4-2014

An Analysis of Active Learning With Uniform Feature Noise

Aaditya Ramdas
Carnegie Mellon University

Barnabas Poczos
Carnegie Mellon University, bapoczos@cs.cmu.edu

Aarti Singh
Carnegie Mellon University, aarti@cs.cmu.edu

Larry Wasserman
Carnegie Mellon University, larry@stat.cmu.edu

Published In

Journal of Machine Learning Research : Workshop and Conference Proceedings, 33, 805-813.

An Analysis of Active Learning With Uniform Feature Noise

Aaditya Ramdas **Barnabas Poczos** **Aarti Singh** **Larry Wasserman**
Carnegie Mellon University, Pittsburgh, USA

Abstract

In active learning, the user sequentially chooses values for feature X and an oracle returns the corresponding label Y. In this paper, we consider the effect of feature noise in active learning, which could arise either because X itself is being measured, or it is corrupted in transmission to the oracle, or the oracle returns the label of a noisy version of the query point. In statistics, feature noise is known as "errors in variables" and has been studied extensively in non-active settings. However, the effect of feature noise in active learning has not been studied before. We consider the well-known Berkson errors-in-variables model with additive uniform noise of width σ.

Our simple but revealing setting is that of one-dimensional binary classification setting where the goal is to learn a threshold (point where the probability of a + label crosses half). We deal with regression functions that are antisymmetric in a region of size σ around the threshold and also satisfy Tsybakov's margin condition around the threshold. We prove minimax lower and upper bounds which demonstrate that when σ is smaller than the minimiax active/passive noiseless error derived in Castro & Nowak (2007), then noise has no effect on the rates and one achieves the same noiseless rates. For larger σ, the *unflattening* of the regression function on convolution with uniform noise, along with its local antisymmetry around the threshold, together yield a behaviour where noise *appears* to be beneficial. Our key result is that active learning can buy significant improvement over a passive strategy even in the presence of feature noise.

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33.

1 Introduction

Active learning is a machine learning paradigm where the algorithm interacts with a label-providing oracle in a feedback driven loop where past training data (features queried and corresponding labels) are used to guide the design of subsequent queries. Typically, the oracle is queried with an exact feature value and the oracle returns the label corresponding precisely to that feature value. However, in many scenarios, the feature value being queried can be noisy and it helps to analyze what would happen in such a setting. Such situations include noisy sensor measurements of features, corrupted transmission of data from source to storage, or just access to a limited noisy oracle.

The errors-in-variables model has been well studied in the statistical literature and their effect can be profound. In density estimation, Gaussian error causes the minimax rate to become logarithmic in sample size instead of polynomial, see Fan (1991). For results in passive regression, refer to Fan et al. (1993); Fuller (2009); Carroll et al. (2010), and for passive classification, see Loustau & Marteau (2012). However, classification has not been studied in the *Berkson* model introduced below. Also, deconvolution estimators require the noise fourier transform to be bounded away from zero, ruling out uniform noise. Finally, to the best of our knowledge, feature noise has not been studied for active learning in any setting.

The *classical errors in variables model* has the graphical form $W \leftarrow X \rightarrow Y$, representing

$$W = X + \delta \ ,$$
$$Y = m(X) + \epsilon \ .$$

Here, the label Y depends on the feature X but we do not observe X; rather we observe the noisy feature W. The *Berkson errors in variables model* is

$$X = W + \delta \ ,$$
$$Y = m(X) + \epsilon \ .$$

The difference is that we start with an observed feature W and then noise is added to determine X. Graphically, this model is $W \rightarrow X \rightarrow Y$.

In this paper, we focus on the Berkson error model since it intuitively makes more sense for active learning - it captures the idea that we request a label for feature W, but the oracle returns the label for X which is a corrupted version generated from W, i.e. the noise occurs between the label request and the oracle output. We use uniform noise since it yields insightful behavior and also has not been addressed in the literature. We conjecture that qualitatively similar results hold for other symmetric error models.

1.1 The Setup

Threshold Classification. Let $\mathcal{X} = [-1, 1]$, $\mathcal{Y} = \{+, -\}$, and $f : \mathcal{X} \to \mathcal{Y}$ denote a classification rule. Assuming 0/1 loss, the risk of the classification rule f is $R(f) = \mathbb{E}[1_{\{f(X)\neq Y\}}] = \mathbb{P}(f(X) \neq Y)$. It is known that the Bayes optimal classifier, the best measurable classifier that minimizes the risk $f^* = \arg\min_f R(f)$, has the following form

$$f^*(x) = \begin{cases} + & \text{if } m(x) \geq 1/2 \ , \\ - & \text{if } m(x) < 1/2 \ , \end{cases}$$

where $m(x) = \mathbb{P}(Y = +|X = x)$ is the unknown regression function. In what follows, we will consider the case where the f^* is a threshold classifier, i.e. there exists a unique $t \in [-1, 1]$ with $m(t) = 1/2$ such that $m(x) < 1/2$ if $x < t$, and $m(x) > 1/2$ if $x > t$.

Berkson Error Model. The model is:

1. User chooses W and requests label.
2. Oracle receives a noisy W namely $X = W + U$.
3. Oracle returns Y where $\mathbb{P}(Y = +|X = x) = m(x)$.

We take the noise to be uniform: $U \sim \text{Unif}[-\sigma, \sigma]$, where the noise width σ is known for simplicity.

Sampling Strategies. In *passive sampling*, assume that we are given a batch of $w_i \sim \text{Unif}[-1, 1]$ and corresponding labels y_i sampled independently of $\{w_j\}_{j\neq i}$ and $\{y_j\}_{j\neq i}$. In this case, a strategy S is just an estimator $S_n : (W \times Y)^n \to [-1, 1]$ that returns a guess $\widehat{t}$ of the threshold t on seeing $\{w_i, y_i\}_{i=1}^n$.

In *active sampling* we are allowed to sequentially choose $w_i = S_i(w_1, \ldots, w_{i-1}, y_1, \ldots, y_{i-1})$, where S_i is a possibly random function of past queries and labels, where the randomness is independent of queries and labels. In this case, a strategy A is a sequence of functions $S_i : (W \times Y)^{i-1} \to [-1, 1]$ returning query points and an estimator $S_n : (W \times Y)^n \to [-1, 1]$ that returns a guess $\widehat{t}$ at the end.

Let $\mathcal{S}_n^P, \mathcal{S}_n^A$ be the set of all passive or active strategies (and estimators) with a total budget of n labels.

To avoid the issue of noise resulting in a point outside the domain, we make a (Q)uerying assumption:

(Q). Querying within σ of the boundary is disallowed.

Loss Measure. Let $\widehat{t} = \widehat{t}(W_1^n, Y_1^n)$ denote an estimator of t using n samples from a passive or active strategy. Our task will be to estimate the location of t, where we measure accuracy of an estimator $\widehat{t}$ by a loss function which is the point error $|\widehat{t} - t|$.

Function Class. In the analysis of rates for classification (among others), it is common to use the *Tsybakov Noise/Margin Condition* (see Tsybakov (2004)), to characterize the behavior of $m(x)$ around the threshold t. Given constants c, C with $C \geq c$, $k \geq 1$, and noise level σ, let $\mathcal{P}(c, C, k, \sigma)$ be the set of regression functions $m(x)$ that satisfy the following conditions (T,M,B) for some threshold t:

(T). $|x - t|^{k-1} \geq |m(x) - 1/2| \geq c|x - t|^{k-1}$ whenever $|m(x) - 1/2| \leq \epsilon_0$ for some constant ϵ_0

(M). $m(t + \delta) - 1/2 = 1/2 - m(t - \delta)$ for all $\delta \leq \sigma$.

(B). t is at least σ away from the boundary.

On adding noise U, the point where $m \star U$ ($\star$ means convolution) crosses half may differ from t, the point where m crosses half. However, the antisymmetry assumption (M) and boundary assumption (B) together imply that the two thresholds are the same. Getting rid of (M,B) seems substantially difficult.

When $\sigma = 0$, (Q), (M) and (B) are vacuously satisfied, and this is exactly the class of functions and strategies considered in Castro & Nowak (2007). Smaller k means that the regression function is steeper, which makes it easier to estimate the threshold and classify future labels (cf. Steinwart & Scovel (2004)). $k = 1$ captures a discontinuous $m(x)$ jumping at t.

Minimax Risk. We are interested in the minimax risk under the point error loss :

$$\mathcal{R}_n(\mathcal{P}(c, C, k, \sigma)) = \inf_{S\in\mathcal{S}_n} \sup_{P\in\mathcal{P}(c,C,k,\sigma)} \mathbb{E}|\widehat{t} - t| \qquad (1)$$

where $\mathcal{S}_n$ is the set of strategies accessing n samples. For brevity, $\mathcal{R}_n^P(k, \sigma)$ or $\mathcal{R}_n^A(k, \sigma)$ denotes risk for (P)assive/(A)ctive sampling stratgies $\mathcal{S}_n^P, \mathcal{S}_n^A$.

Notation $\prec, \succ, \asymp, \preceq, \succeq$. We analyse minimax point error rates in different regimes of σ as a function of n (or equivalently, for a given point error, we can analyse how the sample size n depends on σ) and we write σ_n for emphasis. In this paper, $f_n \prec g_n$ means $f_n/g_n \to 0$, $f_n \asymp g_n$ means $c_1 g_n \leq f_n \leq c_2 g_n$ where c_1, c_2 are constants, $f_n \preceq g_n$ means $f_n \prec g_n$ or $f_n \asymp g_n$, $f_n \succeq g_n$ means $g_n \preceq f_n$ and $f_n \succ g_n$ means $g_n \prec f_n$.

2 Main Result and Comparisions

The main result of this paper is as follows.

Theorem 1. *Under the Berkson error model, when given n labels sampled actively or passively with assumption (Q), and when the true underlying regression function lies in $\mathcal{P}(c, C, k, \sigma_n)$ for known k, σ_n, the minimax risk under the point error loss is:*

1. $\mathcal{R}_n^P(\mathcal{P}(k,\sigma)) \asymp \begin{cases} n^{-\frac{1}{2k-1}} & \text{if } \sigma_n \prec n^{-\frac{1}{2k-1}} \\ \sigma_n^{-(k-\frac{3}{2})}\sqrt{\frac{1}{n}} & \text{otherwise} \end{cases}$

2. $\mathcal{R}_n^A(\mathcal{P}(k,\sigma)) \asymp \begin{cases} n^{-\frac{1}{2k-2}} & \text{if } \sigma_n \prec n^{-\frac{1}{2k-2}} \\ \sigma_n^{-(k-2)}\sqrt{\frac{1}{n}} & \text{otherwise} \end{cases}$

When $k = 1$, $m(x)$ jumps at the threshold, and we interpret the quantity $n^{-\frac{1}{2k-2}}$ as being exponentially small, i.e. being smaller than n^{-p} for any p. We also suppress logarithmic factors in n, σ_n. If the domain was $[-R, R]$, the corresponding passive rates are obtained by substituting n by n/R, but active rates remain the same upto logarithmic factors in R.

Remark. In this paper, we focus on learning the threshold t. This is relevant because the threshold maybe of intrinsic interest, and also of interest for prediction if, for example, future queries could be made with a different noise model or can be obtained (with some cost) noise-free. Similar results can be derived for 0/1-risk.

Zero Noise. When $\sigma = 0$, the assumptions (Q,B,M) are vacuously true, and our class $\mathcal{P}(c, C, k, 0)$ matches the class $\mathcal{P}(c, C, k)$ considered in Castro & Nowak (2007), and our rates for $\sigma = 0$ i.e. $n^{-\frac{1}{2k-1}}$ and $n^{-\frac{1}{2k-2}}$ are precisely the passive and active minimax point error rates in Castro & Nowak (2007).

Small Noise. When the noise is small, we get what we expect - the risk does not change with noise as long as the noise itself is smaller than the noiseless error. In other words, as long as the noise is smaller than the noiseless error rate of $n^{-\frac{1}{2k-1}}$ for passive learning, passive learners will not really be able to notice this tiny noise, and the minimax rate remains $n^{-\frac{1}{2k-1}}$. Similarly, as long as the noise is smaller than the noiseless error rate of $n^{-\frac{1}{2k-2}}$ for active learning, active learners will not really be able to notice this tiny noise, and the minimax rate remains $n^{-\frac{1}{2k-1}}$. Also, the passive rates vary smoothly - at the point when $\sigma_n \asymp n^{-\frac{1}{2k-1}}$, the rates for small and large noise coincide. Similarly, at the point when $\sigma_n \asymp n^{-\frac{1}{2k-2}}$, the aforementioned active rates for small and large noise coincide.

Large Noise and Assumption (M). When the noise is large, we see a curious behaviour of the rates. When $k > 2$, the error rates seem to get smaller/better with larger noise for both active and passive learning, and furthermore the noisy rates can also be better than the noiseless rate! This might seem to violate both the information processing inequality, and our intuition that more noise shouldn't help estimation. Moreover, a noiseless active learner may be able to simulate a noisy situation by adding noise and querying at the resulting point, and get better rates, violating lower bounds in Castro & Nowak (2007).

However, we make the following crucial but subtle observation. Our claimed rates are *not* about a fixed function class - due to assumption (M), the function class changes with σ, and in fact (M) requires the antisymmetry of the regression function to hold over a larger region for larger σ. This set of functions is actually getting smaller with larger σ. Even though the functions can behave quite arbitrarily outside $(t-\sigma, t+\sigma)$, this assumption (M) on a small region of size 2σ actually helps us significantly.

Given that there is no contradiction to the results of Castro & Nowak (2007) or more fundamental information theoretic ideas, there is also an intuitive explanation of why assumption (M) helps when we have large noise. As we will see in a later figure, convolution with noise seems to "stretch/unflatten" the function around the threshold. Specifically, for larger $k > 2$, the regression function can be quite flat around the threshold - convolution with noise makes it less flat and more linear - in fact it behaves linearly over a large region of width nearly 2σ. This is true regardless of whether assumption (M) holds - however if (M) does not hold, then the convolved threshold, which is the point where the convolved function crosses half, need not be the original threshold t. While dropping assumption (M) will not hurt if we only want to find the convolved threshold, but given that our aim is to estimate t, the problem of figuring out how much the threshold shifted can be quite non-trivial.

Hence, large noise ensures a behaviour that is less flat and more linear around the threshold, and assumption (M) ensures that the threshold doesn't shift from t. Intuitively this is why (M) and large noise help, and technically there is no contradiction becasue the function class is getting progressively simpler because of more controlled growth around the threshold.

The main takeaway is that in all settings, active learning yields a gain over passive sampling. We now describe the upper and lower bounds that lead to Theorem 1. The case $k = 1$ is handled in detail for intuitionb but proofs for $k > 1$ are in the Appendix.

2.1 Simulation of Noise Convolution

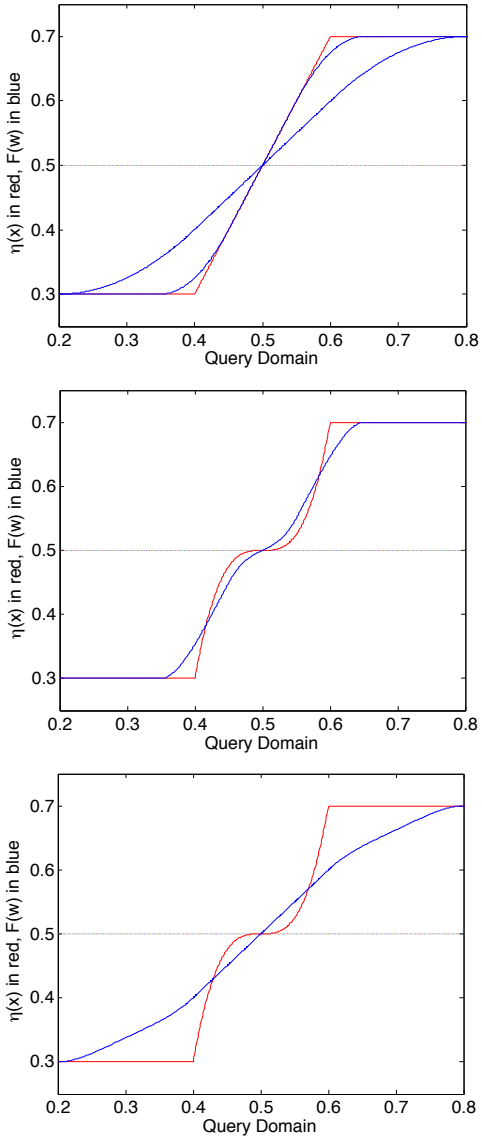

Figure 1: Regression function $\eta(x)$ (red) before and $F(w)$ (blue) after convolution with noise. In all 3 figures, Tsybakov's margin condition holds for $x \in [0.4, 0.6]$. The top plot has a linear regression function ($k = 2$), and its two blue curves are for $\sigma_n = 0.05$ (*narrow*), 0.2 (*wide*), and they show that a linear growth around $t = 0.5$ remains linear. The middle and bottom figure are for a flatter regression function with $k = 4$, and $\sigma_n = 0.05, 0.2$ respectively, plotted separately for clarity. $k = 4$ is harder than for $k = 2$ because the red curve is flatter around t, making it harder to pinpoint the threshold. However, as one can see in both plots, noise actually *helps* by smoothing it out and making it more linear. However, note that the effect of assumption (M) cannot be understated, due to which in all plots the threshold before and after noise cross half at the same point. The effect of noise when $k = 1$ can be seen in the following section.

2.2 Paper Roadmap

We devote the next two sections to proving the lower and upper bounds, in that order, that lead to Theorem 1. While the proofs will be self-contained, we leave some detailed calculations to the appendix.

For easier readibility, we present lower bounds for $k = 1$ first to absorb the technique and then the lower bounds for $k > 1$. In Section 2 we will prove

Theorem 2 (Lower Bounds)**.** *Under the Berkson error model and assumption (Q),*

1. *For $k = 1$, the passive/active lower bounds are*

$$\inf_{S \in \mathcal{S}_n^P} \sup_{P \in \mathcal{P}(1,\sigma_n)} \mathbb{E}|\widehat{t} - t| \;\succeq\; \begin{cases} \frac{1}{n} & \text{if } \sigma_n \prec \frac{1}{n} \\ \sqrt{\frac{\sigma_n}{n}} & \text{otherwise} \end{cases}$$

$$\inf_{S \in \mathcal{S}_n^A} \sup_{P \in \mathcal{P}(1,\sigma_n)} \mathbb{E}|\widehat{t} - t| \;\succeq\; \begin{cases} e^{-n} & \text{if } \sigma_n \prec e^{-n} \\ \frac{\sigma_n}{\sqrt{n}} & \text{otherwise} \end{cases}$$

2. *For $k > 1$, the passive/active lower bounds are*

$$\inf_{S \in \mathcal{S}_n^P} \sup_{P \in \mathcal{P}(k,\sigma_n)} \mathbb{E}|\widehat{t} - t| \;\succeq\; \begin{cases} n^{-\frac{1}{2k-1}} & \text{if } \sigma_n \prec n^{-\frac{1}{2k-1}} \\ \sigma_n^{-(k-\frac{3}{2})}\sqrt{\frac{1}{n}} & \text{otherwise} \end{cases}$$

$$\inf_{S \in \mathcal{S}_n^A} \sup_{P \in \mathcal{P}(k,\sigma_n)} \mathbb{E}|\widehat{t} - t| \;\succeq\; \begin{cases} n^{-\frac{1}{2k-2}} & \text{if } \sigma_n \prec n^{-\frac{1}{2k-2}} \\ \sigma_n^{-(k-2)}\sqrt{\frac{1}{n}} & \text{otherwise} \end{cases}$$

Following that, we again present active and passive algorithms for $k = 1$ first to gather intuition and then generalize them for $k > 1$. In Section 3 we will prove

Theorem 3 (Upper Bounds)**.** *Under the Berkson error model and assumption (Q),*

1. *For $k = 1$, a passive algorithm (WIDEHIST) and an active algorithm (ACTPASS) return $\widehat{t}$ s.t.*

$$\sup_{P \in \mathcal{P}(1,\sigma_n)} \mathbb{E}|\widehat{t} - t| \;\preceq\; \begin{cases} \frac{1}{n} & \text{if } \sigma_n \prec \frac{1}{n} \\ \sqrt{\frac{\sigma_n}{n}} & \text{otherwise} \end{cases}$$

$$\sup_{P \in \mathcal{P}(1,\sigma_n)} \mathbb{E}|\widehat{t} - t| \;\preceq\; \begin{cases} e^{-n} & \text{if } \sigma_n \prec e^{-n} \\ \frac{\sigma_n}{\sqrt{n}} & \text{otherwise} \end{cases}$$

2. *For $k > 1$, a passive algorithm (WIDEHIST) and an active algorithm (ACTPASS) return $\widehat{t}$ s.t.*

$$\sup_{P \in \mathcal{P}(k,\sigma_n)} \mathbb{E}|\widehat{t} - t| \;\preceq\; \begin{cases} n^{-\frac{1}{2k-1}} & \text{if } \sigma_n \prec n^{-\frac{1}{2k-1}} \\ \sigma_n^{-(k-\frac{3}{2})}\sqrt{\frac{1}{n}} & \text{otherwise} \end{cases}$$

$$\sup_{P \in \mathcal{P}(k,\sigma_n)} \mathbb{E}|\widehat{t} - t| \;\preceq\; \begin{cases} n^{-\frac{1}{2k-2}} & \text{if } \sigma_n \prec n^{-\frac{1}{2k-2}} \\ \sigma_n^{-(k-2)}\sqrt{\frac{1}{n}} & \text{otherwise} \end{cases}$$

3 Lower Bounds

To derive lower bounds, we will follow the approach of Ibargimov & Hasminskii (1981); Tsybakov (2009) which were exemplified in lower bounds for active learning problems without feature noise in Castro & Nowak (2007, 2008). The standard methodology is to reduce the problem of classification in the class $P(c, C, k, \sigma)$ to one of hypothesis testing. Similar to Castro & Nowak (2007, 2008), it will suffice to consider two hypotheses and use the following version of Fano's lemma from Tsybakov (2009) (Theorem 2.2).

Theorem 4 (Tsybakov (2009))**.** *Let $\mathcal{F}$ be a class of models. Associated with each $f \in \mathcal{F}$ we have a probability measure P_f defined on a common probability space. Let $d(.,.) : \mathcal{F}, \mathcal{F} \to \mathbb{R}$ be a semi-distance. Let $f_0, f_1 \in \mathcal{F}$ be such that $d(f_0, f_1) \geq 2a$, with $a > 0$. Also assume that $KL(P_{f_0}, P_{f_1}) \leq \gamma$, where KL denotes the Kullback-Leibler divergence. Then, the following bound holds:*

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{F}} P_f(d(\widehat{f}, f) \geq a) \geq \inf_{\widehat{t}} \max_{j \in \{0,1\}} P_{f_j}(d(\widehat{f}, f_j) \geq a) \geq \max\left(\frac{e^{-\gamma}}{4}, \frac{1 - \sqrt{\frac{\gamma}{2}}}{2}\right) =: \rho$$

where the inf *is taken with respect to the collection of all possible estimators of f based on a sample from P_f.*

Corollary 5. *If γ is a constant, then ρ is a constant, and by Markov's inequality, we would get*

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{F}} \mathbb{E} d(\widehat{f}, f) \geq \rho a$$

and the minimax risk under loss d would be $\succeq a$.

Proof of Theorem 2, $k = 1$**.** Choose $\mathcal{F} = \mathcal{P}(1, \sigma_n)$. Let $P_t \in \mathcal{P}(1, \sigma_n)$ denote a regression function with threshold at t. We choose the semi-metric to be the distance between thresholds, i.e. $d(P_r, P_s) = |r - s|$. We now choose two such distributions with thresholds at least $2a_n$ apart (we use a_n to explicitly remind the reader that a will later be set to depend on n) - let them be denoted P_{t_0} and P_{t_1} with $t_0 = -a_n, t_1 = a_n$ and

$$P_t(Y = +|X = x) = \begin{cases} 0.5 - c & x < t \ , \\ 0.5 + c & x \geq t \ . \end{cases}$$

Due to addition of noise, we get convolved distributions $P^0 = P_{t_0}(Y|W)$ and $P^1 := P_{t_1}(Y|W)$.

As hinted by the above corollary, we will choose a_n so that $KL(P^0, P^1)$ is bounded by a constant, to get a lower bound on risk $\succeq a_n$. This follows by the following argument from Castro & Nowak (2008).

The $KL(P^0, P^1)$ can be bounded as

$$\begin{aligned}
& \mathbb{E}^1_{W,Y}\left[\log \frac{P^1(W_1^n, Y_1^n)}{P^0(W_1^n, Y_1^n)}\right] && (2)\\
= \ & \mathbb{E}^1_{W,Y}\left[\log \frac{\prod_i P^1(Y_i|W_i) P(W_i|W_1^{i-1}, Y_1^{i-1})}{\prod_i P^0(Y_i|W_i) P(W_i|W_1^{i-1}, Y_1^{i-1})}\right] \\
= \ & \mathbb{E}^1_{W,Y}\left[\log \frac{\prod_i P^1(Y_i|W_i)}{\prod_i P^0(Y_i|W_i)}\right] && (3)\\
= \ & \sum_i \mathbb{E}^1_W\left[\mathbb{E}^1_Y\left[\log \frac{P^1(Y_i|W_i)}{P^0(Y_i|W_i)} \ \middle| \ W_1, ..., W_n\right]\right] && (4)\\
\leq \ & n \max_{w \in [-1,1]} \mathbb{E}^1_Y\left[\log \frac{P^1(Y|W)}{P^0(Y|W)} \ \middle| \ W = w\right] && (5)\\
\preceq \ & n \max_{w \in [-1,1]} (P^1(Y|w) - P^0(Y|w))^2 && (6)
\end{aligned}$$

where (3) holds for active learning because the algorithm determines W_i when given $\{W_1^{i-1}, Y_1^{i-1}\}$ and is independent of the model, and follows by the independence of future from past for passive learning. (4) holds by law of iterated expectation. (5) is used for active learning but is not needed for passive learning. (6) follows by an approximation

$$KL(Ber(1/2 + p), Ber(1/2 + q)) \preceq (p - q)^2$$

for sufficiently small constants p, q.

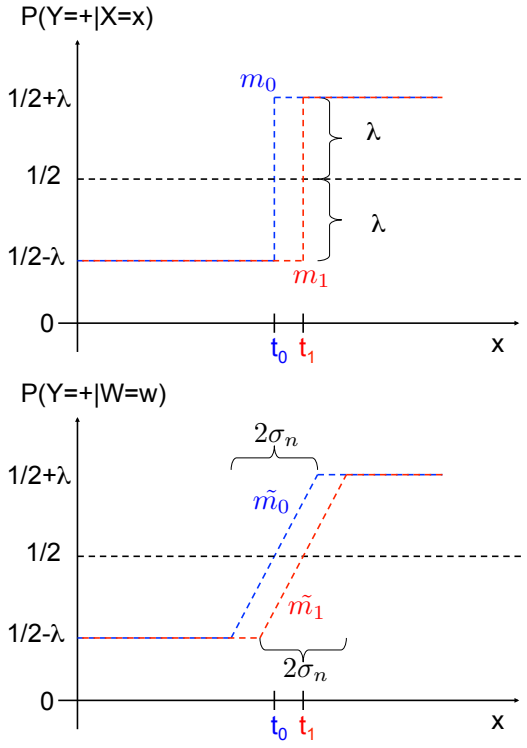

Figure 2: Regression functions before (top) and after (bottom) convolution with noise.

$F_t(w) := P_t(Y|W=w) = \int P_t(Y|X)P(X|W=w)dX$ and a straightforward calculation reveals that

$$F_t(w) = \begin{cases} 0.5 - c & w \le t - \sigma_n \ , \\ 0.5 + \frac{c}{\sigma_n}(w-t) & w \in [t-\sigma_n, t+\sigma_n] \ , \\ 0.5 + c & w \ge t + \sigma_n \ . \end{cases} \quad (7)$$

As depicted in Fig.2, note the behavior before and after convolution with noise: (i) $m(t) = F(t) = 1/2$, hence $F_1(a_n) = 1/2 = F_0(-a_n)$ (ii) Both convolved regression functions grow linearly for a region of width $2\sigma_n$, and differ only on a width of $2(\sigma_n + a_n)$; (iii) For a large region $[a_n - \sigma_n, -a_n + \sigma_n]$ of size $2(\sigma_n - a_n)$, we have $\big|F_1(w) - F_0(w)\big| = 2a_n c/\sigma_n$, a constant. Their gap varies when $\sigma_n \succeq a_n$ as $\big|F_0(w) - F_1(w)\big| =$

$$\begin{cases} \Big(w + a_n + \sigma_n\Big)\frac{c}{\sigma_n} & w \in [-a_n - \sigma_n, a_n - \sigma_n] \\ 2a_n \frac{c}{\sigma_n} & w \in [a_n - \sigma_n, -a_n + \sigma_n] \\ \Big((a_n + \sigma_n) - w\Big)\frac{c}{\sigma_n} & w \in [-a_n + \sigma_n, a_n + \sigma_n] \\ 0 & \text{otherwise.} \end{cases}$$

When $\sigma_n \prec a_n$, $\big|F_1(w) - F_0(w)\big| =$

$$\begin{cases} \Big(w + a_n + \sigma_n\Big)\frac{c}{\sigma_n} & w \in [-a_n - \sigma_n, -a_n + \sigma_n] \\ 2c & w \in [-a_n + \sigma_n, a_n - \sigma_n] \\ \Big((a_n + \sigma_n) - w\Big)\frac{c}{\sigma_n} & w \in [a_n - \sigma_n, a_n + \sigma_n] \\ 0 & \text{otherwise.} \end{cases}$$

For active learning, when $\sigma_n \succeq a_n$ we note

$$\max_{w \in [-1,1]} |P^1(Y|w) - P^0(Y|w)| = \frac{2a_n c}{\sigma_n}$$

and get $KL(P^0, P^1) \preceq n\frac{a_n^2}{\sigma_n^2}$ by Eq.(6). We choose $a_n \asymp \frac{\sigma_n}{\sqrt{n}}$, which becomes our active minimax error rate by Corollary 5 when $\sigma_n \succeq a_n$ i.e. $\sigma_n \succeq e^{-n}$.

Similarly, if $\sigma_n \prec \exp\{-n\}$, setting $a_n \asymp \exp\{-n\}$ easily gives us an exponentially small lower bound.

In the passive setting, Eq.(5) does not apply. Since the two convolved distributions differ only on an interval of size $2(\sigma_n + a_n)$, the effective number of points falling in this interval would be $\asymp n(\sigma_n + a_n)$.

When $\sigma_n \succeq a_n$, a simple calculation shows

$$KL(P^0, P^1) \ \preceq \ n(\sigma_n + a_n)\frac{a_n^2}{\sigma_n^2} \ \asymp \ n\frac{a_n^2}{\sigma_n},$$

giving rise to a choice of $a_n \asymp \sqrt{\frac{\sigma_n}{n}}$, which is the passive minimax rate when $\sigma_n \succeq a_n$ i.e. $\sigma_n \succeq \frac{1}{n}$.

When $\sigma_n \prec \frac{1}{n}$, a similar calculation shows

$$KL(P^0, P^1) \ \preceq \ n(\sigma_n + a_n)4c^2 \ \asymp \ na_n$$

giving rise to a choice of $a_n \asymp \frac{1}{n}$, which is the passive minimax rate when $\sigma_n \succeq a_n$ i.e. $\sigma_n \prec \frac{1}{n}$. ■

Proof of Theorem 2, $k > 1$ We follow a very similar setup to the case $k = 1$. The difference will lie in picking functions that are in $\mathcal{P}(c, C, k, \sigma_n)$ for general $k \neq 1$, and calculating the bounds on KL divergence appropriately. However, for notational convenience, we will assume that the domain is shifted to $[-\sigma_n, 2 - \sigma_n]$ instead of $[-1, 1]$ and that the distance between thresholds is a_n instead of $2a_n$. Define

$$P_0(Y|x) = \begin{cases} 1/2 - c|x|^{k-1} \text{ if } x \in [-\sigma_n, 0] \\ 1/2 + c|x|^{k-1} \text{ if } x > 0 \end{cases}$$

$$P_1(Y|x) = \begin{cases} 1/2 - c|x - a_n|^{k-1} \text{ if } x \in [-\sigma_n, a_n] \\ 1/2 + c|x - a_n|^{k-1} \text{ if } x \in [a_n, \beta a_n + \sigma_n] \\ 1/2 + c|x|^{k-1} \quad \text{if } x > \beta a_n + \sigma_n \end{cases}$$

where $\beta = \frac{1}{1 - (c/C)^{1/(k-1)}} \ge 1$ is a constant chosen such that $P_1 \in \mathcal{P}(c, C, k, \sigma_n)$ (this fact is verified explicitly in the Appendix). For ease of notation, P_0, P_1 are understood to actually saturate at $0, 1$ if need be (i.e. we are implicitly working with $\min\{P_{0/1}, 1\}$, etc). The two thresholds are clearly at $0, a_n$ respectively, and after the point $\beta a_n + \sigma_n$, the two functions are the same. Continuing the same notation as for $k = 1$, we let $P^i = P_i(Y|W) = F_i(w)$ for $i = 0, 1$.

The following claims hold true (Appendix).

1. When $\sigma_n \preceq a_n$, $\max_w |F_1(w) - F_2(w)| \asymp a_n^{k-1}$.
2. When $\sigma_n \succeq a_n$, $\max_w |F_1(w) - F_2(w)| \asymp \sigma_n^{k-2} a_n$.
3. As a subpart of the above cases, when $\sigma_n \asymp a_n$, $\max_w |F_1(w) - F_2(w)| \asymp \sigma_n^{k-2} a_n \asymp a_n^{k-1}$

If the above propositions are true, we can verify:

1. In the first case, $KL(P^0, P^1) \preceq n a_n^{2k-2}$, hence $a_n \asymp n^{-\frac{1}{2k-2}}$ is a lower bound when $\sigma_n \preceq n^{-\frac{1}{2k-2}}$.
2. Otherwise, $KL(P^0, P^1) \preceq n\sigma_n^{2k-4} a_n^2$, hence $a_n \asymp \frac{\sigma_n^{-(k-2)}}{\sqrt{n}}$ is a lower bound when $\sigma_n \succ n^{-\frac{1}{2k-2}}$.

The passive bounds follow by not just considering the maximum difference between $|F_1(w) - F_2(w)|$ but also the length of that difference, since it is directly proportional to the number of points that may randomly fall in that region. Following the same calculations,

1. When $\sigma_n \prec a_n$, $|F_1(w) - F_2(w)| \asymp a_n^{k-1}$ for all $w \in [0, \beta a_n + 2\sigma_n]$. Hence $KL(P^0, P^1) \preceq n(\beta a_n + 2\sigma_n) a_n^{2k-2} \asymp n a_n^{2k-1}$ and $a_n \asymp n^{-\frac{1}{2k-1}}$ is the minimax passive rate when $\sigma_n \prec n^{-\frac{1}{2k-1}}$.
2. When $\sigma_n \succ a_n$, $|F_1(w) - F_2(w)| \asymp \sigma_n^{k-2} a_n$ for all $w \in [0, \beta a_n + 2\sigma_n]$. Hence $KL(P^0, P^1) \preceq n(\beta a_n + 2\sigma_n)\sigma_n^{2k-4} a_n^2$ and $a_n \asymp \sigma_n^{-(k-\frac{3}{2})}\sqrt{\frac{1}{n}}$ is the minimax passive rate when $\sigma_n \succ n^{-\frac{1}{2k-1}}$.

as verified from the Appendix calculation. ■

4 Upper Bounds

For passive sampling, we present a modified histogram estimator, WIDEHIST, when the noise level σ_n is larger than the noiseless minimax rate of $1/n$. Assume for simplicity that the n sampled points on $[-1,1]$ are equally spaced to mimic a uniform distribution, lying at $\frac{(2j-1)}{2n}$, $j = 1, ..., n$.

Algorithm WIDEHIST.

1. Divide $[-1,1]$ into m bins of width $h > \frac{2}{n}$ so $m = \frac{2}{h} < n$. The i^{th} bin covers $[-1+(i-1)h, -1+ih]$, $i \in \{1, ..., m\}$ and hence each bin has $\frac{nh}{2}$ points. Let b_i be the average number of positive labels in bin i of these $\frac{nh}{2}$ points.
2. Let $\widehat{p}_i$ be the average of the b_i's over a all bins within $\pm\sigma_n/2$ of bin i. We "classify" regions with $\widehat{p}_i < 1/2$ as being $-$ and $\widehat{p}_i > 1/2$ as being $+$, and return $\widehat{t}$ as the center of the first bin from left to right where $\widehat{p}_i$ crosses half.

Observe that we need not operate on $[-1,1]$ with n queries - WIDEHIST(D,B) could take as inputs any domain D and any query budget B. The argument below hinges on the fact that the convolved regression function behaves linearly around t.

Proof of Theorem 3, $k = 1$, (Passive). Let $i^* \in \{1, ..., m\}$ denote the true bin $[(i^*-1)h, i^*h]$ that contains t. Let $\widehat{t}$ be from bin $\widehat{i}$, i.e. $\widehat{p}_{\widehat{i}} < 1/2$ and $\widehat{p}_{\widehat{i}+1} > 1/2$. We will argue that $\widehat{i}$ is very close to i^*, in which case the point error we suffer is $|\widehat{i} - i^*|h$. Specifically, we prove that all bins except $I^* = \{i^*-1, i^*, i^*+1\}$ will be "classified" correctly with high probability. In other words, we claim w.h.p. $\widehat{p}_i < 1/2$ if $i < i^*-1$ and $\widehat{p}_i > 1/2$ if $i > i^*+1$.

Indeed, we can show (Appendix)

$$\text{For } i > i^*+2,\ \mathbb{E}[\widehat{p}_i] \geq \mathbb{E}[\widehat{p}_{i^*+2}] \geq 1/2 + \tfrac{c}{\sigma_n}h \qquad (8)$$

$$\text{For } i < i^*-2,\ \mathbb{E}[\widehat{p}_i] \leq \mathbb{E}[\widehat{p}_{i^*-2}] \leq 1/2 - \tfrac{c}{\sigma_n}h \qquad (9)$$

Using Hoeffding's inequality, we get that for bin i, $\Pr(|\widehat{p}_i - p_i| > \epsilon) \leq 2\exp\left\{-2\frac{n\sigma_n}{2}\epsilon^2\right\}$ Taking union bound over all bins other than those in i^*-1, i^*, i^*+1 and setting $\epsilon = \frac{c}{\sigma_n}h$, we get

$$\Pr(\forall i \backslash I^*, |\widehat{p}_i - p_i| > \tfrac{c}{\sigma_n}h) \leq 2m\exp\left\{-2\tfrac{n\sigma_n}{2}\left(\tfrac{ch}{\sigma_n}\right)^2\right\}$$

So we get bins $i \backslash I^*$ correct and $\widehat{i} \in \{i^*-1, i^*, i^*+1\}$ with probability $\geq 1 - 2n\exp\left\{-n\sigma_n\left(\frac{ch}{\sigma_n}\right)^2\right\}$ since $m < n$. Setting $h = \frac{1}{c}\sqrt{\frac{\sigma_n}{n}\log(\frac{2n}{\delta})}$ makes this hold with probability $\geq 1-\delta$ so the point error $|\widehat{i} - i^*|h < 2h$ behaves like $h \preceq \sqrt{\frac{\sigma_n}{n}}$. ■

For active sampling when the noise level σ_n is larger than the minimax noiseless rate e^{-n}, we present a algorithm ACTPASS which makes its n queries on the domain $[-1,1]$ in E different epochs/rounds. As a subroutine, it uses any optimal passive learning algorithm, like WIDEHIST(D,B). In each round, ACTPASS runs WIDEHIST on progressively smaller domains D with a restricted budget B. Hence it "activizes" the WIDEHIST and achieves the optimal active rate in the process. This algorithm was inspired by a similar idea from Ramdas & Singh (2013).

Algorithm ACTPASS.

Let $E = \lceil \log(1/\sigma_n) \rceil$ be the number of epochs and $D_1 = [-1,1]$ denote the domain of "radius" $R_1 = 1$ around $t_0 = 0$. The budget of every epoch is a constant $B = n/E$. For epochs $1 \leq e \leq E$, do:

1. Query for B labels uniformly on D_e.
2. Let $t_e = \text{WIDEHIST}(D_e, B)$ be the returned estimator using the most recent samples and labels.
3. Define $D_{e+1} = [t_e - 2^{-e}, t_e + 2^{-e}] \cap [-1,1]$ with a radius of at most $R_{e+1} = 2^{-e}$ around t_e. Repeat.

Observe that ACTPASS runs while $R_e > \sigma_n$, since by design $E \geq \log(1/\sigma_n)$ so $\sigma_n \leq 2^{-E} = R_{E+1}$.

Proof of Theorem 3, $k = 2$, (Active). The analysis of ACTPASS proceeds in two stages depending on the value of σ_n. Initially, when R_e is large, it is possible that $\sigma_n \preceq R_e/n$ and in this phase, the passive algorithm WIDEHIST will behave as if it is in the noiseless setting since the noise is smaller than its noiseless rate. However, after some point, when R_e becomes quite small, $\sigma_n \succeq R_e/n$ is possible and then WIDEHIST will behave as if it is in the noisy setting since noise is larger than its noiseless rate. Observe that it cannot stay in the first phase till the end of the algorithm, since the first phase runs while $\sigma_n \preceq R_e/n$ but we know that $\sigma_n > R_{E+1}$ by construction, so there must be an epoch where it switches phases, and ends the algorithm in its second phase.

We prove (by a separate induction in each epoch) that with high probability, the true threshold t will always lie inside the domain at the start of every epoch (this is clearly true before the first epoch). We claim:

1. Before all e in phase one, $t \in D_e$ w.h.p.
2. Before all e in phase two, $t \in D_e$ w.h.p.

We prove these in the Appendix. If these are true, then in the second phase, WIDEHIST is in the large noise setting and it gets an error of $\sqrt{\frac{R_e\sigma_n}{B}}$. Hence the final error of the algorithm is $\sqrt{\frac{R_E\sigma_n}{n/E}} \asymp \frac{\sigma_n}{\sqrt{n}}$. ■

Proof of Theorem 3, $k > 1$. The proofs for $k > 1$ are simply generalizations of those for $k = 1$. Again, we present concise arguments here for the settings where the algorithm can actually detect noise, i.e. when the noise level is larger than the noiseless minimax rate (otherwise, one can argue that algorithms which worked for the noiseless case will suffice). In both cases, the algorithm remains unchanged.

1. We outline the proof for WIDEHIST when $\sigma_n \succeq n^{-\frac{1}{2k-1}}$. Using similar notation as before, we will again show that if t is in bin i^* of width $h < \sigma_n$, then except for bins $i^* - 1, i^*, i^* + 1$, we will "classify" all other bins correct with high probability, by averaging over the $n\sigma_n/2$ points to the left and right of that bin. Specifically, we claim

$$\text{For } i > i^* + 2,\ \mathbb{E}[\widehat{p}_i] \geq \mathbb{E}[\widehat{p}_{i^*+2}] \ \geq\ 1/2 + \lambda\sigma_n^{k-2}h \quad (10)$$
$$\text{For } i < i^* - 2,\ \mathbb{E}[\widehat{p}_i] \leq \mathbb{E}[\widehat{p}_{i^*-2}] \ \leq\ 1/2 - \lambda\sigma_n^{k-2}h \quad (11)$$

A similar use of Hoeffding's inequality gives

$$\Pr(\forall i \backslash I^*, |\widehat{p}_i - p_i| \ > \ \lambda\sigma_n^{k-2}h) \ \leq \ 2m\exp\left\{-2(\tfrac{n\sigma_n}{2R})h^2\lambda^2\sigma_n^{2k-4}\right\}.$$

Arguing as before, w.h.p. we get a point error of $h \preceq \sqrt{\frac{R}{\sigma_n^{2k-3}n}} < \sigma_n$ when $\sigma_n \succ n^{-\frac{1}{2k-1}}$.

2. We outline the proof for ACTPASS when $\sigma_n \succeq n^{-\frac{1}{2k-2}}$. As before, the algorithm runs in two phases, and we will prove required properties within each phase by induction.

The first phase is when R_e is large and so σ_n may possibly be smaller than $(R_e/n)^{\frac{1}{2k-1}}$ and WIDEHIST will achieve noiseless rates within each epoch. In the second phase, after R_e has shrunk enough, σ_n will become larger than $(R_e/n)^{\frac{1}{2k-1}}$ and WIDEHIST will achieve noisy rates in these epochs.

One can verify, as before, that the second phase must occur, by design. Intuitively, the second phase must occur because we make a fixed number of queries $n/E \asymp n/\log n$ in a halving domain size (equivalently we make geometrically increasing queries on a rescaled domain), and so relatively in successive epochs this noiseless error shrinks, and at some point σ_n becomes larger than this shrinking noiseless error rate.

As before we make the following claims:

1. Before all e in phase one $t \in D_e$ w.h.p.
2. Before all e in phase two $t \in D_e$ w.h.p.

These are proved in the Appendix by induction.

The final point error is given by WIDEHIST in the last epoch as $\sqrt{\frac{R_E}{\sigma_n^{2k-3}n/E}} \asymp \frac{1}{\sigma_n^{k-2}}\sqrt{\frac{1}{n}}$ since $R_E \asymp \sigma_n$ and $E \asymp \log n$.

5 Conclusion

In this paper, we propose a simple Berkson error model for one-dimensional threshold classification, inspired by the setup and model analysed in Castro & Nowak (2007, 2008), in which we can analyse active learning with additive uniform feature noise. To the best of our knowledge, this is the first attempt at jointly tackling feature noise and label noise in active learning.

This simple setting already yields interesting behaviour depending on the additive feature noise level and the label noise of the underlying regression function. For both passive and active learning, whenever the noise level is smaller than the minimax noiseless rate, the learner cannot notice that there is noise, and will continue to achieve the noiseless rate. As the noise gets larger, the rates do depend on the noise level. Importantly, one can achieve better rates than passive learning in most scenarios, and we propose unique algorithms/estimators to achieve tight rates. The idea of "activizing" passive algorithms, like algorithm ACTPASS did, seems especially powerful and could carry forward to other settings beyond our paper and Ramdas & Singh (2013).

The immediate future work and most direct extension to this paper concerns the main weakness of the paper - the possibility of getting rid of Assumption (M), which is the only hurdle to a fair comparision with the noiseless setting. We would like to re-emphasize that at first glance, the rates may be misleading and counterintuitive because it "appears" as if larger noise could possibly help estimation due to the presence of σ_n in the denominator for larger k.

However, we point out once more that the class of functions is not constant over all σ_n - it depends on σ_n, and in fact it gets "smaller" in some sense with larger σ_n because the assumption (M) becomes more stringent. This observation about the non-constant function class, along with the fact that convolution with uniform noise seems to *unflatten* the regression function as shown in the figures, together cause the rates to seemingly improve with larger noise levels.

Analysing the case without (M) seems to be quite a challenging task since the noiseless and convolved thresholds can be different - we did attempt to formulate a few kernel-based estimators with additional assumptions, but do not presently have tight bounds, and leave those for a future work.

Acknowledgements

We thank Rui Castro for detailed conversations about our model and results. This work is supported in part by NSF Big Data grant IIS-1247658.

References

Carroll, Raymond J, Ruppert, David, Stefanski, Leonard A, and Crainiceanu, Ciprian M. *Measurement error in nonlinear models: a modern perspective*, volume 105. Chapman and Hall/CRC, 2010.

Castro, R. and Nowak, R. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, 2008.

Castro, Rui M. and Nowak, Robert D. Minimax bounds for active learning. In *Proceedings of the 20th annual conference on Learning theory*, COLT'07, pp. 5–19, Berlin, Heidelberg, 2007. Springer-Verlag.

Fan, Jianqing. On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, pp. 1257–1272, 1991.

Fan, Jianqing, Truong, Young K, et al. Nonparametric regression with errors in variables. *The Annals of Statistics*, 21(4):1900–1925, 1993.

Fuller, Wayne A. *Measurement error models*, volume 305. Wiley, 2009.

Ibargimov, I. A. and Hasminskii, R. Z. *Statistical Estimation. Asymptotic Theory*. Springer-Verlag, 1981.

Loustau, Sébastien and Marteau, Clément. Discriminant analysis with errors in variables. *arXiv preprint arXiv:1201.3283*, 2012.

Ramdas, Aaditya and Singh, Aarti. Algorithmic connections between active learning and stochastic convex optimization. In *Algorithmic Learning Theory*, pp. 339–353. Springer, 2013.

Steinwart, Ingo and Scovel, Clint. Fast rates to bayes for kernel machines. In *Advances in neural information processing systems*, pp. 1345–1352, 2004.

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

Tsybakov, Alexandre B. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2009. ISBN 0387790519, 9780387790510.