Minimax Rates for Conditional Density Estimation via Empirical Entropy

Blair Bilodeau[1]
(joint work with Dylan J. Foster[2] and Daniel M. Roy[1])

April 23, 2021

Statistical Sciences Research Day

[1]University of Toronto and Vector Institute
[2]Massachusetts Institute of Technology

Regression with Uncertainty

Regression with Uncertainty

Observe: $X_{1:n} \sim \mu^{\otimes n}$ and $Y_{1:n} = \mathrm{Noise}\Big[f^*(X_{1:n})\Big]$.

Regression with Uncertainty

Observe: $X_{1:n} \sim \mu^{\otimes n}$ and $Y_{1:n} = \mathrm{Noise}\Big[f^*(X_{1:n})\Big]$.
Goal: Given a new $X \sim \mu$, predict Y.

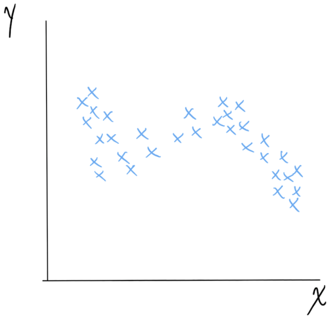
Regression with Uncertainty

Observe: $X_{1:n} \sim \mu^{\otimes n}$ and $Y_{1:n} = \mathrm{Noise}\Big[f^*(X_{1:n})\Big]$.
Goal: Given a new $X \sim \mu$, predict Y.
Regression: Approximate $\mathbb{E}[Y \mid X]$ with $\hat{f} : \mathcal{X} \to \mathcal{Y}$.

Regression with Uncertainty

Observe: $X_{1:n} \sim \mu^{\otimes n}$ and $Y_{1:n} = \mathrm{Noise}\Big[f^*(X_{1:n})\Big]$.
Goal: Given a new $X \sim \mu$, predict Y.
Regression: Approximate $\mathbb{E}[Y \mid X]$ with $\hat{f} : \mathcal{X} \to \mathcal{Y}$.

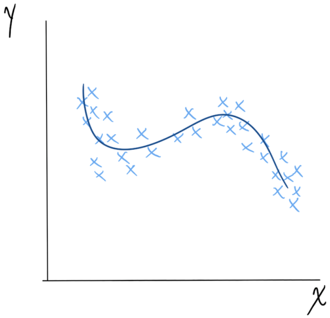
Observe some data...



Regression with Uncertainty

Observe: $X_{1:n} \sim \mu^{\otimes n}$ and $Y_{1:n} = \mathrm{Noise}\Big[f^*(X_{1:n})\Big]$.
Goal: Given a new $X \sim \mu$, predict Y.
Regression: Approximate $\mathbb{E}[Y \mid X]$ with $\hat{f} : \mathcal{X} \to \mathcal{Y}$.

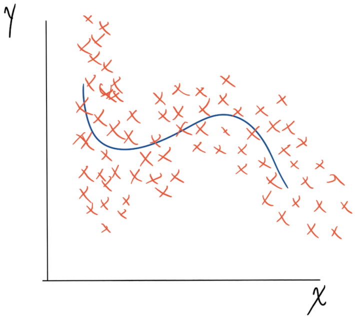
Draw your favourite curve...



Regression with Uncertainty

Observe: $X_{1:n} \sim \mu^{\otimes n}$ and $Y_{1:n} = \mathrm{Noise}\Big[f^*(X_{1:n})\Big]$.
Goal: Given a new $X \sim \mu$, predict Y.
Regression: Approximate $\mathbb{E}[Y \mid X]$ with $\hat{f} : \mathcal{X} \to \mathcal{Y}$.

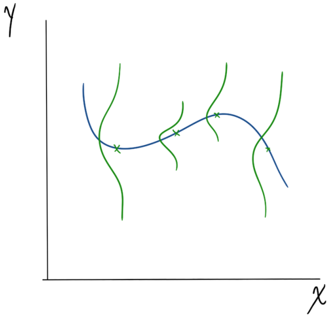
This curve works for lots of data...



Regression with Uncertainty

Observe: $X_{1:n} \sim \mu^{\otimes n}$ and $Y_{1:n} = \mathrm{Noise}\Big[f^*(X_{1:n})\Big]$.
Goal: Given a new $X \sim \mu$, predict Y.
Regression: Approximate $\mathbb{E}[Y \mid X]$ with $\hat{f} : \mathcal{X} \to \mathcal{Y}$.

Idea: Approximate $f^*(Y \mid X)$ with $\hat{f} : \mathcal{X} \to \mathcal{M}(\mathcal{Y}) = \{\text{densities on } \mathcal{Y}\}$



Measuring Performance

Measuring Performance

For regression we could use square / absolute / classification loss...

Measuring Performance

For regression we could use square / absolute / classification loss...

...but now we want to capture the quality of our predictions in the tails.

Measuring Performance

For regression we could use square / absolute / classification loss...

...but now we want to capture the quality of our predictions in the tails.

We use **log loss** to achieve this:

$$\ell\left(\hat{f}, (X, Y)\right) = -\log\left(\hat{f}(Y \mid X)\right).$$

Measuring Performance

For regression we could use square / absolute / classification loss...

...but now we want to capture the quality of our predictions in the tails.

We use **log loss** to achieve this:

$$\ell\left(\hat{f}, (X, Y)\right) = -\log\left(\hat{f}(Y \mid X)\right).$$

This is just the negative log-likelihood of your predictive density.

Measuring Performance

For regression we could use square / absolute / classification loss...

...but now we want to capture the quality of our predictions in the tails.

We use **log loss** to achieve this:

$$\ell\Big(\hat{f},(X,Y)\Big) = -\log\Big(\hat{f}(Y \mid X)\Big).$$

This is just the negative log-likelihood of your predictive density.

Being confidently wrong is worse than being ambivalent (once in a while).

Measuring Performance

For regression we could use square / absolute / classification loss...

...but now we want to capture the quality of our predictions in the tails.

We use **log loss** to achieve this:

$$\ell\Big(\hat{f}, (X, Y)\Big) = -\log\Big(\hat{f}(Y \mid X)\Big).$$

This is just the negative log-likelihood of your predictive density.

Being confidently wrong is worse than being ambivalent (once in a while).

Consider the simple case of estimating the probability of rain, p:

Measuring Performance

For regression we could use square / absolute / classification loss...

...but now we want to capture the quality of our predictions in the tails.

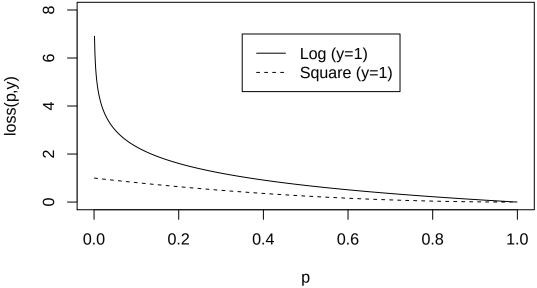
We use **log loss** to achieve this:

$$\ell\Big(\hat{f}, (X, Y)\Big) = -\log\Big(\hat{f}(Y \mid X)\Big).$$

This is just the negative log-likelihood of your predictive density.

Being confidently wrong is worse than being ambivalent (once in a while).

Consider the simple case of estimating the probability of rain, p:



Minimax Performance

Minimax Performance

If $f^*(Y \mid X)$ can vary wildly in X, I can do arbitrarily bad in the tails.

Minimax Performance

If $f^*(Y \mid X)$ can vary wildly in X, I can do arbitrarily bad in the tails.

I need to make some assumption about my data...

Minimax Performance

If $f^*(Y \mid X)$ can vary wildly in X, I can do arbitrarily bad in the tails.

I need to make some assumption about my data...

...but I want to know how this assumption might affect my predictive performance.

Minimax Performance

If $f^*(Y \mid X)$ can vary wildly in X, I can do arbitrarily bad in the tails.

I need to make some assumption about my data...

...but I want to know how this assumption might affect my predictive performance.

The Big Assumption

Suppose the data-generating $f^* : \mathcal{X} \to \mathcal{M}(\mathcal{Y})$ is in some known set: $f^* \in \mathcal{F}$.

Minimax Performance

If $f^*(Y \mid X)$ can vary wildly in X, I can do arbitrarily bad in the tails.

I need to make some assumption about my data...

...but I want to know how this assumption might affect my predictive performance.

The Big Assumption

Suppose the data-generating $f^* : \mathcal{X} \to \mathcal{M}(\mathcal{Y})$ is in some known set: $f^* \in \mathcal{F}$.

This is the classical statistics assumption: i.i.d. data with a well-specified model.

Minimax Performance

If $f^*(Y \mid X)$ can vary wildly in X, I can do arbitrarily bad in the tails.

I need to make some assumption about my data...

...but I want to know how this assumption might affect my predictive performance.

The Big Assumption

Suppose the data-generating $f^* : \mathcal{X} \to \mathcal{M}(\mathcal{Y})$ is in some known set: $f^* \in \mathcal{F}$.

This is the classical statistics assumption: i.i.d. data with a well-specified model.

Minimax Risk

$$\mathcal{R}_n(\mathcal{F}) = \inf_{\hat{f}} \sup_{\mu} \sup_{f^* \in \mathcal{F}} \mathbb{E}_{X_{1:n}, Y_{1:n}} \mathbb{E}_X \mathrm{KL}\Big(f^*(\cdot \mid X) \,\|\, \hat{f}_n(\cdot \mid X)\Big).$$

Minimax Performance

If $f^*(Y \mid X)$ can vary wildly in X, I can do arbitrarily bad in the tails.
I need to make some assumption about my data...
...but I want to know how this assumption might affect my predictive performance.

The Big Assumption
Suppose the data-generating $f^* : \mathcal{X} \to \mathcal{M}(\mathcal{Y})$ is in some known set: $f^* \in \mathcal{F}$.
This is the classical statistics assumption: i.i.d. data with a well-specified model.

Minimax Risk

$$\mathcal{R}_n(\mathcal{F}) = \inf_{\hat{f}} \sup_{\mu} \sup_{f^* \in \mathcal{F}} \mathbb{E}_{X_{1:n}, Y_{1:n}} \mathbb{E}_X \mathrm{KL}\Big(f^*(\cdot \mid X) \,\|\, \hat{f}_n(\cdot \mid X) \Big).$$

Best-case predictions against the worst-case data distribution,
in expectation over data of the accuracy in expectation over a new covariate.

Minimax Performance

If $f^*(Y \mid X)$ can vary wildly in X, I can do arbitrarily bad in the tails.
I need to make some assumption about my data...
...but I want to know how this assumption might affect my predictive performance.

The Big Assumption

Suppose the data-generating $f^* : \mathcal{X} \to \mathcal{M}(\mathcal{Y})$ is in some known set: $f^* \in \mathcal{F}$.
This is the classical statistics assumption: i.i.d. data with a well-specified model.

Minimax Risk

$$\mathcal{R}_n(\mathcal{F}) = \inf_{\hat{f}} \sup_{\mu} \sup_{f^* \in \mathcal{F}} \mathbb{E}_{X_{1:n}, Y_{1:n}} \mathbb{E}_X \mathrm{KL}\Big(f^*(\cdot \mid X) \,\|\, \hat{f}_n(\cdot \mid X)\Big).$$

Best-case predictions against the worst-case data distribution,
in expectation over data of the accuracy in expectation over a new covariate.

We provide an explicit algorithm that achieves the $\inf$ within $\log$ factors.
This works for every $\mathcal{F}$, and looks like a Bayesian mixture density.

Minimax Performance

If $f^*(Y \mid X)$ can vary wildly in X, I can do arbitrarily bad in the tails.
I need to make some assumption about my data...
...but I want to know how this assumption might affect my predictive performance.

The Big Assumption

Suppose the data-generating $f^* : \mathcal{X} \to \mathcal{M}(\mathcal{Y})$ is in some known set: $f^* \in \mathcal{F}$.
This is the classical statistics assumption: i.i.d. data with a well-specified model.

Minimax Risk

$$\mathcal{R}_n(\mathcal{F}) = \inf_{\hat{f}} \sup_{\mu} \sup_{f^* \in \mathcal{F}} \mathbb{E}_{X_{1:n}, Y_{1:n}} \mathbb{E}_X \mathrm{KL}\Big(f^*(\cdot \mid X) \,\|\, \hat{f}_n(\cdot \mid X)\Big).$$

Best-case predictions against the worst-case data distribution,
in expectation over data of the accuracy in expectation over a new covariate.

We provide an explicit algorithm that achieves the $\inf$ within $\log$ factors.
This works for every $\mathcal{F}$, and looks like a Bayesian mixture density.

Examples of Rates (BFR21)

Examples of Rates (BFR21)

Truncated Generalized Linear Model

$Y \mid X$ follows Exponential family distribution truncated to $[-B, B]$.
Location parameter is a linear function of X in the unit $\|\cdot\|_2$-ball on $\mathbb{R}^d$.

$$\mathcal{R}_n(\mathcal{F}) \lesssim \frac{\log(nB)}{\sqrt{n}}.$$

Examples of Rates (BFR21)

Truncated Generalized Linear Model

$Y \mid X$ follows Exponential family distribution truncated to $[-B, B]$.

Location parameter is a linear function of X in the unit $\|\cdot\|_2$-ball on $\mathbb{R}^d$.

$$\mathcal{R}_n(\mathcal{F}) \lesssim \frac{\log(nB)}{\sqrt{n}}.$$

VC-Type Classes (Solved open problem from ALT 2021)

$\mathcal{X}$ arbitrary, $Y \mid X \sim \mathrm{Bernoulli}(p(X))$,

where $p(X) = a + b\,\mathbb{I}\{X \in c\}$ for some $a, b > 0$ and subset $c \in \mathcal{C}$.

$$\mathcal{R}_n(\mathcal{F}) \lesssim \frac{\mathrm{VCdim}(\mathcal{C})\ \log(n)}{n}.$$

Examples of Rates (BFR21)

Truncated Generalized Linear Model

$Y \mid X$ follows Exponential family distribution truncated to $[-B, B]$.

Location parameter is a linear function of X in the unit $\|\cdot\|_2$-ball on $\mathbb{R}^d$.

$$\mathcal{R}_n(\mathcal{F}) \lesssim \frac{\log(nB)}{\sqrt{n}}.$$

VC-Type Classes (Solved open problem from ALT 2021)

$\mathcal{X}$ arbitrary, $Y \mid X \sim \mathrm{Bernoulli}(p(X))$,

where $p(X) = a + b\,\mathbb{I}\{X \in c\}$ for some $a, b > 0$ and subset $c \in \mathcal{C}$.

$$\mathcal{R}_n(\mathcal{F}) \lesssim \frac{\mathrm{VCdim}(\mathcal{C})\,\log(n)}{n}.$$

Nonparametric Conditional Densities

$Y \mid X$ has an α-Hölder continuous conditional density on $\mathcal{X} = [0,1]^d$.

$$\mathcal{R}_n(\mathcal{F}) \lesssim n^{-\frac{\alpha}{\alpha+d}} \log(n).$$

Examples of Rates (BFR21)

Truncated Generalized Linear Model

$Y \mid X$ follows Exponential family distribution truncated to $[-B, B]$.

Location parameter is a linear function of X in the unit $\|\cdot\|_2$-ball on $\mathbb{R}^d$.

$$\mathcal{R}_n(\mathcal{F}) \lesssim \frac{\log(nB)}{\sqrt{n}}.$$

VC-Type Classes (Solved open problem from ALT 2021)

$\mathcal{X}$ arbitrary, $Y \mid X \sim \mathrm{Bernoulli}(p(X))$,

where $p(X) = a + b\,\mathbb{I}\{X \in c\}$ for some $a, b > 0$ and subset $c \in \mathcal{C}$.

$$\mathcal{R}_n(\mathcal{F}) \lesssim \frac{\mathrm{VCdim}(\mathcal{C})\ \log(n)}{n}.$$

Nonparametric Conditional Densities

$Y \mid X$ has an α-Hölder continuous conditional density on $\mathcal{X} = [0, 1]^d$.

$$\mathcal{R}_n(\mathcal{F}) \lesssim n^{-\frac{\alpha}{\alpha+d}} \log(n).$$

Our lower bounds match the polynomial dependence on n.

Examples of Rates (BFR21)

Truncated Generalized Linear Model

$Y \mid X$ follows Exponential family distribution truncated to $[-B, B]$.

Location parameter is a linear function of X in the unit $\|\cdot\|_2$-ball on $\mathbb{R}^d$.

$$\mathcal{R}_n(\mathcal{F}) \lesssim \frac{\log(nB)}{\sqrt{n}}.$$

VC-Type Classes (Solved open problem from ALT 2021)

$\mathcal{X}$ arbitrary, $Y \mid X \sim \mathrm{Bernoulli}(p(X))$,

where $p(X) = a + b\,\mathbb{I}\{X \in c\}$ for some $a, b > 0$ and subset $c \in \mathcal{C}$.

$$\mathcal{R}_n(\mathcal{F}) \lesssim \frac{\mathrm{VCdim}(\mathcal{C})\,\log(n)}{n}.$$

Nonparametric Conditional Densities

$Y \mid X$ has an α-Hölder continuous conditional density on $\mathcal{X} = [0, 1]^d$.

$$\mathcal{R}_n(\mathcal{F}) \lesssim n^{-\frac{\alpha}{\alpha+d}} \log(n).$$

Our lower bounds match the polynomial dependence on n.

Complexity of $\mathcal{F}$

Complexity of $\mathcal{F}$

Well-specified is more reasonable for a complex $\mathcal{F}$, but estimation will be harder.

Complexity of $\mathcal{F}$

Well-specified is more reasonable for a complex $\mathcal{F}$, but estimation will be harder.

What is the formal notion of complexity that determines the minimax rates?

Complexity of $\mathcal{F}$

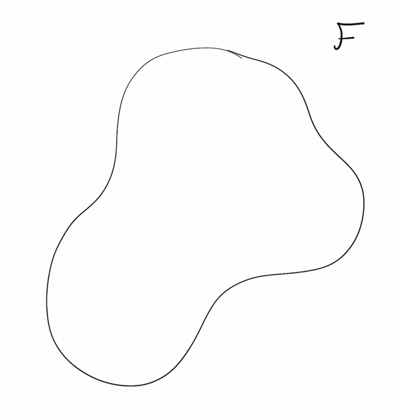
Well-specified is more reasonable for a complex $\mathcal{F}$, but estimation will be harder. What is the formal notion of complexity that determines the minimax rates?

Entropy measures how many functions are needed to discretely approximate $\mathcal{F}$.

Complexity of $\mathcal{F}$

Well-specified is more reasonable for a complex $\mathcal{F}$, but estimation will be harder. What is the formal notion of complexity that determines the minimax rates?

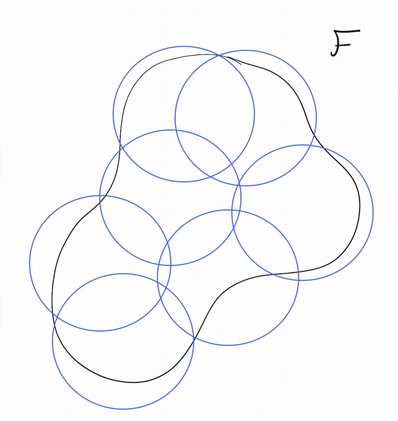
Entropy measures how many functions are needed to discretely approximate $\mathcal{F}$.



Complexity of $\mathcal{F}$

Well-specified is more reasonable for a complex $\mathcal{F}$, but estimation will be harder. What is the formal notion of complexity that determines the minimax rates?

Entropy measures how many functions are needed to discretely approximate $\mathcal{F}$.

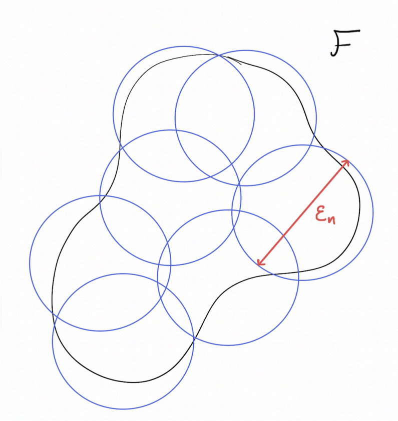


Complexity of $\mathcal{F}$

Well-specified is more reasonable for a complex $\mathcal{F}$, but estimation will be harder. What is the formal notion of complexity that determines the minimax rates?

Entropy measures how many functions are needed to discretely approximate $\mathcal{F}$.

How should the notion of size be chosen?



Existing Results

Existing Results

Nonparametric Regression

Existing Results

Nonparametric Regression

Minimax performance for regression with square loss is well-studied.

Existing Results

Nonparametric Regression

Minimax performance for regression with square loss is well-studied.

Rakhlin et al. (2017) define entropy $\mathcal{H}^{\mathrm{Sq.Loss}}$ satisfying

$$\mathcal{H}_n^{\mathrm{Sq.Loss}}(\mathcal{F}, \varepsilon_n) \asymp n\varepsilon_n^2 \implies \mathcal{R}_n^{\mathrm{Sq.Loss}}(\mathcal{F}) \asymp \varepsilon_n^2.$$

Existing Results

Nonparametric Regression

Minimax performance for regression with square loss is well-studied.

Rakhlin et al. (2017) define entropy $\mathcal{H}^{\mathrm{Sq.Loss}}$ satisfying

$$\mathcal{H}_n^{\mathrm{Sq.Loss}}(\mathcal{F}, \varepsilon_n) \asymp n\varepsilon_n^2 \implies \mathcal{R}_n^{\mathrm{Sq.Loss}}(\mathcal{F}) \asymp \varepsilon_n^2.$$

This type of relationship is also classically known; it appears in LeCam (1973).

Existing Results

Nonparametric Regression

Minimax performance for regression with square loss is well-studied.

Rakhlin et al. (2017) define entropy $\mathcal{H}^{\mathrm{Sq.Loss}}$ satisfying

$$\mathcal{H}_n^{\mathrm{Sq.Loss}}(\mathcal{F}, \varepsilon_n) \asymp n\varepsilon_n^2 \implies \mathcal{R}_n^{\mathrm{Sq.Loss}}(\mathcal{F}) \asymp \varepsilon_n^2.$$

This type of relationship is also classically known; it appears in LeCam (1973).

Problem #1: This entropy is for real-valued $\mathcal{F}$, our regressors are *function-valued*!

Existing Results

Nonparametric Regression

Minimax performance for regression with square loss is well-studied.

Rakhlin et al. (2017) define entropy $\mathcal{H}^{\mathrm{Sq.Loss}}$ satisfying

$$\mathcal{H}_n^{\mathrm{Sq.Loss}}(\mathcal{F}, \varepsilon_n) \asymp n\varepsilon_n^2 \implies \mathcal{R}_n^{\mathrm{Sq.Loss}}(\mathcal{F}) \asymp \varepsilon_n^2.$$

This type of relationship is also classically known; it appears in LeCam (1973).

Problem #1: This entropy is for real-valued $\mathcal{F}$, our regressors are *function-valued*!

Density Estimation

Existing Results

Nonparametric Regression

Minimax performance for regression with square loss is well-studied.

Rakhlin et al. (2017) define entropy $\mathcal{H}^{\mathrm{Sq.Loss}}$ satisfying

$$\mathcal{H}_n^{\mathrm{Sq.Loss}}(\mathcal{F}, \varepsilon_n) \asymp n\varepsilon_n^2 \implies \mathcal{R}_n^{\mathrm{Sq.Loss}}(\mathcal{F}) \asymp \varepsilon_n^2.$$

This type of relationship is also classically known; it appears in LeCam (1973).

Problem #1: This entropy is for real-valued $\mathcal{F}$, our regressors are *function-valued*!

Density Estimation

Joint density estimation (e.g., of $p(X, Y)$) is also well-studied.

Existing Results

Nonparametric Regression

Minimax performance for regression with square loss is well-studied.

Rakhlin et al. (2017) define entropy $\mathcal{H}^{\text{Sq.Loss}}$ satisfying

$$\mathcal{H}_n^{\text{Sq.Loss}}(\mathcal{F}, \varepsilon_n) \asymp n\varepsilon_n^2 \implies \mathcal{R}_n^{\text{Sq.Loss}}(\mathcal{F}) \asymp \varepsilon_n^2.$$

This type of relationship is also classically known; it appears in LeCam (1973).

Problem #1: This entropy is for real-valued $\mathcal{F}$, our regressors are *function-valued*!

Density Estimation

Joint density estimation (e.g., of $p(X, Y)$) is also well-studied.

Yang and Barron (1999) define *a different* entropy $\mathcal{H}^{\text{Joint}}$ satisfying

$$\mathcal{H}_n^{\text{Joint}}(\mathcal{F}, \varepsilon_n) \asymp n\varepsilon_n^2 \implies \mathcal{R}_n^{\text{Joint}}(\mathcal{F}) \asymp \varepsilon_n^2.$$

Existing Results

Nonparametric Regression

Minimax performance for regression with square loss is well-studied.

Rakhlin et al. (2017) define entropy $\mathcal{H}^{\mathrm{Sq.Loss}}$ satisfying

$$\mathcal{H}_n^{\mathrm{Sq.Loss}}(\mathcal{F}, \varepsilon_n) \asymp n\varepsilon_n^2 \implies \mathcal{R}_n^{\mathrm{Sq.Loss}}(\mathcal{F}) \asymp \varepsilon_n^2.$$

This type of relationship is also classically known; it appears in LeCam (1973).

Problem #1: This entropy is for real-valued $\mathcal{F}$, our regressors are *function-valued*!

Density Estimation

Joint density estimation (e.g., of $p(X, Y)$) is also well-studied.

Yang and Barron (1999) define *a different* entropy $\mathcal{H}^{\mathrm{Joint}}$ satisfying

$$\mathcal{H}_n^{\mathrm{Joint}}(\mathcal{F}, \varepsilon_n) \asymp n\varepsilon_n^2 \implies \mathcal{R}_n^{\mathrm{Joint}}(\mathcal{F}) \asymp \varepsilon_n^2.$$

Problem #2: We shouldn't have to estimate the marginal distribution on $\mathcal{X}$!

Existing Results

Nonparametric Regression

Minimax performance for regression with square loss is well-studied.

Rakhlin et al. (2017) define entropy $\mathcal{H}^{\mathrm{Sq.Loss}}$ satisfying

$$\mathcal{H}_n^{\mathrm{Sq.Loss}}(\mathcal{F}, \varepsilon_n) \asymp n\varepsilon_n^2 \implies \mathcal{R}_n^{\mathrm{Sq.Loss}}(\mathcal{F}) \asymp \varepsilon_n^2.$$

This type of relationship is also classically known; it appears in LeCam (1973).

Problem #1: This entropy is for real-valued $\mathcal{F}$, our regressors are *function-valued*!

Density Estimation

Joint density estimation (e.g., of $p(X, Y)$) is also well-studied.

Yang and Barron (1999) define *a different* entropy $\mathcal{H}^{\mathrm{Joint}}$ satisfying

$$\mathcal{H}_n^{\mathrm{Joint}}(\mathcal{F}, \varepsilon_n) \asymp n\varepsilon_n^2 \implies \mathcal{R}_n^{\mathrm{Joint}}(\mathcal{F}) \asymp \varepsilon_n^2.$$

Problem #2: We shouldn't have to estimate the marginal distribution on $\mathcal{X}$!

Main Results

Theorem (BFR21)

We define a new notion of entropy $\mathcal{H}$ such that

$$\mathcal{H}_n(\mathcal{F}, \varepsilon_n) \asymp n\varepsilon_n^2 \implies \mathcal{R}_n(\mathcal{F}) \asymp \varepsilon_n^2$$

for conditional density estimation.

Main Results

Theorem (BFR21)

We define a new notion of entropy $\mathcal{H}$ *such that*

$$\mathcal{H}_n(\mathcal{F}, \varepsilon_n) \asymp n\varepsilon_n^2 \implies \mathcal{R}_n(\mathcal{F}) \asymp \varepsilon_n^2$$

for conditional density estimation.

Highlights of what this means

Main Results

Theorem (BFR21)

We define a new notion of entropy $\mathcal{H}$ such that

$$\mathcal{H}_n(\mathcal{F}, \varepsilon_n) \asymp n\varepsilon_n^2 \implies \mathcal{R}_n(\mathcal{F}) \asymp \varepsilon_n^2$$

for conditional density estimation.

Highlights of what this means

a) We obtain the minimax rates (as a function of n) for all $\mathcal{F}$ simultaneously.

Main Results

Theorem (BFR21)

We define a new notion of entropy $\mathcal{H}$ such that

$$\mathcal{H}_n(\mathcal{F}, \varepsilon_n) \asymp n\varepsilon_n^2 \implies \mathcal{R}_n(\mathcal{F}) \asymp \varepsilon_n^2$$

for conditional density estimation.

Highlights of what this means

a) We obtain the minimax rates (as a function of n) for all $\mathcal{F}$ simultaneously.

b) Existing joint density results require estimating the covariate distribution, which we eliminate for conditional density estimation.

Main Results

Theorem (BFR21)

We define a new notion of entropy $\mathcal{H}$ such that

$$\mathcal{H}_n(\mathcal{F}, \varepsilon_n) \asymp n\varepsilon_n^2 \implies \mathcal{R}_n(\mathcal{F}) \asymp \varepsilon_n^2$$

for conditional density estimation.

Highlights of what this means

a) We obtain the minimax rates (as a function of n) for all $\mathcal{F}$ simultaneously.

b) Existing joint density results require estimating the covariate distribution, which we eliminate for conditional density estimation.

c) Our results allow for infinite dimensional and unbounded covariate spaces.

Main Results

Theorem (BFR21)

We define a new notion of entropy $\mathcal{H}$ such that

$$\mathcal{H}_n(\mathcal{F}, \varepsilon_n) \asymp n\varepsilon_n^2 \implies \mathcal{R}_n(\mathcal{F}) \asymp \varepsilon_n^2$$

for conditional density estimation.

Highlights of what this means

a) We obtain the minimax rates (as a function of n) for all $\mathcal{F}$ simultaneously.

b) Existing joint density results require estimating the covariate distribution, which we eliminate for conditional density estimation.

c) Our results allow for infinite dimensional and unbounded covariate spaces.

d) Our notion of entropy is data-dependent, which leads to an implementable algorithm.

Main Results

Theorem (BFR21)

We define a new notion of entropy $\mathcal{H}$ such that

$$\mathcal{H}_n(\mathcal{F}, \varepsilon_n) \asymp n\varepsilon_n^2 \implies \mathcal{R}_n(\mathcal{F}) \asymp \varepsilon_n^2$$

for conditional density estimation.

Highlights of what this means

a) We obtain the minimax rates (as a function of n) for all $\mathcal{F}$ simultaneously.

b) Existing joint density results require estimating the covariate distribution, which we eliminate for conditional density estimation.

c) Our results allow for infinite dimensional and unbounded covariate spaces.

d) Our notion of entropy is data-dependent, which leads to an implementable algorithm.