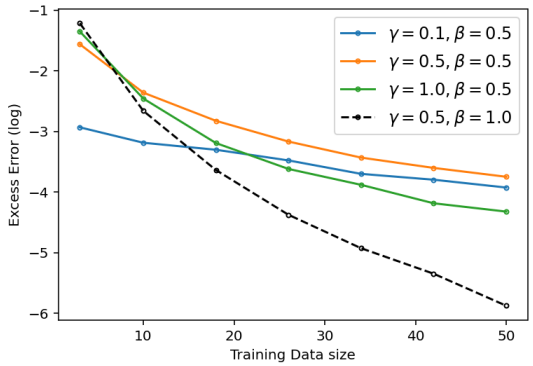

Figure 6: kNN Performance based on γ's

use the first 25% for testing, and sample from the rest for training (to simulate varying training data sizes). For ekNN, the training data sampling procedure follows the conditions in Remark 5.2. All results are averaged over 1000 shuffles.

We evaluate excess error on the test set and estimate the Bayes error using the full data pool. For kNN, we set $k = C_k n^{\frac{\beta+2}{\beta+C_v}}$ according to Theorem 4.1, where $C_v = 2 + d/v = 3$ and C_k is chosen so that $k = 3$ when $n = 5$. For ekNN, we set $m = 5$ and $k = \tilde{C}_k \tilde{n}^{\frac{\beta+2}{\beta+C_v}}$ according to Theorem 5.1, where $\tilde{C}_k$ is chosen in the same way. All k values are rounded up to the nearest odd integers.

Figure 4 shows the excess error of kNN classifier with $k = 3$, $\gamma = 0.5$ and various values of β. The y-axis is plotted on a logarithmic scale, so straighter curves indicate rates closer to exponential. As β increases, the excess error curve becomes increasingly straight, which agrees with our theoretical result that larger β implies convergence closer to exponential.

Figure 5 shows the performance of ensemble kNN classifier with $k = 3$, $m = 5$, $\gamma = 0.5$ and various values of β. As β increases, the excess error curve becomes increasingly straight, suggesting convergence rates closer to exponential.

Another interesting observation arises from comparing Figures 4 and 5. The ekNN error curves appear straighter than those of kNN, suggesting a potential advantage of ekNN. However, neither our present theory nor existing ekNN convergence rates (Table 1) explain this gap, for they largely inherit the rates of kNN. Understanding this gap remains an interesting direction for future research.

Finally, we evaluate the impact of γ. Figure 6 shows the kNN excess error based on $k = 3$ and $\beta = 0.5$. As γ increases, the error decreases more rapidly, but the error curves do not become substantially straighter, especially when compared with the dashed curve obtained by increasing β. This agrees with our theoretical implications: γ can accelerate convergence, but unlike β, it does not make the convergence rate closer to exponential. This explains why the dashed error curve is noticeably straighter.

7 CONCLUSION AND DISCUSSION

This paper introduces Boltzmann margin, a new margin condition that enables near-exponential convergence rates for classification. We demonstrate its application to kNN and ekNN classifiers and establish their first near-exponential rates, as well as the first strong consistency guarantee for ekNN. Extending Boltzmann margin to broader hypothesis classes remains an important direction for future research.

Besides its theoretical contributions, the present study also has practical implications. For example, Lemma 3.2 suggests that, for distributions satisfying a Tsybakov margin, one may expect faster error decay in the small-data regime than that implied by the margin alone. Furthermore, improving error rates may not require finding a feature space (e.g., through data embedding) in which a strong Massart margin holds. Such a representation can be difficult to obtain and may involve substantial tradeoffs. Instead, one may identify a feature space satisfying Boltzmann margin while still achieving near-exponential error convergence.

8 ACKNOWLEDGMENT

We thank all anonymous reviewers for their thorough and constructive feedback that improves the quality of this work.

References

Gérard Biau and Luc Devroye. On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *Journal of Multivariate Analysis*, 101(10):2499–2518, 2010.

Gérard Biau, Luc Devroye, and Gábor Lugosi. Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9(66):2015–2033, 2008.

Gérard Biau, Frédéric Cérou, and Arnaud Guyader. On the rate of convergence of the bagged nearest neighbor estimate. *Journal of Machine Learning Research*, 11(2), 2010.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Number 4. Springer, 2006.

Vivien Cabannnes and Stefano Vigogna. A case of exponential convergence rates for svm. In *International Conference on Artificial Intelligence and Statistics*, pages 359–374. PMLR, 2023.

Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for nearest neighbor classification. *Advances in Neural Information Processing Systems*, 27, 2014.

George H Chen, Devavrat Shah, et al. Explaining the success of nearest neighbor methods in prediction. *Foundations and Trends® in Machine Learning*, 10(5-6):337–588, 2018.

Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

Enyan Dai and Suhang Wang. Towards self-explainable graph neural network. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 302–311, 2021.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

Maik Döring, László Györfi, and Harro Walk. Rate of convergence of k-nearest-neighbor classification rule. *Journal of Machine Learning Research*, 18(227):1–16, 2018.

Sébastien Gadat, Thierry Klein, and Clément Marteau. Classification in general finite dimensional spaces with the k-nearest neighbor rule. *The Annals of Statistics*, 44(3): 982–1009, 2016.

L Gyorfi. The rate of convergence of k_n-nn regression estimates and classification rules (corresp.). *IEEE Transactions on Information Theory*, 27(3):362–364, 2006.

László Györfi and Roi Weiss. Universal consistency and rates of convergence of multiclass prototype algorithms in metric spaces. *Journal of Machine Learning Research*, 22(151):1–25, 2021.

Peter Hall and Richard J Samworth. Properties of bagged nearest neighbour classifiers. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(3): 363–379, 2005.

Hanyuan Hang, Yuchao Cai, Hanfang Yang, and Zhouchen Lin. Under-bagging nearest neighbors for imbalanced classification. *Journal of machine learning research*, 23 (118):1–63, 2022.

Steve Hanneke, Aryeh Kontorovich, Sivan Sabato, and Roi Weiss. Universal bayes consistency in metric spaces. In *2020 Information Theory and Applications Workshop (ITA)*, pages 1–33, 2020.

Trevor Hastie, Robert Tibshirani, Jerome Friedman, et al. The elements of statistical learning, 2009.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2020.

Michael Kohler and Adam Krzyzak. On the rate of convergence of local averaging plug-in classification rules under a margin condition. *IEEE transactions on information theory*, 53(5):1735–1742, 2007.

Aryeh Kontorovich and Roi Weiss. A bayes consistent 1-nn classifier. In *Artificial Intelligence and Statistics*, pages 480–488. PMLR, 2015.

Sanjeev R Kulkarni and Steven E Posner. Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Transactions on Information Theory*, 41 (4):1028–1039, 2002.

Zijian Lei and Liang Lan. Improved subsampled randomized hadamard transform for linear svm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 04, pages 4519–4526, 2020.

Minghan Li, Xilun Chen, Ari Holtzman, Beidi Chen, Jimmy Lin, Scott Yih, and Victoria Lin. Nearest neighbor speculative decoding for llm generation and attribution. *Advances in Neural Information Processing Systems*, 37: 80987–81015, 2024.

Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006.

Xingye Qiao, Jiexin Duan, and Guang Cheng. Rates of convergence for large-scale nearest neighbor classification. *Advances in neural information processing systems*, 32, 2019.

Nazneen Fatema Rajani, Ben Krause, Wengpeng Yin, Tong Niu, Richard Socher, and Caiming Xiong. Explaining and improving model behavior with k nearest neighbor representations. *arXiv preprint arXiv:2010.09030*, 2020.

Henry Reeve and Ata Kabán. Fast rates for a knn classifier robust to unknown asymmetric label noise. In *International Conference on Machine Learning*, pages 5401–5409. PMLR, 2019.

Luka Rimanic, Cedric Renggli, Bo Li, and Ce Zhang. On convergence of nearest neighbor classifiers over feature transformations. *Advances in Neural Information Processing Systems*, 33:12521–12532, 2020.

Richard J Samworth. Optimal weighted nearest neighbour classifiers. *The Annals of Statistics*, pages 2733–2763, 2012a.

Richard J Samworth. Supplement to "optimal weighted nearest neighbour classifiers.". *arXiv preprint arXiv:1101.5783*, 2012b.

Clifton D Sutton. Classification and regression trees, bagging, and boosting. *Handbook of statistics*, 24:303–329, 2005.

Alexander B Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1): 135–166, 2004.

Stefano Vigogna, Giacomo Meanti, Ernesto De Vito, and Lorenzo Rosasco. Multiclass learning with margin: exponential rates with no bias-variance trade-off. In *International Conference on Machine Learning*, pages 22260–22269. PMLR, 2022.

Lirong Xue and Samory Kpotufe. Achieving the time of 1-nn, but the accuracy of k-nn. In *International Conference on Artificial Intelligence and Statistics*, pages 1628–1636. PMLR, 2018.

Pengkun Yang and Jingzhao Zhang. Fast and multiphase rates for nearest neighbor classifiers. *Proceedings of Machine Learning Research vol*, 291:1–2, 2025.

Hang Zhang and Ping Li. Improved bound on generalization error of compressed knn estimator. In *International Conference on Artificial Intelligence and Statistics*, pages 7578–7593. PMLR, 2023.

Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms*. CRC press, 2025.

Near-Exponential Convergence Rates for kNN Classification based on Boltzmann Margin (Supplementary Material)

Luyuan Yang[1] **Shayan Shafaei**[1] **Chao Lan**[1]

[1]School of Computer Science, University of Oklahoma, Norman, Oklahoma, USA

CONTENTS

A BASIC RESULTS AND TOOLS

A.1 RELATIONS BETWEEN MARGINS (LEMMAS 3.1, 3.2, 3.3)

Lemma 3.1. *A t_* Massart margin implies a (α, β, γ) Boltzmann margin for $\alpha, \beta, \gamma > 0$ satisfying $\alpha \exp(-\frac{\gamma}{t_*^\beta}) = 1$. Conversely, for any $t_* > 0$, a $(1, \beta, t_*^\beta)$ Boltzmann margin approaches a t_* Massart margin condition as $\beta \to \infty$, provided that $|1/2 - \eta(X)| \neq t_*$ almost everywhere.*

Proof. Suppose data distribution has a t_* Massart margin and α, β, γ satisfies $\alpha e^{-\gamma t_*^{-\beta}} = 1$. For any $t \geq t_*$,

$$\Pr\{|1/2 - \eta(X)|\} \leq t\} \leq 1 = \alpha e^{-\gamma t_*^{-\beta}} \leq \alpha e^{-\gamma t^{-\beta}}. \tag{6}$$

When $t < t_*$, the above inequality also holds because the left side is zero due to Massart margin. This proves the first claim.

Reversely, for any $t_* > 0$, suppose the distribution has a $(\alpha, \beta, t_*^\beta)$ Boltzmann margin. Then, for any $t > 0$:

$$\Pr\{|1/2 - \eta(X)| \leq t\} \leq \alpha \exp(-t_*^\beta t^{-\beta}) = \alpha \exp\left(-(t_*/t)^\beta\right) := \Delta. \tag{7}$$

Consider three cases for t.

– If $t < t_*$, then $t_*/t > 1$ so that $\Delta \xrightarrow{\beta \to \infty} 0$.

– If $t > t_*$, then $t_*/t < 1$ so that $\Delta \xrightarrow{\beta \to \infty} \alpha$.

– Case $t = t_*$ happens with probability zero by assumption.

Putting all together, we have

$$\lim_{\beta \to \infty} \Pr\{|1/2 - \eta(X)| \leq t\} \leq \begin{cases} 0 & \text{if } t \leq t_* \\ \alpha & \text{if } t > t_* \end{cases} \tag{8}$$

This proves the Boltzmann margin converges to a t_* Massart margin. □

Lemma 3.2. *A (α, β, γ) Boltzmann margin implies a $(\frac{\alpha}{\gamma}, \beta)$ Tsybakov margin. Reversely, a (α, β) Tsybakov margin satisfies the $(\alpha', \beta', \gamma')$ Boltzmann margin property (2) for $t \in [(\frac{\gamma'}{\ln(\alpha'/\alpha)})^{1/\beta'}, 0.5]$, if $\alpha', \beta' > 0$ and $\gamma' \leq \ln(\alpha'/\alpha)$.*

Proof. Suppose a distribution has (α, β, γ) Boltzmann margin. Then for any $t > 0$,

$$\Pr\{|1/2 - \eta(x)|\} \leq t\} \leq \alpha e^{-\gamma t^{-\beta}} \leq \alpha \frac{1}{\gamma t^{-\beta}} = \frac{\alpha}{\gamma} t^\beta, \tag{9}$$

which implies $(\alpha/\gamma, \beta)$ Tsybakov margin. This proves the first claim.

Reversely, suppose a distribution has (α, β) Tsybakov margin. Let $\alpha', \beta', \gamma' > 0$ be any constants satisfying $\gamma' \leq \ln(\alpha'/\alpha)$. Then, one can verify that $(\frac{\gamma'}{\ln(\alpha'/\alpha)})^{1/\beta'} \leq t \leq 0.5$ implies

$$\alpha t^\beta \leq \alpha \leq \alpha' \exp(-\gamma' t^{-\beta'}), \tag{10}$$

where the first inequality follows $t^\beta \leq 1$ and the second follows the lower bound of t. Hence the $(\alpha', \beta', \gamma')$ Boltzmann margin. Note that β is not included in the condition on t because we set $t \leq 0.5$, in which case $t^\beta \leq 1$ for any $\beta > 0$. □

Lemma 3.3. *A (α, β, γ) Boltzmann margin implies a $(\alpha', \beta', \gamma')$ Boltzmann margin whenever (i) $\alpha' \geq \alpha$, (ii) $\beta' \geq \beta$ or (iii) $\gamma' \leq \gamma$, with the remaining parameters fixed.*

Proof. This property follows the definition. Recall (α, β, γ) Boltzmann margin means $\Pr\{|1/2 - \eta(x)|\} \leq t\} \leq \alpha e^{-\gamma t^{-\beta}}$. Apparently, the right side increases if α or β increases, or γ decreases. □

A.2 RELATION BETWEEN EXCESS ERROR AND ERROR PROBABILITY

In this section, we present some useful relations between excess error and error probability, which will be applied later.

Lemma A.1. *Let $\hat{y}$ be any classifier trained on sample S. For any fixed S, there is*

$$er(\hat{y}) - er(\hat{y}_*) \leq \Pr_X\{\hat{y}(X) \neq \hat{y}_*(X)\}. \tag{11}$$

Further, over the random choice of S, there is

$$\mathbb{E}_S[er(\hat{y})] - er(\hat{y}_*) \leq \Pr_{X,S}\{\hat{y}(X) \neq \hat{y}_*(X)\}; \tag{12}$$

and for any $t > 0$,

$$\Pr_S\{er(\hat{y}) - er(\hat{y}_*) > t\} \leq \frac{1}{t}\Pr_{X,S}\{\hat{y}(X) \neq \hat{y}_*(X)\}. \tag{13}$$

Proof. Let Y be the label of X. Since $\hat{y}(X) \neq Y$ implies $\hat{y}(X) \neq \hat{y}_*(X)$ or $\hat{y}_*(X) \neq Y$, there is

$$\Pr_{X,Y}\{\hat{y}(X) \neq Y\} \leq \Pr_X\{\hat{y}(X) \neq \hat{y}_*(X)\} + \Pr_{X,Y}\{\hat{y}_*(X) \neq Y\}, \tag{14}$$

which implies

$$\Pr_X\{\hat{y}(X) \neq \hat{y}_*(X)\} \geq \Pr_{X,Y}\{\hat{y}(X) \neq Y\} - \Pr_{X,Y}\{\hat{y}_*(X) \neq Y\} = er(\hat{y}) - er(\hat{y}_*). \tag{15}$$

This proves the first claim. The second claim follows by taking expectation w.r.t. S on both sides of (15), and the third claim follows by further employing the Markov inequality (based on the fact that $er(\hat{y}) - er(\hat{y}_*) \geq 0$). □

The above lemma shows $\Pr_X\{\hat{y}(X) \neq \hat{y}_*(X)\}$ is key for bounding excess error. The following lemma provides an observation that is useful for bounding $\Pr_X\{\hat{y}(X) \neq \hat{y}_*(X)\}$.

Lemma A.2. *Let $\hat{y}$ be a classifier defined as $\hat{y}(x) = \mathbb{I}\{\hat{y}(x; S) \geq 1/2\}, \forall x \in \mathcal{X}$. Then, almost surely, any $x \in X$ satisfying $|\hat{\eta}(x; S) - \eta(x)| \leq |1/2 - \eta(x)|$ also satisfies $\hat{y}(x) = \hat{y}_*(x)$.*

Proof. It is clear that $\hat{y}(x) = \hat{y}_*(x)$ if $\hat{\eta}(x; S)$ and $\eta(x)$ fall on the same side of 1/2, which is guaranteed if $|\hat{\eta}(x; S) - \eta(x)| \leq |1/2 - \eta(x)|$. This proves the lemma. □

A.3 OTHER TOOLS

We borrow the following kNN regression bound from prior work. The original bound is uniform and very strong, but we only need its implied pointwise bound. The following is a paraphrase of [Xue and Kpotufe, 2018, Proposition 1].

Proposition A.3. *Let $\mathcal{V}$ be the VC dimension of the class of balls on $\mathcal{X}$. Let $\delta < 1$ and $k = O((\ln \frac{2|Z|}{\delta})^{\frac{d}{2\upsilon+d}}(C_d|Z|)^{\frac{2\upsilon}{2\upsilon+d}})$. With probability at least $1 - 2\delta$ over the random sampling of Z, we have simultaneously for all $x \in \mathcal{X}$:*

$$|\hat{\eta}(x; Z) - \eta(x)| \leq C\left(\frac{\mathcal{V}\ln(2|Z|/\delta)}{C_d|Z|}\right)^{\frac{\upsilon}{2\upsilon+d}}, \tag{16}$$

where C is a function of υ, λ, and C_d is a constant depending on intrinsic data dimension d.

Corollary A.4. *Let $C_\upsilon = 2 + \frac{d}{\upsilon}$ and $C' = \frac{C_d}{\mathcal{V}C^{C_\upsilon}}$. Proposition A.3 implies for any $t > 0$, if $k = O(t^{\frac{d}{\upsilon}}|Z|C_d^{\frac{2\upsilon+2d}{2\upsilon+d}})$ then*

$$\Pr_{X,Z}\{|\hat{\eta}(X; Z) - \eta(X)| > t\} \leq 2|Z|\exp\left(-C'|Z|t^{C_\upsilon}\right). \tag{17}$$

Proof. Proposition A.3 implies for any $x \in X$, (16) holds with high probability. Setting the right side of (16) to t implies $\Pr_Z\{|\hat{\eta}(x; Z) - \eta(x)| > t\} \leq 2|Z|\exp\left(-\frac{C_d}{\mathcal{V}}|Z|\left(\frac{t}{C}\right)^{C_\upsilon}\right)$. Taking expectation w.r.t. x on both sides gives the corollary. □

We will also apply the classic Borel-Cantelli Lemma to prove storng consistency. It is quoted as follows.

Lemma A.5. *Let $E_1, E_2, \ldots$ be a sequence of events. If $\sum_{i=1}^{\infty}\Pr\{E_i\} < \infty$, then $\Pr\{\limsup_{i\to\infty} E_i\} = 0$.*

B RESULTS FOR KNN CLASSIFICATION

Theorem 4.1. *Suppose data distribution has a (α, β, γ) Boltzmann margin. Let $C_\upsilon = 2 + d/\upsilon$. If $k = O(n^{\frac{\beta+2}{\beta+C_\upsilon}})$, then for $n \geq (C' \ln n)^{1+C_\upsilon/\beta}$:*

$$||\hat{y} - \hat{y}_*||_e \leq C_\alpha \exp\left(-C_\gamma n^{\frac{\beta}{\beta+C_\upsilon}}\right),$$

where $C_\alpha = 2 + \alpha$, $C_\gamma = \min(\gamma, 1/C')$ and $C' > 0$ is a constant depending on λ, υ and the structure of $\mathcal{X}$.

Proof. By Lemma A.1 and Lemma A.2, for any $t > 0$,

$$\begin{aligned} ||\hat{y} - \hat{y}_*||_e &\leq \Pr_{X,S}\{\hat{y}(X) \neq \hat{y}_*(X)\} \\ &\leq \Pr_{X,S}\{|\hat{\eta}(X;S) - \eta(X)| > |1/2 - \eta(X)|\} \\ &\leq \Pr_{X,S}\{|\hat{\eta}(X;S) - \eta(X)| > t\} + \Pr_X\{|1/2 - \eta(X)| \leq t\}, \end{aligned} \tag{18}$$

where the last inequality is by case-studying $|1/2 - \eta(X)|$ based on t.

On the right side of (18), the second term can be bounded based on Boltzmann margin i.e,

$$\Pr_X\{|1/2 - \eta(X)| \leq t\} \leq \alpha \exp(-\gamma t^{-\beta}). \tag{19}$$

The first term can be bounded using Corollary A.4. This corollary sets $C_\upsilon = 2 + d/\upsilon$ and $C' = \frac{C_d}{\mathcal{V} C^{C_\upsilon}}$, where $C_d > 0$ is a constant depending on d and C is the constant in Lemma A.3 that depends on λ and υ. (This means C' depends on $\lambda, \upsilon, d, \mathcal{V}$.) The corollary says if $k = O(t^{\frac{d}{\upsilon}} n C_d^{\frac{2\upsilon+2d}{2\upsilon+d}})$ then

$$\Pr_{X,S}\{|\hat{\eta}(X;S) - \eta(X)| > t\} \leq 2n \exp(-C' n t^{C_\upsilon}). \tag{20}$$

Plugging this and (19) back to (18), we have

$$\Pr_{X,S}\{\hat{y}(X) \neq \hat{y}_*(X)\} \leq 2n \exp(-C' n t^{C_\upsilon}) + \alpha \exp(-\gamma t^{-\beta}) := \Delta. \tag{21}$$

To merge the two terms in Δ, we first assume $n \geq \frac{2}{C'} t^{-C_\upsilon} \ln n$, which is realizable for large enough n and a proper choice of t. Then, we have $n \leq \exp(\frac{1}{2} C' n t^{C_\upsilon})$ and thus

$$\Delta \leq 2 \exp(-\frac{1}{2} C' n t^{C_\upsilon}) + \alpha \exp(-\gamma t^{-\beta}). \tag{22}$$

Set $n t^{C_\upsilon} = t^{-\beta}$. Then $t = n^{-\frac{1}{C_\upsilon + \beta}}$ and the above implies

$$\Delta \leq (2 + \alpha) \exp\left(-\min(\frac{1}{2} C', \gamma) \cdot n^{\frac{\beta}{\beta+C_\upsilon}}\right). \tag{23}$$

Also, based on the above choice of t, the conditions on n and k become $n \geq (\frac{2}{C'} \ln n)^{\frac{C_\upsilon+\beta}{\beta}}$ and $k = O(n^{\frac{2+\beta}{C_\upsilon+\beta}})$. The theorem is proved by substituting $\frac{2}{C'}$ by C', and noting that C' depends on λ, υ as well as $d, \mathcal{V}$ (structure of $\mathcal{X}$). □

Remark 4.2. *Theorem 4.1 implies that, for $n \geq 1$:*

(i) If $k = O(n^{\frac{\beta+2}{\beta+C_\upsilon}})$, then there exists a constant $\tilde{C}_\alpha > 0$ such that $||\hat{y} - \hat{y}_*||_e \leq \tilde{C}_\alpha \exp\left(-C_\gamma n^{\frac{\beta}{\beta+C_\upsilon}}\right)$.

(ii) Take $\beta \to \infty$. Then $||\hat{y} - \hat{y}_*||_e \leq C'_\alpha \exp(-C_\gamma n)$ for $k = O(n)$, where $C'_\alpha = \lim_{\beta\to\infty} \tilde{C}_\alpha < \infty$.

Proof. To prove (i), we start with arguments (21) and (23) in the proof of Theorem 4.1, which imply

$$\Pr_{X,S}\{\hat{y}(X) \neq \hat{y}_*(X)\} \leq (2 + \alpha) \exp\left(-\min(\frac{1}{2} C', \gamma) \cdot n^{\frac{\beta}{\beta+C_\upsilon}}\right) \tag{24}$$

when $n \geq (\frac{2}{C'} \ln n)^{\frac{C_\upsilon+\beta}{\beta}}$. This constraint can be satisfied for all $n \geq N$, where $N > 0$ is a constant depending on $\lambda, \upsilon, d, \mathcal{V}, \beta$. In other words, when $n \geq N$, we have (24).

Set

$$C_N := (2+\alpha)\exp\left(-\min(\frac{1}{2}C', \gamma)\cdot N^{\frac{\beta}{\beta+C_v}}\right). \tag{25}$$

Then, when $n \leq N$, there is

$$\begin{aligned}\Pr_{X,S}\{\hat{y}(X) \neq \hat{y}_*(X)\} \leq 1 &= \frac{1}{C_N}(2+\alpha)\exp\left(-\min(\frac{1}{2}C', \gamma)\cdot N^{\frac{\beta}{\beta+C_v}}\right) \\ &\leq \frac{1}{C_N}(2+\alpha)\exp\left(-\min(\frac{1}{2}C', \gamma)\cdot n^{\frac{\beta}{\beta+C_v}}\right).\end{aligned} \tag{26}$$

Combining (24) and (26) with $\tilde{C}_\alpha := \max(1, C_N^{-1})\cdot(2+\alpha)$, we have

$$\Pr_{X,S}\{\hat{y}(X) \neq \hat{y}_*(X)\} \leq \tilde{C}_\alpha \exp\left(-\min(\frac{1}{2}C', \gamma)\cdot n^{\frac{\beta}{\beta+C_v}}\right). \tag{27}$$

This proves claim (i) in the remark.

To prove claim (ii), it is tempting to simply take the limit of β on both sides of (27). However, we need to first make sure $\lim_{\beta\to\infty}\tilde{C}_\alpha < \infty$ which is not obvious since $\tilde{C}_\alpha$ also depends on β.

Recall $\tilde{C}_\alpha = \max(1, C_N^{-1})\cdot(2+\alpha)$ where C_N is set in (25), it is clear a sufficient condition for $\lim_{\beta\to\infty}\tilde{C}_\alpha < \infty$ is $\lim_{\beta\to\infty} N^{\frac{\beta}{\beta+C_v}} < \infty$. This is true for the following reason.

First, as $\beta \to \infty$, N remains bounded because now we need every $n \geq N$ to satisfy

$$n \geq \lim_{\beta\to\infty}(\frac{2}{C'}\ln n)^{\frac{C_v+\beta}{\beta}} = \frac{2}{C'}\ln n, \tag{28}$$

which is easy to fulfill by a finite N. Let N_* be such an integer. Then, $\lim_{\beta\to\infty} N^{\frac{\beta}{\beta+C_v}} = N_*$.

Now we can safely take the limit of β on both sides of (27), which proves claim (ii). □

Corollary 4.3. *Suppose data distribution has (α, β, γ) Boltzmann margin and $k = O(n^{\frac{\beta+2}{\beta+C_v}})$. Then for any $\delta > 0$, the following holds for $n \geq 1$:*

$$er(\hat{y}) - er(\hat{y}_*) \leq \tilde{C}_\alpha \delta^{-1}\exp\left(-C_\gamma n^{\frac{\beta}{\beta+C_v}}\right),$$

with probability at least $1-\delta$ over the random choice of S.

Proof. Combining Lemma A.1 and argument (27) for Remark 4.2, for any $t > 0$ and $n \geq 1$,

$$\Pr_S\{er(\hat{y}) - er(\hat{y}_*) > t\} \leq \frac{1}{t}\tilde{C}_\alpha \exp(-C_\gamma n^{\frac{\beta}{\beta+C_v}}), \tag{29}$$

Setting the right side to δ and solving for t proves the corollary. □

Corollary 4.4. *Suppose data distribution has a (α', β') Tsybakov margin and let $\alpha, \beta, \gamma > 0$ be constants satisfying $\gamma \leq \ln(\alpha/\alpha')$. If $k = O(n^{\frac{2+\beta}{C_v+\beta}})$, then*

$$||\hat{y} - \hat{y}_*||_e \leq \tilde{C}_\alpha \exp(-C_\gamma n^{\frac{\beta}{\beta+C_v}}),$$

for $2 \leq n \leq \left(\gamma^{-1}\ln(\alpha/\alpha')\right)^{1+C_v/\beta}$.

Proof. By Lemma 3.2, the distribution satisfies the (α, β, γ) Boltzmann margin property when $t \in [(\gamma/\ln(\alpha/\alpha'))^{1/\beta}, 0.5]$.

Revisit the proof of Theorem 4.1 based on this implied Boltzmann margin. It sets $t = n^{\frac{-1}{C_v+\beta}}$, which implies

$$2 \leq n \leq \left(\frac{1}{\gamma}\ln(\alpha/\alpha')\right)^{1+C_v/\beta}. \tag{30}$$

This proves the corollary. □

Corollary 4.5. *Let $k = O(n)$. If data distribution has a t_* Massart margin, then*

$$||\hat{y} - \hat{y}_*||_e \leq C \exp(-C_* n), \tag{31}$$

for all $n \geq C_*^{-1} \ln n$, where C and $C_* \propto t_*^{C_\upsilon}$ are positive constants. Moreover, for every $t_* > 0$, there exists a data distribution satisfying the t_* Massart margin for which (31) holds for all $n \geq 1$ (possibly with different constants).

Proof. By Lemma 3.1, the distribution also has (α, β, γ) Boltzmann margin as long as $\alpha \exp(-\gamma t_*^{-\beta}) = 1$. Revisit the proof of Theorem 4.1 based on this implied Boltzmann margin.

Arguments (18) (20) imply that, if $k = O(t^{\frac{d}{\upsilon}} n C_d^{\frac{2\upsilon+2d}{2\upsilon+d}})$, then for any $t > 0$:

$$||\hat{y} - \hat{y}_*||_e \leq \Pr_{X,S}\{\hat{y}(X) \neq \hat{y}_*(X)\} \leq 2n \exp(-C' n t^{C_\upsilon}) + \Pr_X\{|1/2 - \eta(X)| \leq t\}. \tag{32}$$

Set $t = t_*$. Then, the right probability vanishes due to Massart margin and the above becomes: if $k = O(t_*^{\frac{d}{\upsilon}} n C_d^{\frac{2\upsilon+2d}{2\upsilon+d}})$, then

$$||\hat{y} - \hat{y}_*||_e \leq 2n \exp(-C' t_*^{C_\upsilon} n) \leq 2 \exp(-\frac{1}{2} C' t_*^{C_\upsilon} n), \tag{33}$$

where the second inequality holds when $n \leq \exp(\frac{1}{2} C' n t_*^{C_\upsilon})$. This proves the first claim of the corollary.

To prove the second claim, start with a data distribution that has $(1, \beta, t_*^\beta)$ Boltzmann margin for any given $t_* > 0$.

Set $\beta \to \infty$. Then two things happen: Lemma 3.1 suggests this data distribution gains a t_* Massart margin, and Remark 4.2 says $||\hat{y} - \hat{y}_*||_e \leq C'_\alpha \exp(-C_\gamma n)$ for $k = O(n)$. For the second event, we should remark that $\lim_{\beta \to \infty} C_\gamma < \infty$ because $C_\gamma = \min(\frac{1}{2} C', \gamma)$ and C' does not depend on margin parameters. Combining them proves the second claim. □

C RESULTS FOR ENSEMBLE KNN CLASSIFICATION

Theorem 5.1. *Suppose data distribution has a (α, β, γ) Boltzmann margin. If $k = O(\tilde{n}^{\frac{\beta+2}{\beta+C_\upsilon}})$, then*

$$||\hat{y}_e - \hat{y}_*||_e \leq C_\alpha \exp(-C_\gamma \tilde{n}^{\frac{\beta}{\beta+C_\upsilon}}), \tag{34}$$

for $\tilde{n} \geq (2 \max\{\ln \tilde{n}, 2 \ln m\}/C')^{1+C_\upsilon/\beta}$.

Moreover, if $m = O(\sqrt{\tilde{n}})$, then (34) holds for all $\tilde{n} \geq 1$, with C_α replaced by a possibly different constant $\tilde{C}_\alpha > 0$.

Proof. Recall $\hat{y}_e$ is trained on $\tilde{S}$, which is a collection of m samples $S_1, \ldots, S_m$ each containing $\tilde{n}$ i.i.d. sampled points.

By Lemma A.1 and Lemma A.2, for any $t > 0$,

$$\begin{aligned} ||\hat{y}_e - \hat{y}_*||_e &\leq \Pr_{X,\tilde{S}}\{\hat{y}_e(X) \neq \hat{y}_*(X)\} \\ &\leq \Pr_{X,\tilde{S}}\{|\hat{\eta}_e(X; \tilde{S}) - \eta(X)| > t\} + \Pr_X\{|1/2 - \eta(X)| \leq t\}. \end{aligned} \tag{35}$$

The second term can be bounded using Boltzmann margin. For the first term,

$$\begin{aligned} \Pr_{X,\tilde{S}}\{|\hat{\eta}_e(X; \tilde{S}) - \eta(X)| > t\} &= \Pr_{X,\tilde{S}}\left\{\frac{1}{m}\left|\sum_{j=1}^m \hat{\eta}(X; S_j) - \eta(X)\right| > t\right\} \\ &\leq \sum_{j=1}^m \Pr_{X,S_j}\{|\hat{\eta}(X; S_j) - \eta(X)| > t\} \\ &= m \cdot \Pr_{X,S_1}\{|\hat{\eta}(X; S_1) - \eta(X)| > t\}, \end{aligned} \tag{36}$$

where the inequality follows a union bound and the last equality is based on the i.i.d. sampling of $S_1, \ldots, S_m$.

The right side of (36) can be bounded using Corollary A.4. Recall $C_\upsilon = 2 + d/\upsilon$ and $C' = \frac{C_d}{\mathcal{V}C^{C_\upsilon}}$. The corollary says if $k = O(t^{\frac{d}{\upsilon}}\tilde{n}C_d^{\frac{2\upsilon+2d}{2\upsilon+d}})$ then $\Pr_{X,S_1}\{|\hat{\eta}(X;S_1) - \eta(X)| > t\} \leq 2\tilde{n}\exp\left(-C'\tilde{n}t^{C_\upsilon}\right)$. Plugging this back to (36), we have

$$\Pr_{X,\tilde{S}}\{|\hat{\eta}_e(X;\tilde{S}) - \eta(X)| > t\} \leq 2m\tilde{n}\exp\left(-C'\tilde{n}t^{C_\upsilon}\right) \leq 2m\exp\left(-\frac{1}{2}C'\tilde{n}t^{C_\upsilon}\right) \leq 2\exp\left(-\frac{1}{4}C'\tilde{n}t^{C_\upsilon}\right), \tag{37}$$

where the second inequality holds when $\tilde{n} \leq \exp(\frac{1}{2}C'\tilde{n}t^{C_\upsilon})$ and the third holds when $m \leq \exp\left(\frac{1}{4}C'\tilde{n}t^{C_\upsilon}\right)$.

Plugging the above back to (35), we have

$$\Pr_{X,\tilde{S}}\{\hat{y}_e(X) \neq \hat{y}_*(X)\} \leq 2\exp\left(-\frac{1}{4}C'\tilde{n}t^{C_\upsilon}\right) + \alpha\exp(-\gamma t^{-\beta}). \tag{38}$$

To merge terms, set $\tilde{n}t^{C_\upsilon} = t^{-\beta}$. Then $t = \tilde{n}^{-\frac{1}{\beta+C_\upsilon}}$ and the above implies

$$\Pr_{X,\tilde{S}}\{\hat{y}_e(X) \neq \hat{y}_*(X)\} \leq (2+\alpha)\exp\left(-\min\left(\frac{1}{4}C', \gamma\right)\cdot\tilde{n}^{\frac{\beta}{\beta+C_\upsilon}}\right). \tag{39}$$

Based on this choice of t, the above constraints on $\tilde{n}$ and n can be merged into

$$\tilde{n} \geq \left(\frac{2}{C'}\max\{\ln\tilde{n}, 2\ln m\}\right)^{1+C_\upsilon/\beta}. \tag{40}$$

This proves the first claim of the theorem.

The second claim follows from the fact that, if $m = O(\sqrt{\tilde{n}})$, then (40) becomes $\tilde{n} \geq (C''\ln\tilde{n})^{1+C_\upsilon\beta}$ for some constant $C'' > 0$. It is clear this new constraint can be satisfied for large enough $\tilde{n}$. This means there exists an integer $\tilde{N} > 0$ such that $n \geq \tilde{N}$ implies $\tilde{n} \geq (C''\ln\tilde{n})^{1+C_\upsilon\beta}$ and thus (39). Set

$$C_{\tilde{N}} = (2+\alpha)\exp\left(-\min\left(\frac{1}{4}C', \gamma\right)\cdot\tilde{N}^{\frac{\beta}{\beta+C_\upsilon}}\right). \tag{41}$$

Then, when $n \leq \tilde{N}$, we have

$$\begin{aligned}\Pr_{X,\tilde{S}}\{\hat{y}_e(X) \neq \hat{y}_*(X)\} \leq 1 &= \frac{1}{C_{\tilde{N}}}(2+\alpha)\exp\left(-\min\left(\frac{1}{4}C', \gamma\right)\cdot\tilde{N}^{\frac{\beta}{\beta+C_\upsilon}}\right) \\ &\leq \frac{1}{C_{\tilde{N}}}(2+\alpha)\exp\left(-\min\left(\frac{1}{4}C', \gamma\right)\cdot\tilde{n}^{\frac{\beta}{\beta+C_\upsilon}}\right).\end{aligned} \tag{42}$$

Finally, combining (39) and (42) with $\tilde{C}_\alpha := \max\{2+\alpha, \frac{1}{C_{\tilde{N}}}(2+\alpha)\}$ proves the second claim. □

Remark C.1. *Theorem 5.1 implies that, if data distribution has (α, β, γ) Boltzmann margin and $k = O(\tilde{n}^{\frac{\beta+2}{\beta+C_\upsilon}})$, then*

$$\Pr_{X,\tilde{S}}\{\hat{y}_e(X) \neq \hat{y}_*(X)\} \leq \tilde{C}_\alpha\exp(-C_\gamma\tilde{n}^{\frac{\beta}{\beta+C_\upsilon}}) \tag{43}$$

for either (i) $m = O(\sqrt{\tilde{n}})$ and $\tilde{n} \geq 1$, or (ii) $m = \tilde{O}(e^{\tilde{n}})$ and $\tilde{n} \geq (2\max\{\ln\tilde{n}, 2\ln m\}/C')^{1+C_\upsilon/\beta}$.

Remark 5.2. *Theorem 5.1 also applies when (i) a training set S is sampled i.i.d. from the population, (ii) subsets $S_1, \ldots, S_m$ are drawn i.i.d. from S, and (iii) in each subset, points are sampled without replacement.*

Proof. Consider the alternative setting. Let S be a training set, whose elements are sampled i.i.d. from the population. Let $S_1, \ldots, S_m$ subsets of S that are independently sampled, and points in each subset are sampled without replacement. Let $\tilde{S}$ be the collection of these subsets. Let $\hat{y}_b$ be the ensemble kNN classifier built on above $\tilde{S}$. Its expected excess error is

$$||\hat{y}_b - \hat{y}_*||_e = \mathbb{E}_{\tilde{S},S}[er(\hat{y}_b)] - er(\hat{y}_*). \tag{44}$$

We will show this error has the same bound presented in Theorem 5.1.

First, by the arguments in Lemma A.1 proof, we have

$$er(\hat{y}) - er(\hat{y}_*) \leq \Pr_X\{\hat{y}(X) \neq \hat{y}_*(X)\}. \tag{45}$$

Taking expectation on both sides of the above w.r.t. S and $\tilde{S}$, we have

$$||\hat{y}_b - \hat{y}_*||_e \leq \Pr_{X,\tilde{S},S}\{\hat{y}_b(X) \neq \hat{y}_*(X)\} = \mathbb{E}_S \Pr_{X,\tilde{S}|S}\{\hat{y}_b(X) \neq \hat{y}_*(X)\}. \tag{46}$$

Second, by the arguments in Lemma A.2 proof, we have

$$\mathbb{E}_S \Pr_{X,\tilde{S}|S}\{\hat{y}_b(X) \neq \hat{y}_*(X)\} \leq \mathbb{E}_S \Pr_{X,\tilde{S}|S}\{|\hat{\eta}_b(X;\tilde{S}) - \eta(X)| > t\} + \Pr_X\{|1/2 - \eta(X)| \leq t\}. \tag{47}$$

Next, by following argument (36) in Theorem 5.1 proof, we have

$$\begin{aligned}
\mathbb{E}_S \Pr_{X,\tilde{S}|S}\{|\hat{\eta}_b(X;\tilde{S}) - \eta(X)| > t\} &= \mathbb{E}_S \Pr_{X,\tilde{S}|S}\left\{\frac{1}{m}\left|\sum_{j=1}^{m}\hat{\eta}(X;S_j) - \eta(X)\right| > t\right\} \\
&\leq \mathbb{E}_S \sum_{j=1}^{m} \Pr_{X,S_j|S}\{|\hat{\eta}(X;S_j) - \eta(X)| > t\} \\
&= \sum_{j=1}^{m} \mathbb{E}_S \Pr_{X,S_j|S}\{|\hat{\eta}(X;S_j) - \eta(X)| > t\} \\
&= \sum_{j=1}^{m} \Pr_{X,S_j}\{|\hat{\eta}(X;S_j) - \eta(X)| > t\} \\
&= m \Pr_{X,S_1}\{|\hat{\eta}(X;S_1) - \eta(X)| > t\}.
\end{aligned} \tag{48}$$

In above, $\Pr_{X,S_j|S}$ only treats S_j as a random subset given S, while $\Pr_{X,S_j}$ treats S_j as an i.i.d. sample of the population (because its elements are sampled from S without replacement). The last equality holds because S_j's are sampled i.i.d. from S and thus now treated as i.i.d. samples of the population. Plugging the above back to (47) then (49), we have

$$||\hat{y}_b - \hat{y}_*||_e \leq m \Pr_{X,S_1}\{|\hat{\eta}(X;S_1) - \eta(X)| > t\} + \Pr_X\{|1/2 - \eta(X)| \leq t\}. \tag{49}$$

Comparing this bound with the bound (35 + 36) presented in Theorem 5.1 proof i.e.,

$$||\hat{y}_e - \hat{y}_*||_e \leq m \cdot \Pr_{X,S_1}\{|\hat{\eta}(X;S_1) - \eta(X)| > t\} + \Pr_X\{|1/2 - \eta(X)| \leq t\}, \tag{50}$$

we see $\hat{y}_b$ has the same excess error bound as $\hat{y}_e$. It is also noted the rest analysis on $\Pr_{X,S_1}\{|\hat{\eta}(X;S_1) - \eta(X)| > t\}$ is the same for both classifiers, since S_1 is treated as an i.i.d. sample from the population in both cases. □

Corollary 5.3. *Suppose data distribution has (α, β, γ) Boltzmann margin. If $k = O(\tilde{n}^{\frac{\beta+2}{\beta+C_v}})$ and $m = O(\sqrt{\tilde{n}})$, then for any $\delta > 0$ and $\tilde{n} \geq 1$, there is*

$$er(\hat{y}) - er(\hat{y}_*) \leq \tilde{C}_\alpha \delta^{-1} \exp(-C_\gamma \tilde{n}^{\frac{\beta}{\beta+C_v}}),$$

with probability at least $1 - \delta$ over the random choice of $\tilde{S}$, where $\tilde{C}_\alpha$ is specified in Theorem 5.1.

Proof. By Lemma A.1 and Remark C.1, we have for any $t > 0$:

$$\Pr_{\tilde{S}}\{er(\hat{y}_e) - er(\hat{y}_*) > t\} \leq \frac{1}{t}\Pr_{X,\tilde{S}}\{\hat{y}_e(X) \neq \hat{y}_*(X)\} \leq \frac{1}{t}\tilde{C}_\alpha \exp(-C_\gamma \tilde{n}^{\frac{\beta}{\beta+C_v}}). \tag{51}$$

Setting the right side to δ and solving for t proves the corollary. □

Theorem 5.4. *Suppose data distribution has a (α, β, γ) Boltzmann margin. Then, $\hat{y}_e$ is strongly Bayes consistent if $k = O(\tilde{n}^{\frac{\beta+2}{\beta+C_v}})$ and $m = \tilde{O}(e^{\tilde{n}})$.*

Proof. By Lemma A.1, for any $\tilde{S}$,

$$er(\hat{y}_e) - er(\hat{y}_*) \leq \Pr_X\{\hat{y}_e(X) \neq \hat{y}_*(X)\} := Q(\tilde{S}). \tag{52}$$

We will prove strong consistency based on the following definition i.e., for any $t > 0$,

$$\Pr_{\tilde{S}}\{\lim_{|\tilde{S}|\to\infty} Q(\tilde{S}) > t\} = 0. \tag{53}$$

According to the Borel-Cantelli Lemma A.5, it suffices to show

$$\sum_{|\tilde{S}|=N_*}^{\infty} \Pr_{\tilde{S}}\{Q(\tilde{S}) > t\} < \infty, \tag{54}$$

for some integer N_*.

Recall $|\tilde{S}| = m\tilde{n}$. Our observation is that, in order to verify (54), it is sufficient to verify the following with $m = 1$:

$$\sum_{\tilde{n}=N_*}^{\infty} \Pr_{\tilde{S}}\{Q(\tilde{S}) > t\} < \infty, \tag{55}$$

The reason is, the sequence of $|\tilde{S}|$ generated with divergent m and divergent $\tilde{n}$ is a subsequence of $|\tilde{S}|$ generated with $m = 1$ and divergent $\tilde{n}$. Therefore, if the latter sequence converges, the former sequence also converges.

Fix $m = 1$. We will set N_* in (55) from two perspectives.

Let $t_s > 0$ be a variable depending on $\tilde{S}$, and inspect each $\Pr_{\tilde{S}}\{Q(\tilde{S}) > t_s\}$. Remark C.1 implies that for $k = O(\tilde{n}^{\frac{\beta+2}{\beta+C_v}})$, $m = \tilde{O}(e^{\tilde{n}})$ (which includes $m = 1$) and large enough $\tilde{n}$ that satisfies $\tilde{n} \geq (2\ln\tilde{n}/C')^{1+C_v/\beta}$ (which sets N_*):

$$\Pr_{X,\tilde{S}}\{\hat{y}_e(X) \neq \hat{y}_*(X)\} \leq \tilde{C}_\alpha \exp(-C_\gamma \tilde{n}^{\frac{\beta}{\beta+C_v}}). \tag{56}$$

Then, by Lemma A.1, we have

$$\Pr_{\tilde{S}}\{Q(\tilde{S}) > t_s\} \leq \frac{1}{t_s}\mathbb{E}_{\tilde{S}} Q(\tilde{S}) = \frac{1}{t_s}\Pr_{X,\tilde{S}}\{\hat{y}_e(X) \neq \hat{y}_*(X)\} \leq \frac{\tilde{C}_\alpha}{t_s}\exp(-C_\gamma \tilde{n}^{\frac{\beta}{\beta+C_v}}). \tag{57}$$

Set $t_s = \exp(-\frac{C_\gamma}{2}\tilde{n}^{\frac{\beta}{\beta+C_v}})$. Then $t_s \to 0$ as $\tilde{n} \to \infty$. This means for any $t > 0$, when $\tilde{n}$ is large enough (which sets N_* from another perspective), there is $t_s < t$ and hence

$$\Pr_{\tilde{S}}\{Q(\tilde{S}) > t\} \leq \Pr_{\tilde{S}}\{Q(\tilde{S}) > t_s\} \leq \frac{\tilde{C}_\alpha}{t_s}\exp(-C_\gamma \tilde{n}^{\frac{\beta}{\beta+C_v}}) = \tilde{C}_\alpha \exp(-\frac{1}{2}C_\gamma \tilde{n}^{\frac{\beta}{\beta+C_v}}). \tag{58}$$

Plugging the above back to (55), we have that for any $t > 0$,

$$\sum_{\tilde{n}=N_*}^{\infty} \Pr_{\tilde{S}}\{Q(\tilde{S}) > t\} \leq \sum_{\tilde{n}=N_*}^{\infty} \Pr_{\tilde{S}}\{Q(\tilde{S}) > t_s\} \leq \tilde{C}_\alpha \sum_{\tilde{n}=N_*}^{\infty} \exp(-\frac{1}{2}C_\gamma \tilde{n}^{\frac{\beta}{\beta+C_v}}) < \infty. \tag{59}$$

This verifies (55) and further (54) with any divergent m. Finally, note we still need $m = \tilde{O}(e^{\tilde{n}})$ in order to apply Remark C.1 and obtain (56). This proves the theorem. □

D NUMERICAL RESULTS

D.1 DENSITY DESIGN

Recall $\mathcal{X} = [0, 1]$ and $\eta(X) = X$. Given constants $\alpha, \beta, \gamma > 0$, our goal is to design a density function f_b on $[0, 1]$ that satisfies (α, β, γ) Boltzmann margin. The basic idea is threefold: data decay near decision boundary, data are constant elsewhere (for better visualization and does not affect Boltzmann margin), and density is Hölder continuous.

Pick a threshold $T \in (0, 0.5)$ and consider two cases.

When $|X - 0.5| \leq T$, we ask data to decay as if under Boltzmann margin. This implies

$$\Pr\{|\eta(x) - 0.5| \leq t\} = \Pr\{|x - 0.5| \leq t\} = 2\int_{0.5}^{0.5+t} f_b(x)dx \leq \alpha \exp(-\gamma t^{-\beta}). \tag{60}$$

For the last inequality, taking derivative of t on both sides and setting $t' = 0.5 + t$ yields

$$f_b(t') \leq \frac{1}{2}\alpha\beta\gamma e^{-\gamma(t'-0.5)^{-\beta}}(t' - 0.5)^{-\beta-1}. \tag{61}$$

A mirrored result of (60) can be obtained by taking integral on $[0.5 - t, 0.5]$ instead of $[0.5, 0.5 + t]$, which yields

$$f_b(t') \leq \frac{1}{2}\alpha\beta\gamma e^{-\gamma(0.5-t')^{-\beta}}(0.5 - t')^{-\beta-1}. \tag{62}$$

Combining both results, we can set

$$f_b(x) = \frac{1}{N_T} \cdot \frac{1}{2}\alpha\beta\gamma e^{-\gamma|x-0.5|^{-\beta}}|x - 0.5|^{-\beta-1}, \tag{63}$$

where N_T is a normalizer (to be specified later). This completes the density design for $|X - 0.5| \leq T$.

When $|x - 0.5| \geq T$, we ask density to stay constant i.e., $f_b(x) = \frac{C_T}{N_T}$ for some constant C_T.

It remains to specify N_T and C_T. Since $f_b(x)$ is continuous at $|0.5 - X| = T$, based on (63) we have

$$C_T = \frac{1}{2}\alpha\beta\gamma e^{-\gamma T^{-\beta}} T^{-\beta-1}. \tag{64}$$

Further, $\int_0^1 f_b(x) = 1$ implies that (through straight calculations)

$$N_T = \alpha \exp(-\gamma T^{-\beta}) + 2C_T(0.5 - T).$$

Putting all together, we complete the density design on $\mathcal{X}$, i.e.,

$$f_b(x) = \begin{cases} \frac{\alpha\beta\gamma}{2N_T} \cdot \frac{e^{-\frac{\gamma}{|x-0.5|^{\beta}}}}{|x-0.5|^{\beta+1}}, & |x - 0.5| \leq T \\ \frac{\alpha\beta\gamma}{2N_T} \cdot \frac{e^{-\gamma T^{-\beta}}}{T^{\beta+1}}, & |x - 0.5| > T. \end{cases} \tag{65}$$

D.2 DATA SAMPLING

Our goal is to sample 1000 points in $\mathcal{X}$ according to the Boltzmann-margin density 5, by applying the rejection sampling Bishop and Nasrabadi [2006] technique.

To be specific, recall $\mathcal{X} = [0, 1]$. Let $g(x) = 1$ be the density of a uniform distribution on $\mathcal{X}$, and $M > 0$ be a constant such that $f_b(x) \leq Mg(x) = M$ for all $x \in \mathcal{X}$. We estimate M from $f_b(x)$ based on 1000 points uniformly sampled in $\mathcal{X}$. Detail sampling and estimation process is described in Algorithm 1.

Algorithm 1 Data Generation

Input: Data pool $\mathcal{X} = \emptyset$, pool size N_p, parameter β.
1: Sample a point $x \in [0, 1]$ uniformly at random.
2: Sample a point $u \in [0, M]$ uniformly at random.
3: If $u \leq f_\beta(x)$, add x to $\mathcal{X}$ (allow duplicate points).
4: Repeat (1-3) until $|\mathcal{X}| \geq N_p$.
5: **for** $x \in \mathcal{X}$ **do**
6: Sample a point $v \in [0, 1]$ uniformly at random.
7: Label x by 1 if $v \leq \eta(x)$; label it by 0 otherwise.
8: **end for**