Minimax Rates for Conditional Density Estimation via Empirical Entropy

Blair Bilodeau¹
(joint work with Dylan J. Foster² and Daniel M. Roy¹)
April 23, 2021
Statistical Sciences Research Day

¹University of Toronto and Vector Institute
²Massachusetts Institute of Technology
Regression with Uncertainty

Observe:

\[ X_1: n \sim \mu^\otimes n \] and 
\[ Y_1: n = \text{Noise} \left[ f^* \left( X_1: n \right) \right]. \]

Goal:

Given a new \( X \sim \mu \), predict \( Y \).

Regression:

Approximate \( \mathbb{E}[Y|X] \) with \( \hat{f}: X \rightarrow Y \).
Observe: $X_{1:n} \sim \mu^\otimes n$ and $Y_{1:n} = \text{Noise}[f^*(X_{1:n})]$. 
Observe: $X_{1:n} \sim \mu \otimes n$ and $Y_{1:n} = \text{Noise}[f^*(X_{1:n})]$.

Goal: Given a new $X \sim \mu$, predict $Y$. 

Regression: Approximate $\mathbb{E}[Y | X]$ with $\hat{f}: X \rightarrow Y$. 

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

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 $Y_{1:n} = \text{Noise}\left[ f^\ast(X_{1:n}) \right]$.

**Goal:** Given a new $X \sim \mu$, predict $Y$.

**Regression:** Approximate $\mathbb{E}[Y \mid X]$ with $\hat{f}: \mathcal{X} \rightarrow \mathcal{Y}$.

Observe some data...
Regression with Uncertainty

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

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 $Y_{1:n} = \text{Noise}[f^*(X_{1:n})]$.

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 \( Y_{1:n} = \text{Noise} \left[ f^* (X_{1:n}) \right] \).

Goal: Given a new \( X \sim \mu \), predict \( Y \).

Regression: Approximate \( \mathbb{E}[Y \mid X] \) with \( \hat{f} : \mathcal{X} \rightarrow \mathcal{Y} \).

Idea: Approximate \( f^*(Y \mid X) \) with \( \hat{f} : \mathcal{X} \rightarrow \mathcal{M}(\mathcal{Y}) = \{ \text{densities on } \mathcal{Y} \} \).
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 \( \text{log loss} \) to achieve this:

\[
\ell(\hat{f}, (X,Y)) = -\log(\hat{f}(Y|X)).
\]

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...

We use log loss to achieve this:

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

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, $p$:...
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:

$$\ell(\hat{f}, (X,Y)) = -\log(\hat{f}(Y|X))$$

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:

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

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

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:

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

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

*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:

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

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

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

Consider the simple case of estimating the probability of rain, \(p\):

![Graph showing the log and square loss functions for rain probability](image-url)
Minimax Performance

If $f^*(Y|X)$ 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 $f^*: X \rightarrow M(Y)$ is in some known set: $f^* \in F$. This is the classical statistics assumption: i.i.d. data with a well-specified model.

Minimax Risk

$$R_n(F) = \inf \hat{f} \sup \mu \sup f^* \in F E_{X_1:n, Y_1:n} E_X KL(f^*(\cdot|X) \| \hat{f}_n(\cdot|X)).$$

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 $F$, 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.
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...

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

Minimax Risk
$R_n(F) = \inf \hat{f} \sup \mu \sup f^* \in F \mathbb{E}_{X_1:n, Y_1:n} \mathbb{E}_{X} \text{KL}(f^*(\cdot \mid X) \parallel \hat{f}_n(\cdot \mid X))$.

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 $F$, 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.
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 $f^* : \mathcal{X} \rightarrow \mathcal{M}(\mathcal{Y})$ 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 $f^* : \mathcal{X} \rightarrow \mathcal{M}(\mathcal{Y})$ 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 $f^*: \mathcal{X} \rightarrow \mathcal{M}(\mathcal{Y})$ 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

$$R_n(\mathcal{F}) = \inf_{\hat{f}} \sup_{\mu} \sup_{f^* \in \mathcal{F}} \mathbb{E}_{X_1:n, Y_1:n} \mathbb{E}_X \text{KL}(f^*(\cdot \mid X) \parallel \hat{f}_n(\cdot \mid X)).$$

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**

\[
\mathcal{R}_n(\mathcal{F}) = \inf_{\hat{f}} \sup_{\mu} \sup_{f^* \in \mathcal{F}} \mathbb{E}_{X_1:n, Y_1:n} \mathbb{E}_X \text{KL} \left( f^*(\cdot \mid X) \parallel \hat{f}_n(\cdot \mid X) \right).
\]

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 $f^* : \mathcal{X} \rightarrow \mathcal{M}(\mathcal{Y})$ 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

$$\mathcal{R}_n(\mathcal{F}) = \inf_{\hat{f}} \sup_\mu \sup_{f^* \in \mathcal{F}} \mathbb{E}_{X_1:n,Y_1:n} \mathbb{E}_X \text{KL}\left( f^*(\cdot \mid X) \parallel \hat{f}_n(\cdot \mid X) \right).$$

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 \( f^* : \mathcal{X} \rightarrow \mathcal{M}(\mathcal{Y}) \) 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

\[
\mathcal{R}_n(\mathcal{F}) = \inf_{\hat{f}} \sup_{\mu} \sup_{f^* \in \mathcal{F}} \mathbb{E}_{X_1:n, Y_1:n} \mathbb{E}_X \text{KL}\left( f^*(\cdot \mid X) \parallel \hat{f}_n(\cdot \mid X) \right).
\]

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)

Truncated Generalized Linear Model

\( Y | X \) follows Exponential family distribution truncated to \([-B, B] \).

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

\[ \mathbb{R}^n(\mathcal{F}) \lesssim \log(nB) \sqrt{n} . \]

VC-Type Classes

(Solved open problem from ALT 2021)

\( X \) arbitrary, \( Y | X \sim \text{Bernoulli}(p(X)) \), where \( p(X) = a + bI\{X \in c\} \) for some \( a, b > 0 \) and subset \( c \in \mathcal{C} \).

\[ \mathbb{R}^n(\mathcal{F}) \lesssim \text{VCdim}(\mathcal{C}) \log(n) \]

Nonparametric Conditional Densities

\( Y | X \) has an \( \alpha \)-Hölder continuous conditional density on \( X = [0, 1]^d \).

\[ \mathbb{R}^n(\mathcal{F}) \lesssim n^{-\alpha}(\alpha + d) \log(n) . \]

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 $\mathbb{R}^d$.

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

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

$$\mathcal{R}_n(\mathcal{F}) \lesssim \frac{\text{VCdim}(\mathcal{C}) \log(n)}{n}.$$
Examples of Rates (BFR21)

**Truncated Generalized Linear Model**

\(Y \ | \ X\) 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, \(Y \ | \ X \sim \text{Bernoulli}(p(X))\),

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

\[
\mathcal{R}_n(\mathcal{F}) \lesssim \frac{\text{VCdim}(\mathcal{C}) \log(n)}{n}.
\]

**Nonparametric Conditional Densities**

\(Y \ | \ X\) has an \(\alpha\)-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, \( Y \mid X \sim \text{Bernoulli}(p(X)) \),

where \( p(X) = a + b \mathbb{1}\{X \in c\} \) for some \( a, b > 0 \) and subset \( c \subseteq \mathcal{C} \).

\[
\mathcal{R}_n(\mathcal{F}) \lesssim \frac{\text{VCdim}(\mathcal{C}) \log(n)}{n}.
\]

**Nonparametric Conditional Densities**

\( Y \mid X \) has an \( \alpha \)-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, $Y \mid X \sim \text{Bernoulli}(p(X))$, where $p(X) = a + b \mathbb{1}\{X \in c\}$ for some $a, b > 0$ and subset $c \in \mathcal{C}$.

\[
\mathcal{R}_n(\mathcal{F}) \lesssim \frac{\text{VCdim}(\mathcal{C}) \log(n)}{n}.
\]

**Nonparametric Conditional Densities**

$Y \mid X$ has an $\alpha$-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 $\mathcal{F}$

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?
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?
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 $\mathcal{F}$**
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 $\mathcal{F}$. 
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

Nonparametric Regression

Minimax performance for regression with square loss is well-studied. Rakhlin et al. (2017) define entropy $H_{Sq, Loss}$ satisfying $H_{Sq, Loss}(F, \varepsilon_n) \approx n\varepsilon_n^2 \Rightarrow R_{Sq, Loss}(F) \approx \varepsilon_n^2$.

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

Problem #1: This entropy is for real-valued $F$, 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 $H_{Joint}$ satisfying $H_{Joint}(F, \varepsilon_n) \approx n\varepsilon_n^2 \Rightarrow R_{Joint}(F) \approx \varepsilon_n^2$.

Problem #2: We shouldn't have to estimate the marginal distribution on $X$!
Existing Results

Nonparametric Regression

Minimax performance for regression with square loss is well-studied. Rakhlin et al. (2017) define entropy $H_{\text{Sq}}$. Loss satisfying $H_{\text{Sq}} \left( F_{\epsilon_n} \right) \approx n \epsilon_n^2 \Rightarrow R_{\text{Sq}} \left( F_{\epsilon_n} \right) \approx \epsilon_n^2$. This type of relationship is also classically known; it appears in LeCam (1973).

Problem #1: This entropy is for real-valued $F$, 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 $H_{\text{Joint}}$ satisfying $H_{\text{Joint}} \left( F_{\epsilon_n} \right) \approx n \epsilon_n^2 \Rightarrow R_{\text{Joint}} \left( F_{\epsilon_n} \right) \approx \epsilon_n^2$. Problem #2: We shouldn't have to estimate the marginal distribution on $X$!
Nonparametric Regression
Minimax performance for regression with square loss is well-studied.
Nonparametric Regression

Minimax performance for regression with square loss is well-studied. Rakhlin et al. (2017) define entropy $\mathcal{H}^{\text{Sq.Loss}}$ satisfying

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

Minimax performance for regression with square loss is well-studied. Rakhlin et al. (2017) define entropy $\mathcal{H}^{\text{Sq.Loss}}$ satisfying

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

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

Minimax performance for regression with square loss is well-studied. Rakhlin et al. (2017) define entropy $H_{n}^{\text{Sq.Loss}}$ satisfying

$$H_{n}^{\text{Sq.Loss}}(\mathcal{F}, \varepsilon_{n}) \asymp n\varepsilon_{n}^2 \implies R_{n}^{\text{Sq.Loss}}(\mathcal{F}) \asymp \varepsilon_{n}^2.$$

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!
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

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

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

Problem #1: This entropy is for real-valued $F$, our regressors are function-valued!

Density Estimation
Nonparametric Regression

Minimax performance for regression with square loss is well-studied. Rakhlin et al. (2017) define entropy $\mathcal{H}^{\text{Sq.Loss}}$ satisfying

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

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 $H_{n}^{\text{Sq. Loss}}$ satisfying

$$H_{n}^{\text{Sq. Loss}} (\mathcal{F}, \varepsilon) \asymp n\varepsilon^2_n \implies R_{n}^{\text{Sq. Loss}} (\mathcal{F}) \asymp \varepsilon^2_n.$$  

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 $H_{n}^{\text{Joint}}$ satisfying

$$H_{n}^{\text{Joint}} (\mathcal{F}, \varepsilon) \asymp n\varepsilon^2_n \implies R_{n}^{\text{Joint}} (\mathcal{F}) \asymp \varepsilon^2_n.$$
Existing Results

Nonparametric Regression
Minimax performance for regression with square loss is well-studied. Rakhlin et al. (2017) define entropy $H^{\text{Sq.Loss}}$ satisfying
\[
H_n^{\text{Sq.Loss}}(\mathcal{F}, \varepsilon_n) \asymp n\varepsilon_n^2 \implies R_n^{\text{Sq.Loss}}(\mathcal{F}) \asymp \varepsilon_n^2.
\]
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 $H^{\text{Joint}}$ satisfying
\[
H_n^{\text{Joint}}(\mathcal{F}, \varepsilon_n) \asymp n\varepsilon_n^2 \implies R_n^{\text{Joint}}(\mathcal{F}) \asymp \varepsilon_n^2.
\]
Problem #2: We shouldn’t have to estimate the marginal distribution on $\mathcal{X}’$!
Nonparametric Regression

Minimax performance for regression with square loss is well-studied. Rakhlin et al. (2017) define entropy $H_{\text{Sq.Loss}}$ satisfying

$$H_{n \text{Sq.Loss}}(\mathcal{F}, \varepsilon_n) \asymp n\varepsilon_n^2 \implies R_n^{\text{Sq.Loss}}(\mathcal{F}) \asymp \varepsilon_n^2.$$  

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 $H_{\text{Joint}}$ satisfying

$$H_{n \text{Joint}}(\mathcal{F}, \varepsilon_n) \asymp n\varepsilon_n^2 \implies R_n^{\text{Joint}}(\mathcal{F}) \asymp \varepsilon_n^2.$$  

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

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

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*

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

*for conditional density estimation.*

**Highlights of what this means**

a) We obtain the minimax rates (as a function of $n$) for all $F$ 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

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

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.