Adapting to Failure of the I.I.D. Assumption

Blair Bilodeau

Based on joint work with:

Jeffrey Negrea, Daniel M. Roy, Nicolò Campolongo, and Francesco Orabona

February 17, 2022

Learning in the Presence of Strategic Behaviour Reading Group, Simons Institute

Motivation

Motivation

Assumptions are used to develop statistical methods and provide guarantees,

Motivation

Assumptions are used to develop statistical methods and provide guarantees,
leaving us susceptible to sharply degrading performance under failure of assumptions.

Motivation

Assumptions are used to develop statistical methods and provide guarantees,
leaving us susceptible to sharply degrading performance under failure of assumptions.

Want to act optimally without having to know how data arise.

Motivation

Assumptions are used to develop statistical methods and provide guarantees, leaving us susceptible to sharply degrading performance under failure of assumptions.

Want to act optimally without having to know how data arise.

Can we...

Motivation

Assumptions are used to develop statistical methods and provide guarantees,
leaving us susceptible to sharply degrading performance under failure of assumptions.

Want to act optimally without having to know how data arise.

Can we...

1. quantify the degree to which particular assumptions fail for a decision task?

Motivation

Assumptions are used to develop statistical methods and provide guarantees,
leaving us susceptible to sharply degrading performance under failure of assumptions.

Want to act optimally without having to know how data arise.

Can we...

1. quantify the degree to which particular assumptions fail for a decision task?
2. design **robust** decision methods that **adapt** to the failure of those assumptions?

Motivation

Assumptions are used to develop statistical methods and provide guarantees,
leaving us susceptible to sharply degrading performance under failure of assumptions.

Want to act optimally without having to know how data arise.

Can we...

1. quantify the degree to which particular assumptions fail for a decision task?
2. design **robust** decision methods that **adapt** to the failure of those assumptions?

Adapting means we simultaneously...

Motivation

Assumptions are used to develop statistical methods and provide guarantees, leaving us susceptible to sharply degrading performance under failure of assumptions.

Want to act optimally without having to know how data arise.

Can we...

1. quantify the degree to which particular assumptions fail for a decision task?
2. design **robust** decision methods that **adapt** to the failure of those assumptions?

Adapting means we simultaneously...

...benefit from assumptions when they hold,

Motivation

Assumptions are used to develop statistical methods and provide guarantees,
leaving us susceptible to sharply degrading performance under failure of assumptions.

Want to act optimally without having to know how data arise.

Can we...

1. quantify the degree to which particular assumptions fail for a decision task?
2. design **robust** decision methods that **adapt** to the failure of those assumptions?

Adapting means we simultaneously...

...benefit from assumptions when they hold,

...but still do "as well as possible" when they fail,

Motivation

Assumptions are used to develop statistical methods and provide guarantees,
leaving us susceptible to sharply degrading performance under failure of assumptions.

Want to act optimally without having to know how data arise.

Can we...

1. quantify the degree to which particular assumptions fail for a decision task?
2. design **robust** decision methods that **adapt** to the failure of those assumptions?

Adapting means we simultaneously...

...benefit from assumptions when they hold,

...but still do "as well as possible" when they fail,

...without knowing which case we are in.

The Role of Online Learning

The Role of Online Learning

This semester, many of you have convincingly motivated the study of sequential games:

e.g., GAN training, economic markets, adversarial corruptions, reinforcement learning.

The Role of Online Learning

This semester, many of you have convincingly motivated the study of sequential games:

e.g., GAN training, economic markets, adversarial corruptions, reinforcement learning.

Everyday, decisions are made using statistical methods tuned to “batch data”.

The Role of Online Learning

This semester, many of you have convincingly motivated the study of sequential games:

e.g., GAN training, economic markets, adversarial corruptions, reinforcement learning.

Everyday, decisions are made using statistical methods tuned to “batch data”.

Can sequential methods and analyses help us make more robust decisions?

Escaping the I.I.D. assumption

Escaping the I.I.D. assumption

I.I.D. is one of the most common assumptions made in statistics.

Escaping the I.I.D. assumption

I.I.D. is one of the most common assumptions made in statistics.

It is unverifiable, and outside of controlled settings, intuitively false.

Escaping the I.I.D. assumption

I.I.D. is one of the most common assumptions made in statistics.

It is unverifiable, and outside of controlled settings, intuitively false.

At the same time, it is often intuitively “good enough”.

Escaping the I.I.D. assumption

I.I.D. is one of the most common assumptions made in statistics.

It is unverifiable, and outside of controlled settings, intuitively false.

At the same time, it is often intuitively “good enough”.

Can we quantify this?

Escaping the I.I.D. assumption

I.I.D. is one of the most common assumptions made in statistics.

It is unverifiable, and outside of controlled settings, intuitively false.

At the same time, it is often intuitively “good enough”.

Can we quantify this?

How do we avoid specific dependence assumptions, and both:

Escaping the I.I.D. assumption

I.I.D. is one of the most common assumptions made in statistics.

It is unverifiable, and outside of controlled settings, intuitively false.

At the same time, it is often intuitively “good enough”.

Can we quantify this?

How do we avoid specific dependence assumptions, and both:

1. match improved performance of I.I.D. methods when data $\approx$ I.I.D.,

Escaping the I.I.D. assumption

I.I.D. is one of the most common assumptions made in statistics.

It is unverifiable, and outside of controlled settings, intuitively false.

At the same time, it is often intuitively "good enough".

Can we quantify this?

How do we avoid specific dependence assumptions, and both:

1. match improved performance of I.I.D. methods when data $\approx$ I.I.D.,
2. ensure methods perform "as well as possible" when I.I.D. fails?

Escaping the I.I.D. assumption

I.I.D. is one of the most common assumptions made in statistics.

It is unverifiable, and outside of controlled settings, intuitively false.

At the same time, it is often intuitively "good enough".

Can we quantify this?

How do we avoid specific dependence assumptions, and both:

1. match improved performance of I.I.D. methods when data $\approx$ I.I.D.,
2. ensure methods perform "as well as possible" when I.I.D. fails?

Without such assumptions:

Escaping the I.I.D. assumption

I.I.D. is one of the most common assumptions made in statistics.

It is unverifiable, and outside of controlled settings, intuitively false.

At the same time, it is often intuitively "good enough".

Can we quantify this?

How do we avoid specific dependence assumptions, and both:

1. match improved performance of I.I.D. methods when data $\approx$ I.I.D.,
2. ensure methods perform "as well as possible" when I.I.D. fails?

Without such assumptions:

1. Data can be fundamentally unpredictable.

Escaping the I.I.D. assumption

I.I.D. is one of the most common assumptions made in statistics.

It is unverifiable, and outside of controlled settings, intuitively false.

At the same time, it is often intuitively "good enough".

Can we quantify this?

How do we avoid specific dependence assumptions, and both:

1. match improved performance of I.I.D. methods when data $\approx$ I.I.D.,
2. ensure methods perform "as well as possible" when I.I.D. fails?

Without such assumptions:

1. Data can be fundamentally unpredictable.
2. *Absolute*, *point-in-time* notions of "good performance" may not be attainable....

   ...but good *relative*, *cumulative* performance might be possible.

Defining Performance

The I.I.D. assumption is intrinsic to static notions of performance:

Defining Performance

The I.I.D. assumption is intrinsic to static notions of performance:

e.g., MSE is the $\mathbb{E}[\text{loss}]$ of a learned model on a "new, test sample".

Defining Performance

The I.I.D. assumption is intrinsic to static notions of performance:

e.g., MSE is the $\mathbb{E}$[loss] of a learned model on a “new, test sample”.

How do we even define “good performance”...

...if we don’t make assumptions linking past and future data?

Defining Performance

The I.I.D. assumption is intrinsic to static notions of performance:

e.g., MSE is the $\mathbb{E}$[loss] of a learned model on a “new, test sample”.

How do we even define “good performance”...

...if we don’t make assumptions linking past and future data?

In order to move away from I.I.D. we leverage the *temporal structure* of the data...

Defining Performance

The I.I.D. assumption is intrinsic to static notions of performance:

e.g., MSE is the $\mathbb{E}$[loss] of a learned model on a "new, test sample".

How do we even define "good performance"...

...if we don't make assumptions linking past and future data?

In order to move away from I.I.D. we leverage the *temporal structure* of the data...

...and turn to *cumulative* measures of performance.

Defining Performance

The I.I.D. assumption is intrinsic to static notions of performance:

e.g., MSE is the $\mathbb{E}$[loss] of a learned model on a “new, test sample”.

How do we even define “good performance”...

...if we don’t make assumptions linking past and future data?

In order to move away from I.I.D. we leverage the *temporal structure* of the data...

...and turn to *cumulative* measures of performance.

Value Proposition

This adaptivity may seem like an impossible goal...

... but we show that it is possible.

This adaptivity may seem like an impossible goal...

... but we show that it is possible.

We need to focus on a concrete setting where we have some hope to achieve it.

This adaptivity may seem like an impossible goal...

... but we show that it is possible.

We need to focus on a concrete setting where we have some hope to achieve it.

Sequential Prediction with Expert Advice

- bounded loss functions
- sequential structure
- a relative & cumulative notion of performance (a.k.a. Regret)

Let's formalize the setting we're working in.

Sequential Prediction a.k.a. Online Learning

Sequential Prediction with Expert Advice

Sequential Prediction a.k.a. Online Learning

For rounds $t = 1, \ldots, T$:

Sequential Prediction with Expert Advice

Sequential Prediction a.k.a. Online Learning

For rounds $t = 1, \ldots, T$:

- Predict $\hat{y}(t) \in \hat{\mathcal{Y}}$ based on historical data before time t

Sequential Prediction with Expert Advice

Sequential Prediction a.k.a. Online Learning

For rounds $t = 1, \ldots, T$:

- Predict $\hat{y}(t) \in \hat{\mathcal{Y}}$ based on historical data before time t
- Observe response data $y(t) \in \mathcal{Y}$

Sequential Prediction a.k.a. Online Learning

For rounds $t = 1, \ldots, T$:

- Predict $\hat{y}(t) \in \hat{\mathcal{Y}}$ based on historical data before time t
- Observe response data $y(t) \in \mathcal{Y}$
- Incur loss $\ell(\hat{y}(t), y(t)) \in [0, 1]$

Sequential Prediction with Expert Advice

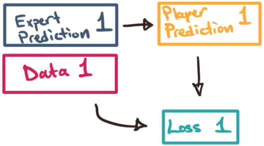
For rounds $t = 1, \ldots, T$:

- Receive $x(t) = (x_1(t), \ldots, x_N(t)) \subseteq \hat{\mathcal{Y}}$ expert predictions
- Predict $\hat{y}(t) \in \hat{\mathcal{Y}}$ based on historical data before time t and expert predictions
- Observe response data $y(t) \in \mathcal{Y}$
- Incur loss $\ell(\hat{y}(t), y(t)) \in [0, 1]$

Sequential Prediction with Expert Advice

For rounds $t = 1, \ldots, T$:

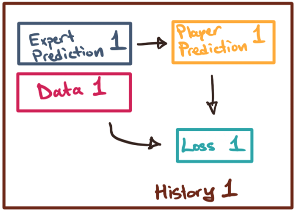
- Receive $x(t) = (x_1(t), \ldots, x_N(t)) \subseteq \hat{\mathcal{Y}}$ expert predictions
- Predict $\hat{y}(t) \in \hat{\mathcal{Y}}$ based on historical data before time t and expert predictions
- Observe response data $y(t) \in \mathcal{Y}$
- Incur loss $\ell(\hat{y}(t), y(t)) \in [0, 1]$



Sequential Prediction with Expert Advice

For rounds $t = 1, \ldots, T$:

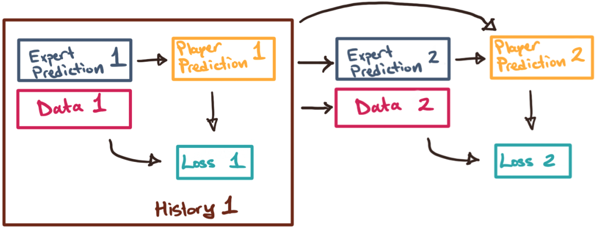
- Receive $x(t) = (x_1(t), \ldots, x_N(t)) \subseteq \hat{\mathcal{Y}}$ expert predictions
- Predict $\hat{y}(t) \in \hat{\mathcal{Y}}$ based on historical data before time t and expert predictions
- Observe response data $y(t) \in \mathcal{Y}$
- Incur loss $\ell(\hat{y}(t), y(t)) \in [0, 1]$



Sequential Prediction with Expert Advice

For rounds $t = 1, \ldots, T$:

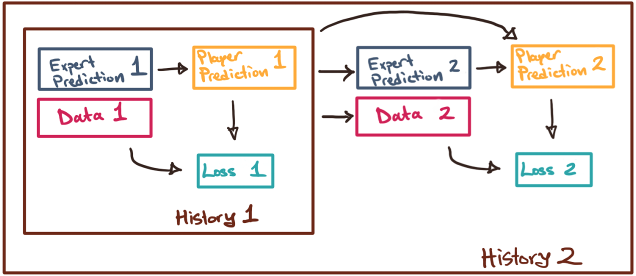
- Receive $x(t) = (x_1(t), \ldots, x_N(t)) \subseteq \hat{\mathcal{Y}}$ expert predictions
- Predict $\hat{y}(t) \in \hat{\mathcal{Y}}$ based on historical data before time t and expert predictions
- Observe response data $y(t) \in \mathcal{Y}$
- Incur loss $\ell(\hat{y}(t), y(t)) \in [0, 1]$



Sequential Prediction with Expert Advice

For rounds $t = 1, \ldots, T$:

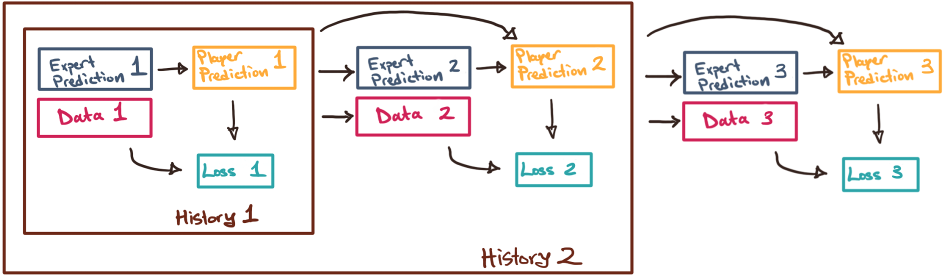
- Receive $x(t) = (x_1(t), \ldots, x_N(t)) \subseteq \hat{\mathcal{Y}}$ expert predictions
- Predict $\hat{y}(t) \in \hat{\mathcal{Y}}$ based on historical data before time t and expert predictions
- Observe response data $y(t) \in \mathcal{Y}$
- Incur loss $\ell(\hat{y}(t), y(t)) \in [0, 1]$



Sequential Prediction with Expert Advice

For rounds $t = 1, \ldots, T$:

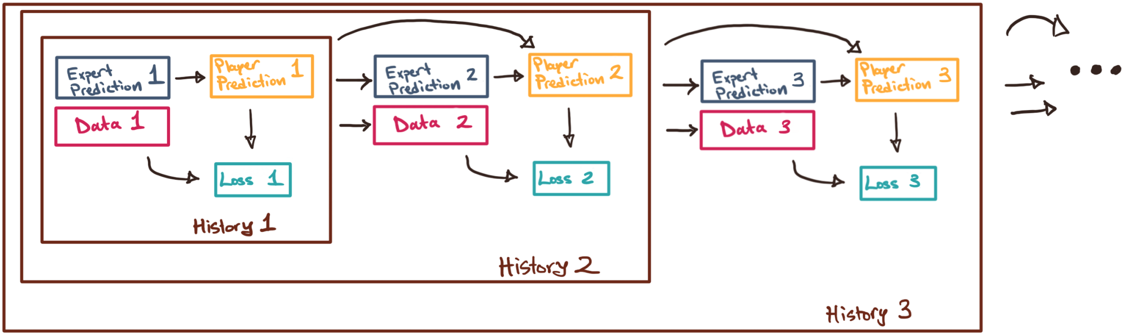
- Receive $x(t) = (x_1(t), \ldots, x_N(t)) \subseteq \hat{\mathcal{Y}}$ expert predictions
- Predict $\hat{y}(t) \in \hat{\mathcal{Y}}$ based on historical data before time t and expert predictions
- Observe response data $y(t) \in \mathcal{Y}$
- Incur loss $\ell(\hat{y}(t), y(t)) \in [0, 1]$



Sequential Prediction with Expert Advice

For rounds $t = 1, \ldots, T$:

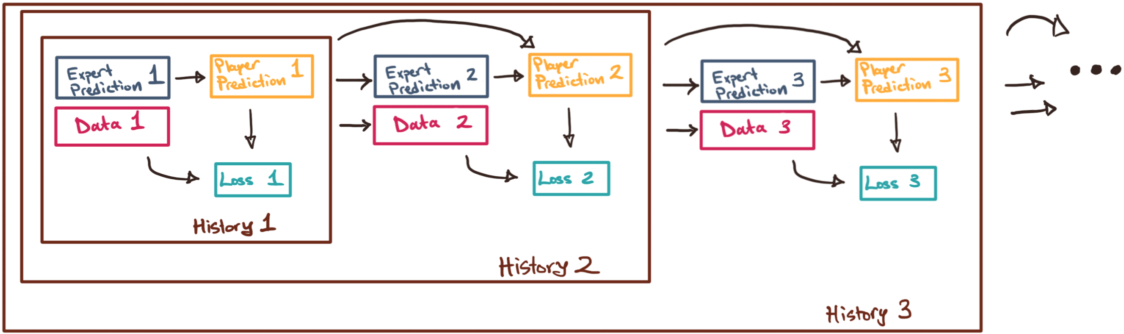
- Receive $x(t) = (x_1(t), \ldots, x_N(t)) \subseteq \hat{\mathcal{Y}}$ expert predictions
- Predict $\hat{y}(t) \in \hat{\mathcal{Y}}$ based on historical data before time t and expert predictions
- Observe response data $y(t) \in \mathcal{Y}$
- Incur loss $\ell(\hat{y}(t), y(t)) \in [0, 1]$



Sequential Prediction with Expert Advice

For rounds $t = 1, \ldots, T$:

- Receive $x(t) = (x_1(t), \ldots, x_N(t)) \subseteq \hat{\mathcal{Y}}$ expert predictions
- Predict $\hat{y}(t) \in \hat{\mathcal{Y}}$ based on historical data before time t and expert predictions
- Observe response data $y(t) \in \mathcal{Y}$
- Incur loss $\ell(\hat{y}(t), y(t)) \in [0, 1]$



Measuring Performance

Even under I.I.D. assumption, if model not well-specified, we measure performance relatively

$$\textbf{Excess risk:}\quad R(T) = \mathbb{E}\,\ell(\hat{\theta}_T, y_{T+1}) - \min_{\theta} \mathbb{E}\,\ell(\theta, y_{T+1})$$

Measuring Performance

Even under I.I.D. assumption, if model not well-specified, we measure performance relatively

$$\textbf{Excess risk:}\quad R(T) = \mathbb{E}\,\ell(\hat{\theta}_T, y_{T+1}) - \min_{\theta} \mathbb{E}\,\ell(\theta, y_{T+1})$$

Without I.I.D. assumption, we cannot look at just the next instance.

Measuring Performance

Even under I.I.D. assumption, if model not well-specified, we measure performance relatively

$$\textbf{Excess risk:} \quad R(T) = \mathbb{E}\,\ell(\hat{\theta}_T, y_{T+1}) - \min_{\theta} \mathbb{E}\,\ell(\theta, y_{T+1})$$

Without I.I.D. assumption, we cannot look at just the next instance.

The measure of the player's performance is...

Measuring Performance

Even under I.I.D. assumption, if model not well-specified, we measure performance relatively

$$\textbf{Excess risk:} \quad R(T) = \mathbb{E}\,\ell(\hat{\theta}_T, y_{T+1}) - \min_{\theta} \mathbb{E}\,\ell(\theta, y_{T+1})$$

Without I.I.D. assumption, we cannot look at just the next instance.

The measure of the player's performance is...

- Relative to the class of N reference *experts*, in hindsight;

Measuring Performance

Even under I.I.D. assumption, if model not well-specified, we measure performance relatively

$$\textbf{Excess risk:} \quad R(T) = \mathbb{E}\,\ell(\hat{\theta}_T, y_{T+1}) - \min_{\theta} \mathbb{E}\,\ell(\theta, y_{T+1})$$

Without I.I.D. assumption, we cannot look at just the next instance.

The measure of the player's performance is...

- Relative to the class of N reference *experts*, in hindsight;
- Given by the excess cumulative loss of the player over the best expert;

Measuring Performance

Even under I.I.D. assumption, if model not well-specified, we measure performance relatively

$$\textbf{Excess risk:}\quad R(T) = \mathbb{E}\,\ell(\hat{\theta}_T, y_{T+1}) - \min_{\theta} \mathbb{E}\,\ell(\theta, y_{T+1})$$

Without I.I.D. assumption, we cannot look at just the next instance.

The measure of the player's performance is...

- Relative to the class of N reference *experts*, in hindsight;
- Given by the excess cumulative loss of the player over the best expert;

$$\textbf{Expected Regret:}\quad R(T) = \mathbb{E}\left[\sum_{t=1}^{T} \ell(\hat{y}(t), y(t)) - \min_{i\in[N]} \sum_{t=1}^{T} \ell(x_i(t), y(t))\right]$$

Measuring Performance

Even under I.I.D. assumption, if model not well-specified, we measure performance relatively

$$\textbf{Excess risk:} \quad R(T) = \mathbb{E}\,\ell(\hat{\theta}_T, y_{T+1}) - \min_{\theta} \mathbb{E}\,\ell(\theta, y_{T+1})$$

Without I.I.D. assumption, we cannot look at just the next instance.

The measure of the player's performance is...

- Relative to the class of N reference *experts*, in hindsight;
- Given by the excess cumulative loss of the player over the best expert;

$$\textbf{Expected Regret:} \quad R(T) = \mathbb{E}\left[\sum_{t=1}^{T} \ell(\hat{y}(t), y(t)) - \min_{i \in [N]} \sum_{t=1}^{T} \ell(x_i(t), y(t))\right]$$

Where the $\mathbb{E}$ is taken with respect to the randomness in the player's and experts' predictions, and the data-generating mechanism for $(y(t))_{t \in \mathbb{N}}$.

Measuring Performance

Even under I.I.D. assumption, if model not well-specified, we measure performance relatively

$$\textbf{Excess risk:} \quad R(T) = \mathbb{E}\,\ell(\hat{\theta}_T, y_{T+1}) - \min_{\theta} \mathbb{E}\,\ell(\theta, y_{T+1})$$

Without I.I.D. assumption, we cannot look at just the next instance.

The measure of the player's performance is...

- Relative to the class of N reference *experts*, in hindsight;
- Given by the excess cumulative loss of the player over the best expert;

$$\textbf{Expected Regret:} \quad R(T) = \mathbb{E}\left[\sum_{t=1}^{T} \ell(\hat{y}(t), y(t)) - \min_{i\in[N]} \sum_{t=1}^{T} \ell(x_i(t), y(t))\right]$$

Where the $\mathbb{E}$ is taken with respect to the randomness in the player's and experts' predictions, and the data-generating mechanism for $(y(t))_{t\in\mathbb{N}}$.

(The $\mathbb{E}$ may be under a complicated, non-I.I.D. measure.)

Cut to the chase: What do we achieve in this setting we just described?

The Punchline: High-Level Overview of Results

The Punchline: High-Level Overview of Results

We define a spectrum of adversaries with I.I.D. at one end and adversarial at the other.

The Punchline: High-Level Overview of Results

We define a spectrum of adversaries with I.I.D. at one end and adversarial at the other.

The Hedge algorithm was recently shown to be optimal at these two endpoints [MG19].

The Punchline: High-Level Overview of Results

We define a spectrum of adversaries with I.I.D. at one end and adversarial at the other.

The Hedge algorithm was recently shown to be optimal at these two endpoints [MG19].
We show Hedge is suboptimal strictly in between.

The Punchline: High-Level Overview of Results

We define a spectrum of adversaries with I.I.D. at one end and adversarial at the other.

The Hedge algorithm was recently shown to be optimal at these two endpoints [MG19].
We show Hedge is suboptimal strictly in between.

Theorem

Hedge is not simultaneously minimax optimal at all settings between I.I.D. and adversarial.
With standard tuning, as soon as data is not I.I.D., Hedge can incur worst-case regret.

The Punchline: High-Level Overview of Results

We define a spectrum of adversaries with I.I.D. at one end and adversarial at the other.

The Hedge algorithm was recently shown to be optimal at these two endpoints [MG19].
We show Hedge is suboptimal strictly in between.

Theorem

Hedge is not simultaneously minimax optimal at all settings between I.I.D. and adversarial.
With standard tuning, as soon as data is not I.I.D., Hedge can incur worst-case regret.

We provide a new algorithm that achieves the minimax rate in all settings...

The Punchline: High-Level Overview of Results

We define a spectrum of adversaries with I.I.D. at one end and adversarial at the other.

The Hedge algorithm was recently shown to be optimal at these two endpoints [MG19].
We show Hedge is suboptimal strictly in between.

Theorem

Hedge is not simultaneously minimax optimal at all settings between I.I.D. and adversarial.
With standard tuning, as soon as data is not I.I.D., Hedge can incur worst-case regret.

We provide a new algorithm that achieves the minimax rate in all settings...
...without knowledge of which setting prevails!

The Punchline: High-Level Overview of Results

We define a spectrum of adversaries with I.I.D. at one end and adversarial at the other.

The Hedge algorithm was recently shown to be optimal at these two endpoints [MG19].
We show Hedge is suboptimal strictly in between.

Theorem

Hedge is not simultaneously minimax optimal at all settings between I.I.D. and adversarial.
With standard tuning, as soon as data is not I.I.D., Hedge can incur worst-case regret.

We provide a new algorithm that achieves the minimax rate in all settings...
...without knowledge of which setting prevails!

Theorem

There is an adaptively minimax optimal algorithm: FTRL-CARL.

The Punchline: High-Level Overview of Results

We define a spectrum of adversaries with I.I.D. at one end and adversarial at the other.

The Hedge algorithm was recently shown to be optimal at these two endpoints [MG19].
We show Hedge is suboptimal strictly in between.

Theorem

Hedge is not simultaneously minimax optimal at all settings between I.I.D. and adversarial.
With standard tuning, as soon as data is not I.I.D., Hedge can incur worst-case regret.

We provide a new algorithm that achieves the minimax rate in all settings...
...without knowledge of which setting prevails!

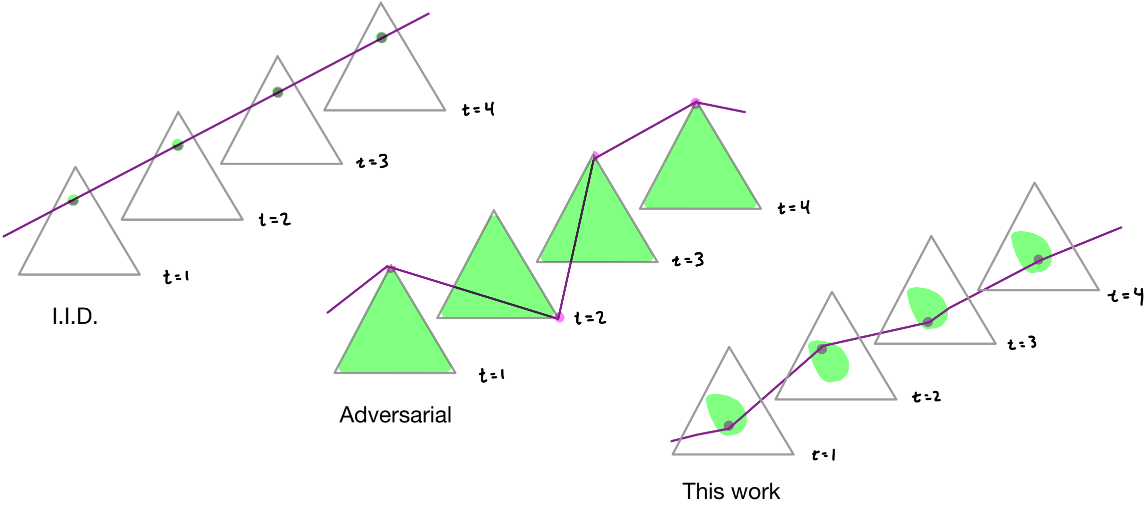
Theorem

There is an adaptively minimax optimal algorithm: FTRL-CARL.

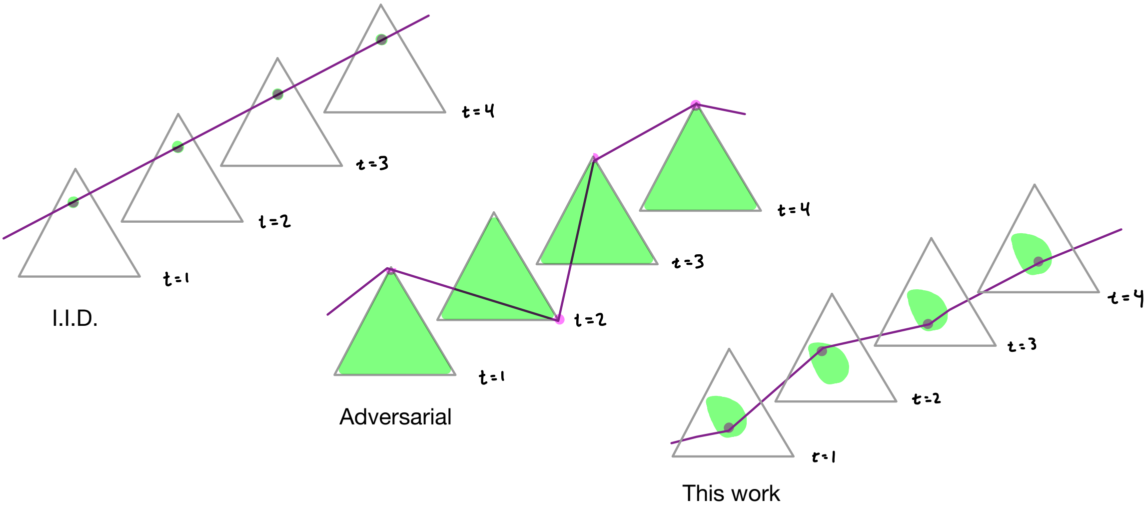
FTRL-CARL is like Hedge with the regularizer chosen to optimize a local-norm bound.

Now that we know what we achieve, let's formally define our constraint framework.

Beyond I.I.D. and Adversarial



Beyond I.I.D. and Adversarial



Semi-Adversarial Framework: Time-Homogeneous Convex Constraints

Formal Framework

- Fix a convex set of distributions $\mathcal{D} \subseteq \mathcal{M}(\hat{\mathcal{Y}}^N \times \mathcal{Y})$.

Semi-Adversarial Framework: Time-Homogeneous Convex Constraints

Formal Framework

- Fix a convex set of distributions $\mathcal{D} \subseteq \mathcal{M}(\hat{\mathcal{Y}}^N \times \mathcal{Y})$.
- $(x(t), y(t))$ drawn from an element of $\mathcal{D}$ given the history prior to t.
    - The choice of distribution is made based on outcomes of the previous rounds.

Semi-Adversarial Framework: Time-Homogeneous Convex Constraints

Formal Framework

- Fix a convex set of distributions $\mathcal{D} \subseteq \mathcal{M}(\hat{\mathcal{Y}}^N \times \mathcal{Y})$.
- $(x(t), y(t))$ drawn from an element of $\mathcal{D}$ given the history prior to t.
  - The choice of distribution is made based on outcomes of the previous rounds.

More Details

- Time-Homogeneous: $\mathcal{D}$ does not depend on t.

Semi-Adversarial Framework: Time-Homogeneous Convex Constraints

Formal Framework

- Fix a convex set of distributions $\mathcal{D} \subseteq \mathcal{M}(\hat{\mathcal{Y}}^N \times \mathcal{Y})$.
- $(x(t), y(t))$ drawn from an element of $\mathcal{D}$ given the history prior to t.
  - The choice of distribution is made based on outcomes of the previous rounds.

More Details

- Time-Homogeneous: $\mathcal{D}$ does not depend on t.
- Convex: environment can flip a coin to select between basic elements of $\mathcal{D}$.

Semi-Adversarial Framework: Time-Homogeneous Convex Constraints

Formal Framework

- Fix a convex set of distributions $\mathcal{D} \subseteq \mathcal{M}(\hat{\mathcal{Y}}^N \times \mathcal{Y})$.
- $(x(t), y(t))$ drawn from an element of $\mathcal{D}$ given the history prior to t.
  - The choice of distribution is made based on outcomes of the previous rounds.

More Details

- Time-Homogeneous: $\mathcal{D}$ does not depend on t.
- Convex: environment can flip a coin to select between basic elements of $\mathcal{D}$.
- Environment may aim to maximize regret subject to the constraint.

How do we study regret bounds for this constraint framework?

Adaptively Minimax Optimal Algorithms

Algorithms should be robust to a spectrum of data-generating mechanisms.

Adaptively Minimax Optimal Algorithms

Algorithms should be robust to a spectrum of data-generating mechanisms.

Definition

An algorithm is **adaptively minimax optimal** for a spectrum of settings if:

- it achieves the minimax optimal performance in each setting; and
- it does not require knowledge of the true setting in advance.

Adaptively Minimax Optimal Algorithms

Algorithms should be robust to a spectrum of data-generating mechanisms.

Definition

An algorithm is **adaptively minimax optimal** for a spectrum of settings if:

- it achieves the minimax optimal performance in each setting; and
- it does not require knowledge of the true setting in advance.

Each **setting** here is a collection of environments we might face.

Adaptively Minimax Optimal Algorithms

Algorithms should be robust to a spectrum of data-generating mechanisms.

Definition

An algorithm is **adaptively minimax optimal** for a spectrum of settings if:

- it achieves the minimax optimal performance in each setting; and
- it does not require knowledge of the true setting in advance.

Each **setting** here is a collection of environments we might face.

Minimax optimal $\equiv \underbrace{\text{best possible}}\ \underbrace{\text{worst-case}}$ performance.

Adaptively Minimax Optimal Algorithms

Algorithms should be robust to a spectrum of data-generating mechanisms.

Definition

An algorithm is **adaptively minimax optimal** for a spectrum of settings if:

- it achieves the minimax optimal performance in each setting; and
- it does not require knowledge of the true setting in advance.

Each **setting** here is a collection of environments we might face.

Minimax optimal $\equiv \underbrace{\text{best possible}}_{\text{players policy}}\ \underbrace{\text{worst-case}}$ performance.

Adaptively Minimax Optimal Algorithms

Algorithms should be robust to a spectrum of data-generating mechanisms.

Definition

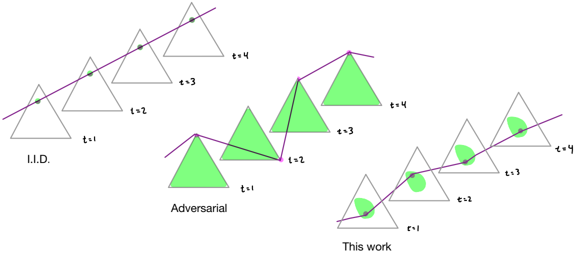
An algorithm is **adaptively minimax optimal** for a spectrum of settings if:

- it achieves the minimax optimal performance in each setting; and
- it does not require knowledge of the true setting in advance.

Each **setting** here is a collection of environments we might face.

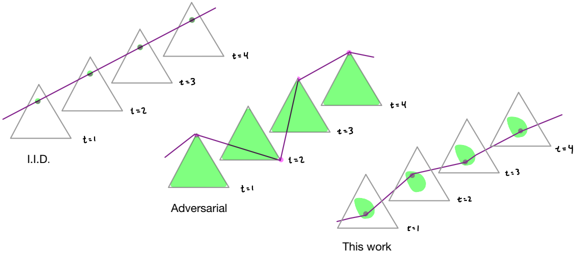
Minimax optimal $\equiv \underbrace{\text{best possible}}_{\text{players policy}}\ \underbrace{\text{worst-case}}_{\text{environment}}$ performance.

Adapting to tiers of problem hardness



Adaptivity is a well established notion in statistics, especially nonparametrics.

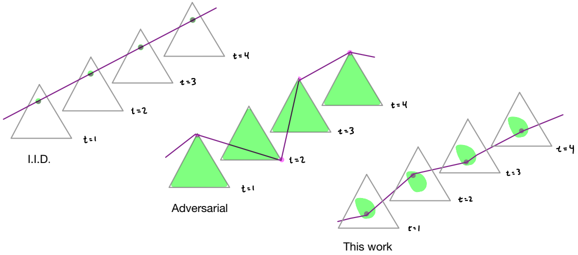
Adapting to tiers of problem hardness



Adaptivity is a well established notion in statistics, especially nonparametrics.

E.g. adapting to smoothness in density estimation.

Adapting to tiers of problem hardness

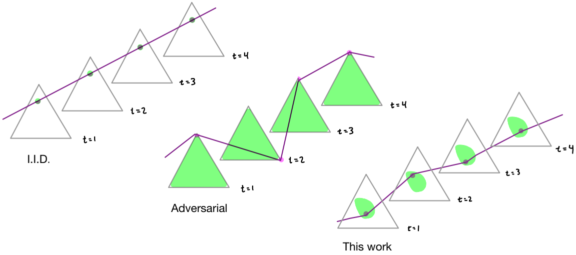


Adaptivity is a well established notion in statistics, especially nonparametrics.

E.g. adapting to smoothness in density estimation.

Do not adapt to a constraint set:

Adapting to tiers of problem hardness



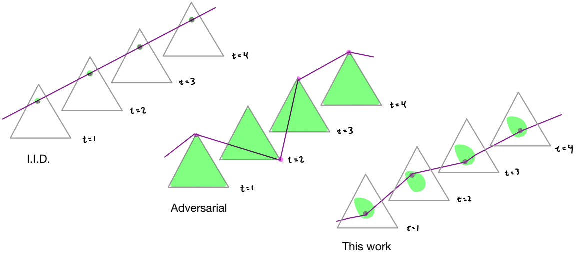
Adaptivity is a well established notion in statistics, especially nonparametrics.

E.g. adapting to smoothness in density estimation.

Do not adapt to a constraint set:

like trying to do as well as if you knew the true density in advance.

Adapting to tiers of problem hardness



Adaptivity is a well established notion in statistics, especially nonparametrics.

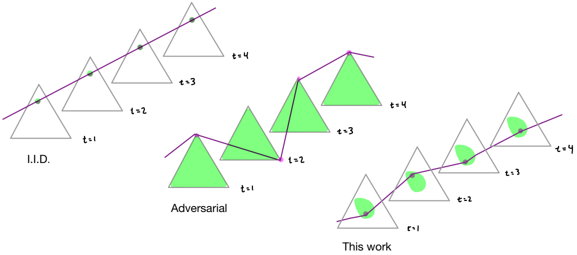
E.g. adapting to smoothness in density estimation.

Do not adapt to a constraint set:

like trying to do as well as if you knew the true density in advance.

We aim to adapt to **a notion of hardness** for the constraint set.

Adapting to tiers of problem hardness



Adaptivity is a well established notion in statistics, especially nonparametrics.

E.g. adapting to smoothness in density estimation.

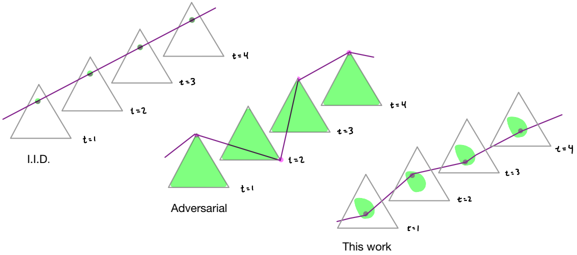
Do not adapt to a constraint set:

like trying to do as well as if you knew the true density in advance.

We aim to adapt to **a notion of hardness** for the constraint set.

do as well as if we know the smoothness level in advance.

Adapting to tiers of problem hardness



Adaptivity is a well established notion in statistics, especially nonparametrics.

E.g. adapting to smoothness in density estimation.

Do not adapt to a constraint set:

like trying to do as well as if you knew the true density in advance.

We aim to adapt to **a notion of hardness** for the constraint set.

do as well as if we know the smoothness level in advance.

What governs the hardness of prediction in a semi-adversarial environment?

Existing Work: Optimality in the I.I.D. and Adversarial Regimes

Existing Work: Optimality in the I.I.D. and Adversarial Regimes

I.I.D.-with-a-gap data

Experts and data are I.I.D. realizations independent of how the player behaves.

There is an expert whose mean loss is Δ smaller than the others.

Existing Work: Optimality in the I.I.D. and Adversarial Regimes

I.I.D.-with-a-gap data

Experts and data are I.I.D. realizations independent of how the player behaves.

There is an expert whose mean loss is Δ smaller than the others.

Minimax $R(T) \asymp (\log N)/\Delta$

Existing Work: Optimality in the I.I.D. and Adversarial Regimes

I.I.D.-with-a-gap data

Experts and data are I.I.D. realizations independent of how the player behaves.

There is an expert whose mean loss is Δ smaller than the others.

Minimax $R(T) \asymp (\log N)/\Delta$

No assumptions on features or responses (adversarial)

Compete against expert predictions and data that maximize $R(T)$.

Existing Work: Optimality in the I.I.D. and Adversarial Regimes

I.I.D.-with-a-gap data

Experts and data are I.I.D. realizations independent of how the player behaves.

There is an expert whose mean loss is Δ smaller than the others.

Minimax $R(T) \asymp (\log N)/\Delta$

No assumptions on features or responses (adversarial)

Compete against expert predictions and data that maximize $R(T)$.

Minimax $R(T) \asymp \sqrt{T \log N}$

Existing Work: Optimality in the I.I.D. and Adversarial Regimes

I.I.D.-with-a-gap data

Experts and data are I.I.D. realizations independent of how the player behaves.

There is an expert whose mean loss is Δ smaller than the others.

Minimax $R(T) \asymp (\log N)/\Delta$

No assumptions on features or responses (adversarial)

Compete against expert predictions and data that maximize $R(T)$.

Minimax $R(T) \asymp \sqrt{T \log N}$

At the endpoints, $\log N$ and Δ govern regret.

Our Work: Constraint-Characterizing Quantities

We want to characterize the hardness of the constraint using quantities that:

Our Work: Constraint-Characterizing Quantities

We want to characterize the hardness of the constraint using quantities that:

- differentiate whether the data is “easy” or not, independent of algorithms;

Our Work: Constraint-Characterizing Quantities

We want to characterize the hardness of the constraint using quantities that:

- differentiate whether the data is "easy" or not, independent of algorithms;
- yield matching lower and upper bounds on regret.

Our Work: Constraint-Characterizing Quantities

We want to characterize the hardness of the constraint using quantities that:

- differentiate whether the data is "easy" or not, independent of algorithms;
- yield matching lower and upper bounds on regret.

Effective Experts

Our Work: Constraint-Characterizing Quantities

We want to characterize the hardness of the constraint using quantities that:

- differentiate whether the data is “easy” or not, independent of algorithms;
- yield matching lower and upper bounds on regret.

Effective Experts

Analogous to the single best expert in the I.I.D.-with-a-gap setting.

Our Work: Constraint-Characterizing Quantities

We want to characterize the hardness of the constraint using quantities that:

- differentiate whether the data is "easy" or not, independent of algorithms;
- yield matching lower and upper bounds on regret.

Effective Experts

Analogous to the single best expert in the I.I.D.-with-a-gap setting.

$$\mathcal{I}_0 = \{\text{experts that are optimal in } \mathbb{E} \text{ for some } \mu \in \mathcal{D}\} \subseteq [N]$$
$$N_0 = |\mathcal{I}_0|$$

Our Work: Constraint-Characterizing Quantities

We want to characterize the hardness of the constraint using quantities that:

- differentiate whether the data is "easy" or not, independent of algorithms;
- yield matching lower and upper bounds on regret.

Effective Experts

Analogous to the single best expert in the I.I.D.-with-a-gap setting.

$$\mathcal{I}_0 = \{\text{experts that are optimal in } \mathbb{E} \text{ for some } \mu \in \mathcal{D}\} \subseteq [N]$$

$$N_0 = |\mathcal{I}_0|$$

Effective Stochastic Gap

Our Work: Constraint-Characterizing Quantities

We want to characterize the hardness of the constraint using quantities that:

- differentiate whether the data is "easy" or not, independent of algorithms;
- yield matching lower and upper bounds on regret.

Effective Experts

Analogous to the single best expert in the I.I.D.-with-a-gap setting.

$$\mathcal{I}_0 = \{\text{experts that are optimal in } \mathbb{E} \text{ for some } \mu \in \mathcal{D}\} \subseteq [N]$$

$$N_0 = |\mathcal{I}_0|$$

Effective Stochastic Gap

Analogous to the gap in the I.I.D.-with-a-gap setting.

Our Work: Constraint-Characterizing Quantities

We want to characterize the hardness of the constraint using quantities that:

- differentiate whether the data is "easy" or not, independent of algorithms;
- yield matching lower and upper bounds on regret.

Effective Experts

Analogous to the single best expert in the I.I.D.-with-a-gap setting.

$$\mathcal{I}_0 = \{\text{experts that are optimal in } \mathbb{E} \text{ for some } \mu \in \mathcal{D}\} \subseteq [N]$$

$$N_0 = |\mathcal{I}_0|$$

Effective Stochastic Gap

Analogous to the gap in the I.I.D.-with-a-gap setting.

$$\Delta_0 = \inf_{\mu \in \mathcal{D}} \{\mu\text{-expected difference in loss of best expert and best expert not in } \mathcal{I}_0\}$$

Main Result

Motivating Intuition

Main Result

Motivating Intuition

- In the adversarial case the minimax optimal regret is $\Theta(\sqrt{T \log N})$

Main Result

Motivating Intuition

- In the adversarial case the minimax optimal regret is $\Theta(\sqrt{T \log N})$
- If we know only N_0 of the experts can ever be "the best", and which ones, ...

Main Result

Motivating Intuition

- In the adversarial case the minimax optimal regret is $\Theta(\sqrt{T \log N})$
- If we know only N_0 of the experts can ever be "the best", and which ones, ...
  - we could restrict an adversarially optimal algorithm to the "best experts"

Main Result

Motivating Intuition

- In the adversarial case the minimax optimal regret is $\Theta(\sqrt{T \log N})$
- If we know only N_0 of the experts can ever be "the best", and which ones, ...
  - we could restrict an adversarially optimal algorithm to the "best experts"
  - so we might strive to have *long run* regret $\Theta(\sqrt{T \log N_0})$ in (T, N_0)

Main Result

Motivating Intuition

- In the adversarial case the minimax optimal regret is $\Theta(\sqrt{T \log N})$
- If we know only N_0 of the experts can ever be "the best", and which ones, ...
  - we could restrict an adversarially optimal algorithm to the "best experts"
  - so we might strive to have *long run* regret $\Theta(\sqrt{T \log N_0})$ in (T, N_0)
- We also need to learn *which* experts are better than the rest by Δ_0

Main Result

Motivating Intuition

- In the adversarial case the minimax optimal regret is $\Theta(\sqrt{T \log N})$
- If we know only N_0 of the experts can ever be "the best", and which ones, ...
  - we could restrict an adversarially optimal algorithm to the "best experts"
  - so we might strive to have *long run* regret $\Theta(\sqrt{T \log N_0})$ in (T, N_0)
- We also need to learn *which* experts are better than the rest by Δ_0
  - at best we could hope for a *fixed* regret $\Theta((\log N)/\Delta_0)$ from the I.I.D. case.

Main Result

Motivating Intuition

- In the adversarial case the minimax optimal regret is $\Theta(\sqrt{T \log N})$
- If we know only N_0 of the experts can ever be "the best", and which ones, ...
  - we could restrict an adversarially optimal algorithm to the "best experts"
  - so we might strive to have *long run* regret $\Theta(\sqrt{T \log N_0})$ in (T, N_0)
- We also need to learn *which* experts are better than the rest by Δ_0
  - at best we could hope for a *fixed* regret $\Theta((\log N)/\Delta_0)$ from the I.I.D. case.

Theorem

FTRL-CARL is adaptively minimax optimal, and achieves

$$\mathbb{E}R(T) \asymp \underbrace{\sqrt{T \log N_0}}_{\text{long run cost}} + \underbrace{(\log N)/\Delta_0}_{\text{fixed cost}}$$

Let's understand N_0 and Δ_0 using some examples.

Constraint Example

$\mathcal{I}_0 = \{\text{experts that are optimal for some } \mu \in \mathcal{D}\}$ $\qquad N_0 = |\mathcal{I}_0|$

$$\Delta_0 = \inf_{\mu \in \mathcal{D}} \{\mu\text{-expected difference in loss of best expert and best expert not in } \mathcal{I}_0\}$$

Constraint Example

$\mathcal{I}_0 = \{\text{experts that are optimal for some } \mu \in \mathcal{D}\}$ $\qquad N_0 = |\mathcal{I}_0|$

$$\Delta_0 = \inf_{\mu \in \mathcal{D}} \{\mu\text{-expected difference in loss of best expert and best expert not in } \mathcal{I}_0\}$$

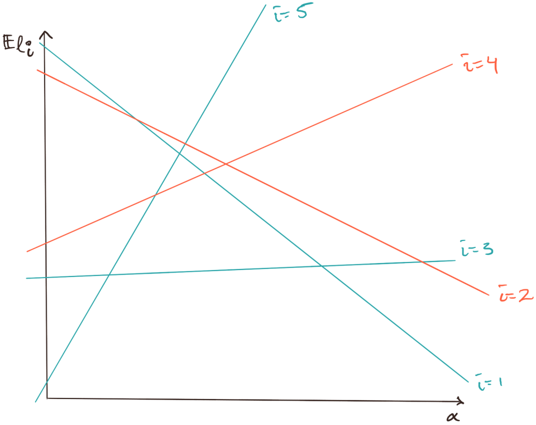
Setting: the means for each expert are jointly defined by a parameter α,
$N = 5$, $\mathcal{I}_0 = \{1, 3, 5\}$, $N_0 = 3$.

Constraint Example

$\mathcal{I}_0 = \{\text{experts that are optimal for some } \mu \in \mathcal{D}\}$ $\qquad N_0 = |\mathcal{I}_0|$

$$\Delta_0 = \inf_{\mu \in \mathcal{D}} \{\mu\text{-expected difference in loss of best expert and best expert not in } \mathcal{I}_0\}$$

Setting: the means for each expert are jointly defined by a parameter α,
$N = 5$, $\mathcal{I}_0 = \{1, 3, 5\}$, $N_0 = 3$.

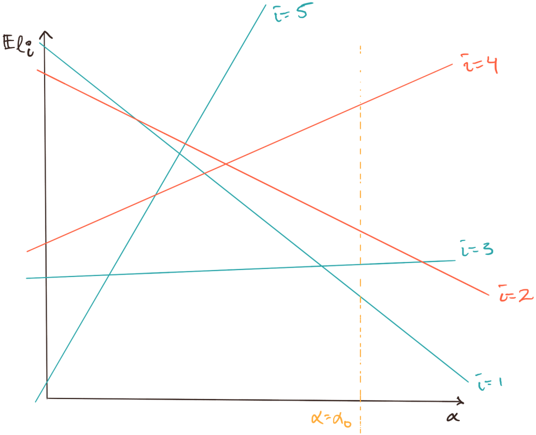


Constraint Example

$\mathcal{I}_0 = \{\text{experts that are optimal for some } \mu \in \mathcal{D}\}$ $\qquad N_0 = |\mathcal{I}_0|$

$$\Delta_0 = \inf_{\mu \in \mathcal{D}} \{\mu\text{-expected difference in loss of best expert and best expert not in } \mathcal{I}_0\}$$

Setting: the means for each expert are jointly defined by a parameter α,
$N = 5$, $\mathcal{I}_0 = \{1, 3, 5\}$, $N_0 = 3$.

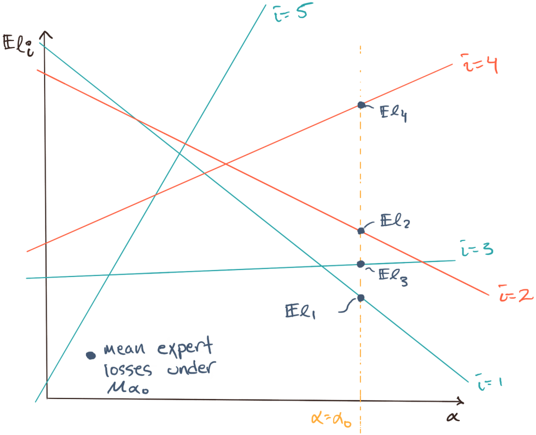


Constraint Example

$\mathcal{I}_0 = \{\text{experts that are optimal for some } \mu \in \mathcal{D}\}$ $\qquad N_0 = |\mathcal{I}_0|$

$$\Delta_0 = \inf_{\mu \in \mathcal{D}} \{\mu\text{-expected difference in loss of best expert and best expert not in } \mathcal{I}_0\}$$

Setting: the means for each expert are jointly defined by a parameter α,
$N = 5$, $\mathcal{I}_0 = \{1, 3, 5\}$, $N_0 = 3$.

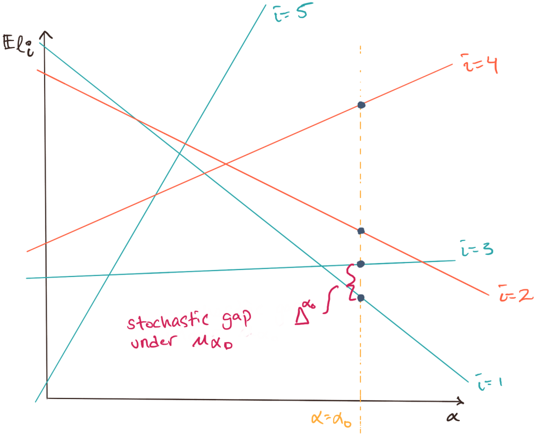


Constraint Example

$$\mathcal{I}_0 = \{\text{experts that are optimal for some } \mu \in \mathcal{D}\} \qquad N_0 = |\mathcal{I}_0|$$

$$\Delta_0 = \inf_{\mu \in \mathcal{D}} \{\mu\text{-expected difference in loss of best expert and best expert not in } \mathcal{I}_0\}$$

Setting: the means for each expert are jointly defined by a parameter α,
$N = 5$, $\mathcal{I}_0 = \{1, 3, 5\}$, $N_0 = 3$.

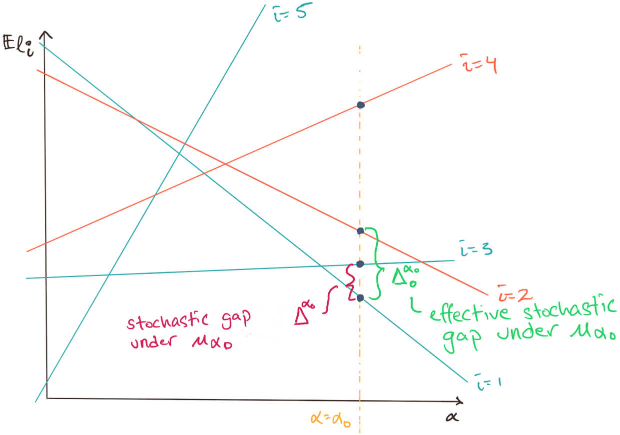


Constraint Example

$\mathcal{I}_0 = \{\text{experts that are optimal for some } \mu \in \mathcal{D}\}$ $\qquad N_0 = |\mathcal{I}_0|$

$$\Delta_0 = \inf_{\mu \in \mathcal{D}} \{\mu\text{-expected difference in loss of best expert and best expert not in } \mathcal{I}_0\}$$

Setting: the means for each expert are jointly defined by a parameter α,
$N = 5$, $\mathcal{I}_0 = \{1, 3, 5\}$, $N_0 = 3$.

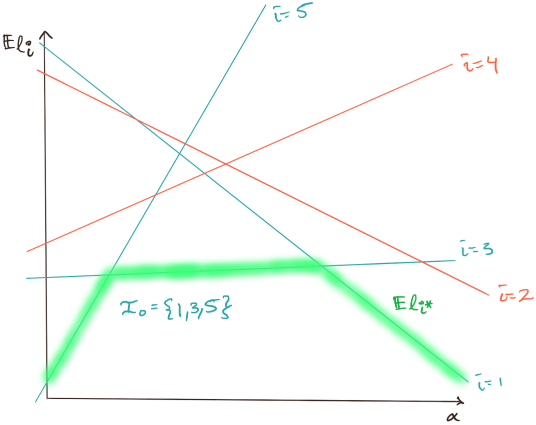


Constraint Example

$\mathcal{I}_0 = \{\text{experts that are optimal for some } \mu \in \mathcal{D}\}$ $\qquad N_0 = |\mathcal{I}_0|$

$$\Delta_0 = \inf_{\mu \in \mathcal{D}} \{\mu\text{-expected difference in loss of best expert and best expert not in } \mathcal{I}_0\}$$

Setting: the means for each expert are jointly defined by a parameter α,
$N = 5$, $\mathcal{I}_0 = \{1, 3, 5\}$, $N_0 = 3$.

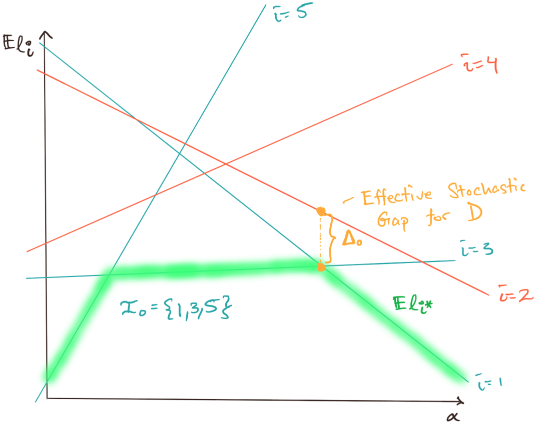


Constraint Example

$\mathcal{I}_0 = \{$experts that are optimal for some $\mu \in \mathcal{D}\}$ $\qquad N_0 = |\mathcal{I}_0|$

$$\Delta_0 = \inf_{\mu \in \mathcal{D}} \{\mu\text{-expected difference in loss of best expert and best expert not in } \mathcal{I}_0\}$$

Setting: the means for each expert are jointly defined by a parameter α,
$N = 5$, $\mathcal{I}_0 = \{1, 3, 5\}$, $N_0 = 3$.



Examples I

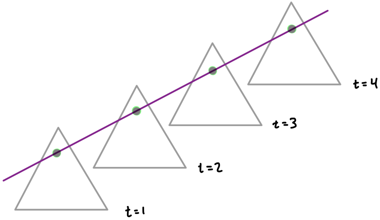
Examples I

I.I.D.-with-a-gap: $\mathcal{D} = \{\mu_0\}$,

Examples I

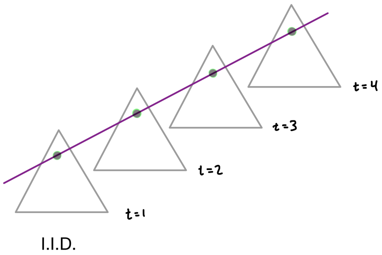
I.I.D.-with-a-gap: $\mathcal{D} = \{\mu_0\}$,

- $N_0 = 1$,
- $\Delta_0 = \Delta$



I.I.D.

Examples I

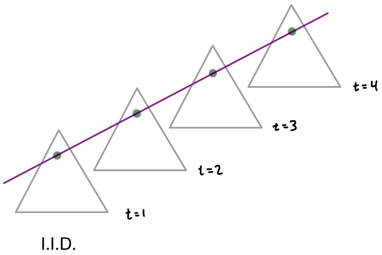


I.I.D.-with-a-gap: $\mathcal{D} = \{\mu_0\}$,

- $N_0 = 1$,
- $\Delta_0 = \Delta$

Adversarial: $\mathcal{D} = \mathcal{M}(\hat{\mathcal{Y}}^N \times \mathcal{Y})$

Examples I



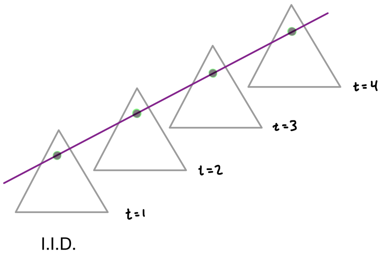
I.I.D.-with-a-gap: $\mathcal{D} = \{\mu_0\}$,

- $N_0 = 1$,
- $\Delta_0 = \Delta$

Adversarial: $\mathcal{D} = \mathcal{M}(\hat{\mathcal{Y}}^N \times \mathcal{Y}) \leftarrow$ contains point masses!

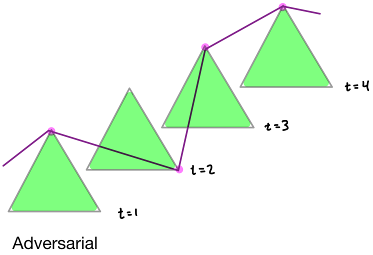
I.I.D.-with-a-gap: $\mathcal{D} = \{\mu_0\}$,

- $N_0 = 1$,
- $\Delta_0 = \Delta$



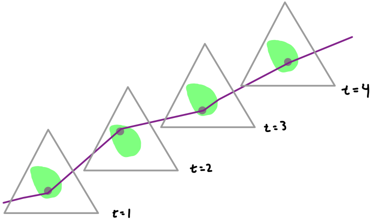
Adversarial: $\mathcal{D} = \mathcal{M}(\hat{\mathcal{Y}}^N \times \mathcal{Y}) \leftarrow$ contains point masses!

- $N_0 = N$,



Examples II

Adversarial-with-an-$\mathbb{E}$-gap [MG19]



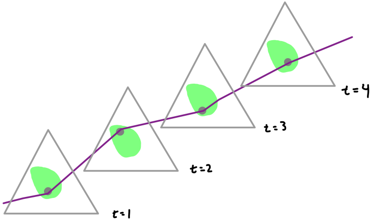
This work

Examples II

Adversarial-with-an-$\mathbb{E}$-gap [MG19]

- All measures where a common expert is better...

  ... than others in $\mathbb{E}$ by $\Delta > 0$.



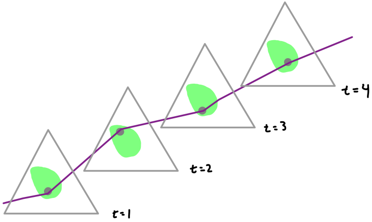
This work

Examples II

Adversarial-with-an-$\mathbb{E}$-gap [MG19]

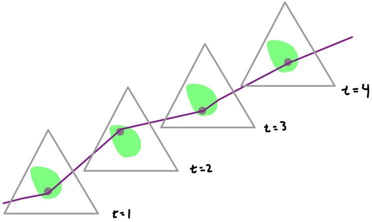
- All measures where a common expert is better…

  … than others in $\mathbb{E}$ by $\Delta > 0$.
- By design, $N_0 = 1$ and $\Delta_0 = \Delta$.



This work

Examples II



This work

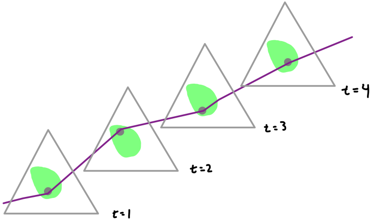
Adversarial-with-an-$\mathbb{E}$-gap [MG19]

- All measures where a common expert is better...

  ... than others in $\mathbb{E}$ by $\Delta > 0$.
- By design, $N_0 = 1$ and $\Delta_0 = \Delta$.

Non-creative adversary

Examples II



This work

Adversarial-with-an-$\mathbb{E}$-gap [MG19]

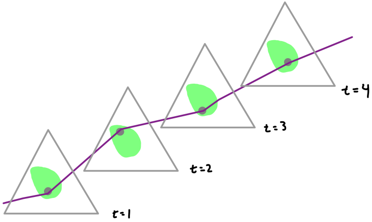
- All measures where a common expert is better...

  ... than others in $\mathbb{E}$ by $\Delta > 0$.
- By design, $N_0 = 1$ and $\Delta_0 = \Delta$.

Non-creative adversary

- The adversary has access to N_0 simple I.I.D. data sources,

Examples II



This work

Adversarial-with-an-$\mathbb{E}$-gap [MG19]

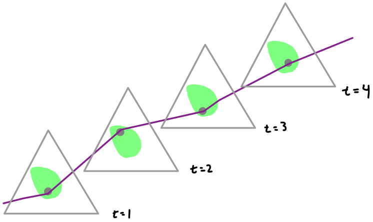
- All measures where a common expert is better...

  ... than others in $\mathbb{E}$ by $\Delta > 0$.
- By design, $N_0 = 1$ and $\Delta_0 = \Delta$.

Non-creative adversary

- The adversary has access to N_0 simple I.I.D. data sources,
- it can select from these sources adversarially.

Examples II



This work

Adversarial-with-an-$\mathbb{E}$-gap [MG19]

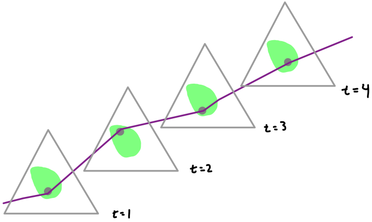
- All measures where a common expert is better...

  ... than others in $\mathbb{E}$ by $\Delta > 0$.
- By design, $N_0 = 1$ and $\Delta_0 = \Delta$.

Non-creative adversary

- The adversary has access to N_0 simple I.I.D. data sources,
- it can select from these sources adversarially.
- Assumption free way to model heterogeneous data sources.

Examples II



This work

Adversarial-with-an-$\mathbb{E}$-gap [MG19]

- All measures where a common expert is better...

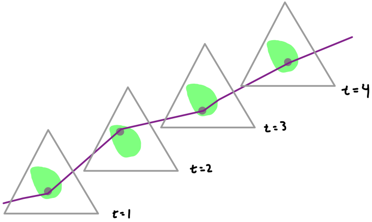
  ... than others in $\mathbb{E}$ by $\Delta > 0$.
- By design, $N_0 = 1$ and $\Delta_0 = \Delta$.

Non-creative adversary

- The adversary has access to N_0 simple I.I.D. data sources,
- it can select from these sources adversarially.
- Assumption free way to model heterogeneous data sources.

Neighborhood-of-I.I.D.

Examples II



This work

Adversarial-with-an-$\mathbb{E}$-gap [MG19]

- All measures where a common expert is better...

  ... than others in $\mathbb{E}$ by $\Delta > 0$.
- By design, $N_0 = 1$ and $\Delta_0 = \Delta$.

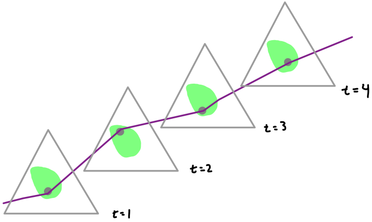
Non-creative adversary

- The adversary has access to N_0 simple I.I.D. data sources,
- it can select from these sources adversarially.
- Assumption free way to model heterogeneous data sources.

Neighborhood-of-I.I.D.

- Pick any distribution μ_0, and any radius, $r \geq 0$.

Examples II



This work

Adversarial-with-an-$\mathbb{E}$-gap [MG19]

- All measures where a common expert is better...

  ... than others in $\mathbb{E}$ by $\Delta > 0$.
- By design, $N_0 = 1$ and $\Delta_0 = \Delta$.

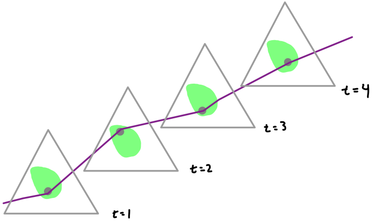
Non-creative adversary

- The adversary has access to N_0 simple I.I.D. data sources,
- it can select from these sources adversarially.
- Assumption free way to model heterogeneous data sources.

Neighborhood-of-I.I.D.

- Pick any distribution μ_0, and any radius, $r \geq 0$.
- $\mathcal{D} = \text{Ball}(\mu_0, r)$

Examples II



This work

Adversarial-with-an-$\mathbb{E}$-gap [MG19]

- All measures where a common expert is better...

  ... than others in $\mathbb{E}$ by $\Delta > 0$.
- By design, $N_0 = 1$ and $\Delta_0 = \Delta$.

Non-creative adversary

- The adversary has access to N_0 simple I.I.D. data sources,
- it can select from these sources adversarially.
- Assumption free way to model heterogeneous data sources.

Neighborhood-of-I.I.D.

- Pick any distribution μ_0, and any radius, $r \geq 0$.
- $\mathcal{D} = \text{Ball}(\mu_0, r)$
- N_0, Δ_0 depend on μ_0 and the radius of the ball...

Now we can get precise about the algorithms we study.

Follow the Regularized Leader

Follow the Regularized Leader

All explicit algorithms we will consider are *proper*:

the player randomly selects an expert to emulate at each time.

Follow the Regularized Leader

All explicit algorithms we will consider are *proper*:

the player randomly selects an expert to emulate at each time.

A proper algorithm assigns probability $w_i(t)$ to expert i at time t.

Follow the Regularized Leader

All explicit algorithms we will consider are *proper*:

the player randomly selects an expert to emulate at each time.

A proper algorithm assigns probability $w_i(t)$ to expert i at time t.

FTRL $\approx$ penalized empirical risk minimization.

Follow the Regularized Leader

All explicit algorithms we will consider are *proper*:

the player randomly selects an expert to emulate at each time.

A proper algorithm assigns probability $w_i(t)$ to expert i at time t.

FTRL $\approx$ penalized empirical risk minimization.

Parametrized by a sequence of regularizers $(\psi_t)_{t\in\mathbb{N}} \subseteq \mathtt{simp}([N]) \to \mathbb{R}$,

Follow the Regularized Leader

All explicit algorithms we will consider are *proper*:

the player randomly selects an expert to emulate at each time.

A proper algorithm assigns probability $w_i(t)$ to expert i at time t.

FTRL $\approx$ penalized empirical risk minimization.

Parametrized by a sequence of regularizers $(\psi_t)_{t\in\mathbb{N}} \subseteq \mathtt{simp}([N]) \to \mathbb{R}$,

$$w(t+1) = \operatorname*{arg\,min}_{w\in\mathtt{simp}([N])} \left(\langle w,\ L(t)\rangle + \psi_{t+1}(w)\right).$$

Linearly Decomposable FTRL

$$w(t+1) = \operatorname*{arg\,min}_{w \in \mathtt{simp}([N])} \left(\langle w, \; L(t) \rangle + \psi_{t+1}(w) \right).$$

Linearly Decomposable FTRL

$$w(t+1) = \operatorname*{arg\,min}_{w \in \mathtt{simp}([N])} \left(\langle w, \; L(t) \rangle + \psi_{t+1}(w) \right) .$$

Decompose the regularizer using a learning rate $\eta_t > 0$ and a function $f : [0,1] \to \mathbb{R}$

$$\psi_t(w) = \eta_t^{-1} \sum_{i=1}^{N} f(w_i).$$

Linearly Decomposable FTRL

$$w(t+1) = \arg\min_{w \in \texttt{simp}([N])} \left(\langle w, \ L(t) \rangle + \psi_{t+1}(w) \right).$$

Decompose the regularizer using a learning rate $\eta_t > 0$ and a function $f : [0,1] \to \mathbb{R}$

$$\psi_t(w) = \eta_t^{-1} \sum_{i=1}^{N} f(w_i).$$

With appropriate scaling by N, this looks like an f-divergence against a uniform over $[N]$.

Linearly Decomposable FTRL

$$w(t+1) = \operatorname*{arg\,min}_{w \in \mathtt{simp}([N])} \left(\langle w, \; L(t) \rangle + \psi_{t+1}(w) \right).$$

Decompose the regularizer using a learning rate $\eta_t > 0$ and a function $f : [0,1] \to \mathbb{R}$

$$\psi_t(w) = \eta_t^{-1} \sum_{i=1}^{N} f(w_i).$$

With appropriate scaling by N, this looks like an f-divergence against a uniform over $[N]$.

[MG19]: Hedge ($f(x) = x \log x$) with $\eta_t = \sqrt{(\log N)/t}$ is optimal for I.I.D. and adversarial.

Linearly Decomposable FTRL

$$w(t+1) = \operatorname*{arg\,min}_{w \in \mathtt{simp}([N])} \left(\langle w, \ L(t) \rangle + \psi_{t+1}(w) \right).$$

Decompose the regularizer using a learning rate $\eta_t > 0$ and a function $f : [0,1] \to \mathbb{R}$

$$\psi_t(w) = \eta_t^{-1} \sum_{i=1}^{N} f(w_i).$$

With appropriate scaling by N, this looks like an f-divergence against a uniform over $[N]$.

[MG19]: Hedge ($f(x) = x \log x$) with $\eta_t = \sqrt{(\log N)/t}$ is optimal for I.I.D. and adversarial.

What other f functions are useful?

Linearly Decomposable FTRL Regret Bounds

Generic local norm bound:

Theorem

For strictly convex f, we have almost surely that

$$R(T) \leq \underbrace{\eta_{T+1}^{-1} f(1)}_{\text{regularizer at comparator}} + \sum_{t=1}^{T} \left[\underbrace{\frac{\eta_t}{2} \sum_{i=1}^{N} \frac{(\ell_i(t) - m_t)^2}{f''(w_i(t))}}_{\text{local-norm curvature}} - \underbrace{\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) \sum_{i=1}^{N} f(w_i(t+1))}_{\text{regularizer increment}} \right]$$

Linearly Decomposable FTRL Regret Bounds

Generic local norm bound:

Theorem

For strictly convex f, we have almost surely that

$$R(T) \leq \underbrace{\eta_{T+1}^{-1} f(1)}_{\text{regularizer at comparator}} + \sum_{t=1}^{T} \left[\underbrace{\frac{\eta_t}{2} \sum_{i=1}^{N} \frac{(\ell_i(t) - m_t)^2}{f''(w_i(t))}}_{\text{local-norm curvature}} - \underbrace{\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \sum_{i=1}^{N} f(w_i(t+1))}_{\text{regularizer increment}} \right]$$

Choosing $f \cdot f'' \approx -1$ balances the local-norm curvature with the regularizer increment.

Linearly Decomposable FTRL Regret Bounds

Generic local norm bound:

Theorem

For strictly convex f, we have almost surely that

$$R(T) \leq \underbrace{\eta_{T+1}^{-1} f(1)}_{\text{regularizer at comparator}} + \sum_{t=1}^{T} \left[\underbrace{\frac{\eta_t}{2} \sum_{i=1}^{N} \frac{(\ell_i(t) - m_t)^2}{f''(w_i(t))}}_{\text{local-norm curvature}} - \underbrace{\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \sum_{i=1}^{N} f(w_i(t+1))}_{\text{regularizer increment}} \right]$$

Choosing $f \cdot f'' \approx -1$ balances the local-norm curvature with the regularizer increment.

This f makes these the correct order without needing $\sqrt{\log N}$ in η_t.

Linearly Decomposable FTRL Regret Bounds

Generic local norm bound:

Theorem

For strictly convex f, we have almost surely that

$$R(T) \leq \underbrace{\eta_{T+1}^{-1} f(1)}_{\text{regularizer at comparator}} + \sum_{t=1}^{T} \left[\underbrace{\frac{\eta_t}{2} \sum_{i=1}^{N} \frac{(\ell_i(t) - m_t)^2}{f''(w_i(t))}}_{\text{local-norm curvature}} - \underbrace{\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) \sum_{i=1}^{N} f(w_i(t+1))}_{\text{regularizer increment}} \right]$$

Choosing $f \cdot f'' \approx -1$ balances the local-norm curvature with the regularizer increment.
This f makes these the correct order without needing $\sqrt{\log N}$ in η_t.

Introducing **FTRL-CARL**:

Linearly Decomposable FTRL Regret Bounds

Generic local norm bound:

Theorem

For strictly convex f, we have almost surely that

$$R(T) \leq \underbrace{\eta_{T+1}^{-1} f(1)}_{\text{regularizer at comparator}} + \sum_{t=1}^{T} \left[\underbrace{\frac{\eta_t}{2} \sum_{i=1}^{N} \frac{(\ell_i(t) - m_t)^2}{f''(w_i(t))}}_{\text{local-norm curvature}} - \underbrace{\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \sum_{i=1}^{N} f(w_i(t+1))}_{\text{regularizer increment}} \right]$$

Choosing $f \cdot f'' \approx -1$ balances the local-norm curvature with the regularizer increment.
This f makes these the correct order without needing $\sqrt{\log N}$ in η_t.

Introducing **FTRL-CARL**:

Follow the Regularized Leader with Constraint-Adaptive Root-Logarithmic regularization

$$w(t+1) \in \mathop{\arg\min}_{w \in \mathtt{simp}([N])} \left(\langle w,\ L(t) \rangle - \sqrt{t+1} \sum_{i \in [N]} \int_0^{w_i} \sqrt{\log(1/s)}\, \mathrm{d}s \right).$$

Intuition for Improving on Hedge I

Thought Experiment

Intuition for Improving on Hedge I

Thought Experiment

Imagine we could run Hedge, but with a per-expert "learning rate" that...

Intuition for Improving on Hedge I

Thought Experiment

Imagine we could run Hedge, but with a per-expert “learning rate” that...

1. was $\propto \sqrt{(\log N_0)/t}$ for the effective experts,

Intuition for Improving on Hedge I

Thought Experiment

Imagine we could run Hedge, but with a per-expert “learning rate” that...

1. was $\propto \sqrt{(\log N_0)/t}$ for the effective experts,
2. and at least $\propto \sqrt{(\log N)/t}$ for the ineffective experts.

Intuition for Improving on Hedge I

Thought Experiment

Imagine we could run Hedge, but with a per-expert "learning rate" that...

1. was $\propto \sqrt{(\log N_0)/t}$ for the effective experts,
2. and at least $\propto \sqrt{(\log N)/t}$ for the ineffective experts.

How to achieve this?

Intuition for Improving on Hedge I

Thought Experiment

Imagine we could run Hedge, but with a per-expert "learning rate" that...

1. was $\propto \sqrt{(\log N_0)/t}$ for the effective experts,
2. and at least $\propto \sqrt{(\log N)/t}$ for the ineffective experts.

How to achieve this? Worst-case adversary forces weights to concentrate to $\mathrm{Unif}(\mathcal{I}_0)$, so

$$w_i(t) \asymp 1/N_0 \text{ for } i \in \mathcal{I}_0,$$
$$w_i(t) \prec 1/N \text{ else.}$$

Intuition for Improving on Hedge I

Thought Experiment

Imagine we could run Hedge, but with a per-expert "learning rate" that...

1. was $\propto \sqrt{(\log N_0)/t}$ for the effective experts,
2. and at least $\propto \sqrt{(\log N)/t}$ for the ineffective experts.

How to achieve this? Worst-case adversary forces weights to concentrate to $\mathrm{Unif}(\mathcal{I}_0)$, so

$$w_i(t) \asymp 1/N_0 \text{ for } i \in \mathcal{I}_0,$$
$$w_i(t) \prec 1/N \text{ else.}$$

This "idealized implicit learning rate" for expert i at time t, $\eta_i(t)$, could look like

$$\eta_i(t) = \sqrt{\frac{\log(1/w_i(t))}{t}} \quad .$$

Intuition for Improving on Hedge II

This “idealized implicit learning rate” for expert i at time t, $\eta_i(t)$, could look like

$$\eta_i(t) = \sqrt{\frac{\log(1/w_i(t))}{t}} \quad .$$

Intuition for Improving on Hedge II

This "idealized implicit learning rate" for expert i at time t, $\eta_i(t)$, could look like

$$\eta_i(t) = \sqrt{\frac{\log(1/w_i(t))}{t}} \quad .$$

In particular, imagine FTRL with per-expert "learning rates",

$$w(t+1) = \operatorname*{arg\,min}_{w \in \mathtt{simp}([N])} \left(\langle w,\ L(t) \rangle + \sum_{i=1}^{N} \eta_i(t+1)^{-1} f(w_i) \right).$$

Intuition for Improving on Hedge II

This "idealized implicit learning rate" for expert i at time t, $\eta_i(t)$, could look like

$$\eta_i(t) = \sqrt{\frac{\log(1/w_i(t))}{t}} \quad .$$

In particular, imagine *Hedge* with *implicit*, per-expert learning rates,

$$w(t+1) = \operatorname*{arg\,min}_{w \in \texttt{simp}([N])} \left(\langle w,\ L(t) \rangle - \sum_{i=1}^{N} \sqrt{\frac{t+1}{\log(1/w_i(t+1))}}\, w_i \log(1/w_i) \right) .$$

Intuition for Improving on Hedge II

This "idealized implicit learning rate" for expert i at time t, $\eta_i(t)$, could look like

$$\eta_i(t) = \sqrt{\frac{\log(1/w_i(t))}{t}} \quad .$$

In particular, imagine *Hedge* with *implicit*, per-expert learning rates,

$$w(t+1) = \underset{w \in \mathtt{simp}([N])}{\arg\min} \left(\langle w,\ L(t) \rangle - \sum_{i=1}^{N} \sqrt{\frac{t+1}{\log(1/w_i(t+1))}}\, w_i \log(1/w_i) \right).$$

This doesn't naively fit into the FTRL framework or analysis.

Intuition for Improving on Hedge II

This "idealized implicit learning rate" for expert i at time t, $\eta_i(t)$, could look like

$$\eta_i(t) = \sqrt{\frac{\log(1/w_i(t))}{t}} \ .$$

In particular, imagine *Hedge* with *implicit*, per-expert learning rates,

$$w(t+1) = \operatorname*{arg\,min}_{w \in \mathtt{simp}([N])} \left(\langle w,\ L(t) \rangle - \sum_{i=1}^{N} \sqrt{\frac{t+1}{\log(1/w_i(t+1))}}\, w_i \log(1/w_i) \right).$$

This doesn't naively fit into the FTRL framework or analysis.

FTRL-CARL with $\eta_t = 1/\sqrt{t}$ approximates Hedge with these implicit learning rates.

Intuition for Improving on Hedge II

This "idealized implicit learning rate" for expert i at time t, $\eta_i(t)$, could look like

$$\eta_i(t) = \sqrt{\frac{\log(1/w_i(t))}{t}} \quad .$$

In particular, imagine *Hedge* with *implicit*, per-expert learning rates,

$$w(t+1) = \operatorname*{arg\,min}_{w \in \mathtt{simp}([N])} \left(\langle w,\ L(t) \rangle - \sum_{i=1}^{N} \sqrt{\frac{t+1}{\log(1/w_i(t+1))}}\, w_i \log(1/w_i) \right) .$$

This doesn't naively fit into the FTRL framework or analysis.

FTRL-CARL with $\eta_t = 1/\sqrt{t}$ approximates Hedge with these implicit learning rates.

The $f(x) = -\int_0^x \sqrt{\log(1/s)}\,\mathrm{d}s$ for FTRL-CARL approximates $f(x) = -x\sqrt{\log(1/x)}$.

Regret Bounds

Regret Bounds

FTRL-CARL is adaptively minimax optimal.

Theorem

For any T and convex $\mathcal{D}$, FTRL-CARL with $\eta_t = 2/\sqrt{t}$ achieves

$$\mathbb{E}R(T) \leq \min\left(\sqrt{2T\log N_0} + 25\frac{\log N}{\Delta_0}, \sqrt{2T\log N}\right).$$

Regret Bounds

FTRL-CARL is adaptively minimax optimal.

Theorem

For any T and convex $\mathcal{D}$, FTRL-CARL with $\eta_t = 2/\sqrt{t}$ achieves

$$\mathbb{E}R(T) \leq \min\left(\sqrt{2T\log N_0} + 25\frac{\log N}{\Delta_0}, \sqrt{2T\log N}\right).$$

Hedge is not.

Theorem

Hedge with $\eta_t = \sqrt{(\log N)/t}$: for every $N_0 \geq 2$, there exists a convex $\mathcal{D}$ with

$$\mathbb{E}R(T) \gtrsim \sqrt{T\log N}.$$

Technical Bits: Minimax Concentration Inequality

When $N_0 \geq 2$, loss differences are not (sub/super)-martingales.

Technical Bits: Minimax Concentration Inequality

When $N_0 \geq 2$, loss differences are not (sub/super)-martingales.

Need a new way to show concentration of measure:

Technical Bits: Minimax Concentration Inequality

When $N_0 \geq 2$, loss differences are not (sub/super)-martingales.

Need a new way to show concentration of measure:

Lemma

For any prediction algorithm, constraint $\mathcal{D}$, and data-generating mechanism,

$$\sup_{i \in [N] \setminus \mathcal{I}_0} \mathbb{E} \min_{i_0 \in \mathcal{I}_0} \exp\left\{ \lambda \sum_{t=0}^{T} [\ell_{i_0}(t) - \ell_i(t)] \right\} \leq \exp\left\{ T \left[\lambda^2/2 - \lambda \Delta_0 \right] \right\}.$$

Technical Bits: Minimax Concentration Inequality

When $N_0 \geq 2$, loss differences are not (sub/super)-martingales.

Need a new way to show concentration of measure:

Lemma

For any prediction algorithm, constraint $\mathcal{D}$, and data-generating mechanism,

$$\sup_{i \in [N] \setminus \mathcal{I}_0} \mathbb{E} \min_{i_0 \in \mathcal{I}_0} \exp \left\{ \lambda \sum_{t=0}^{T} [\ell_{i_0}(t) - \ell_i(t)] \right\} \leq \exp \left\{ T \left[\lambda^2/2 - \lambda \Delta_0 \right] \right\}.$$

Relies on minimaxity. Not implied by Azuma-Hoeffding for $N_0 \geq 2$.

Our Contributions

Our Contributions

1. Introduced a spectrum of relaxations of the I.I.D. assumption.

Our Contributions

1. Introduced a spectrum of relaxations of the I.I.D. assumption.
    - Defined what it means for data to be nearly I.I.D.

Our Contributions

1. Introduced a spectrum of relaxations of the I.I.D. assumption.
   - Defined what it means for data to be nearly I.I.D.
   - We want to know that we do well even when I.I.D. fails.

Our Contributions

1. Introduced a spectrum of relaxations of the I.I.D. assumption.
   - Defined what it means for data to be nearly I.I.D.
   - We want to know that we do well even when I.I.D. fails.

2. Characterized minimax regret under time-homogeneous convex constraints.

Our Contributions

1. Introduced a spectrum of relaxations of the I.I.D. assumption.
   - Defined what it means for data to be nearly I.I.D.
   - We want to know that we do well even when I.I.D. fails.

2. Characterized minimax regret under time-homogeneous convex constraints.
   - Depends on the number of effective experts, N_0,
     and the effective stochastic gap, Δ_0.

Our Contributions

1. Introduced a spectrum of relaxations of the I.I.D. assumption.
   - Defined what it means for data to be nearly I.I.D.
   - We want to know that we do well even when I.I.D. fails.

2. Characterized minimax regret under time-homogeneous convex constraints.
   - Depends on the number of effective experts, N_0,
     and the effective stochastic gap, Δ_0.

3. Formalized the notion of adaptive minimax optimality.

Our Contributions

1. Introduced a spectrum of relaxations of the I.I.D. assumption.
   - Defined what it means for data to be nearly I.I.D.
   - We want to know that we do well even when I.I.D. fails.

2. Characterized minimax regret under time-homogeneous convex constraints.
   - Depends on the number of effective experts, N_0,
     and the effective stochastic gap, Δ_0.

3. Formalized the notion of adaptive minimax optimality.

4. Proved prevailing methods **are not** adaptively minimax optimal

Our Contributions

1. Introduced a spectrum of relaxations of the I.I.D. assumption.
   - Defined what it means for data to be nearly I.I.D.
   - We want to know that we do well even when I.I.D. fails.

2. Characterized minimax regret under time-homogeneous convex constraints.
   - Depends on the number of effective experts, N_0,
     and the effective stochastic gap, Δ_0.

3. Formalized the notion of adaptive minimax optimality.

4. Proved prevailing methods **are not** adaptively minimax optimal

5. Provided a new algorithm that **is** adaptively minimax optimal.

Our Contributions

1. Introduced a spectrum of relaxations of the I.I.D. assumption.
   - Defined what it means for data to be nearly I.I.D.
   - We want to know that we do well even when I.I.D. fails.

2. Characterized minimax regret under time-homogeneous convex constraints.
   - Depends on the number of effective experts, N_0,
     and the effective stochastic gap, Δ_0.

3. Formalized the notion of adaptive minimax optimality.

4. Proved prevailing methods **are not** adaptively minimax optimal

5. Provided a new algorithm that **is** adaptively minimax optimal.
   - Performs as well as possible relative to the constraint on the adversary,

Our Contributions

1. Introduced a spectrum of relaxations of the I.I.D. assumption.
   - Defined what it means for data to be nearly I.I.D.
   - We want to know that we do well even when I.I.D. fails.

2. Characterized minimax regret under time-homogeneous convex constraints.
   - Depends on the number of effective experts, N_0,
     and the effective stochastic gap, Δ_0.

3. Formalized the notion of adaptive minimax optimality.

4. Proved prevailing methods **are not** adaptively minimax optimal

5. Provided a new algorithm that **is** adaptively minimax optimal.
   - Performs as well as possible relative to the constraint on the adversary,
     without knowledge of the constraint.

Interpreting (N_0, Δ_0^{-1})

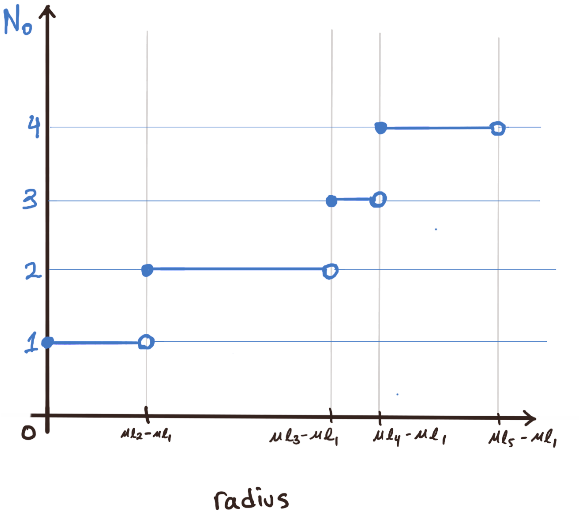
$\mathcal{D} = \mathsf{Ball}(\mu, \texttt{radius})$ w/ $\mathbb{E}_\mu \ell_1 < \mathbb{E}_\mu \ell_2 < \dots$

Interpreting (N_0, Δ_0^{-1})

$\mathcal{D} = \mathsf{Ball}(\mu, \texttt{radius})$ w/ $\mathbb{E}_\mu \ell_1 < \mathbb{E}_\mu \ell_2 < \dots$

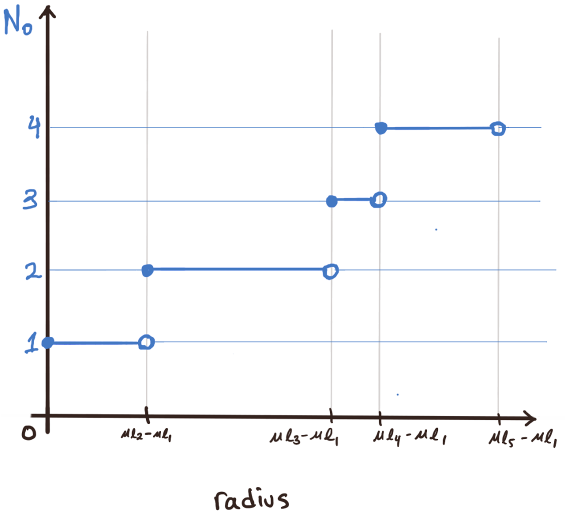
Interpreting (N_0, Δ_0^{-1})

$\mathcal{D} = \text{Ball}(\mu, \texttt{radius})$ w/ $\mathbb{E}_\mu \ell_1 < \mathbb{E}_\mu \ell_2 < \dots$



Interpreting (N_0, Δ_0^{-1})

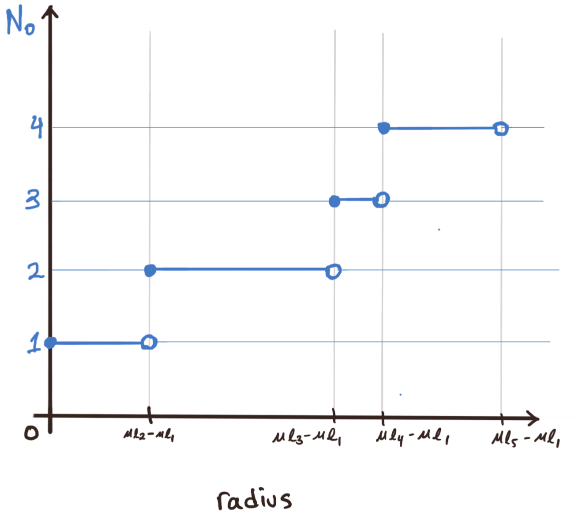
$\mathcal{D} = \text{Ball}(\mu, \texttt{radius})$ w/ $\mathbb{E}_\mu \ell_1 < \mathbb{E}_\mu \ell_2 < \dots$



- N_0 non-decreasing with radius

Interpreting (N_0, Δ_0^{-1})

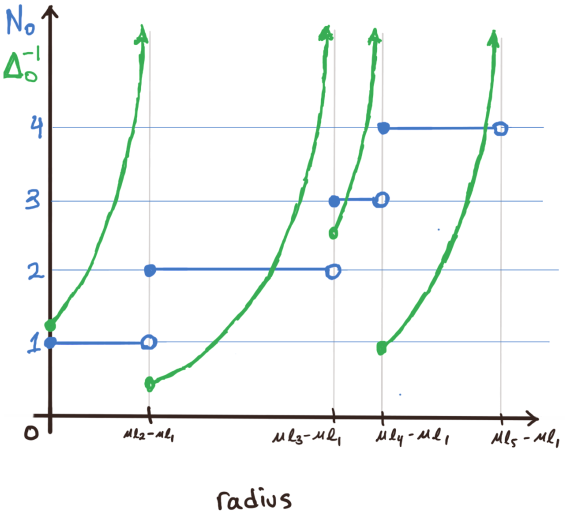
$\mathcal{D} = \text{Ball}(\mu, \texttt{radius})$ w/ $\mathbb{E}_\mu \ell_1 < \mathbb{E}_\mu \ell_2 < \dots$



- N_0 non-decreasing with radius
- N_0 increases discretely

Interpreting (N_0, Δ_0^{-1})

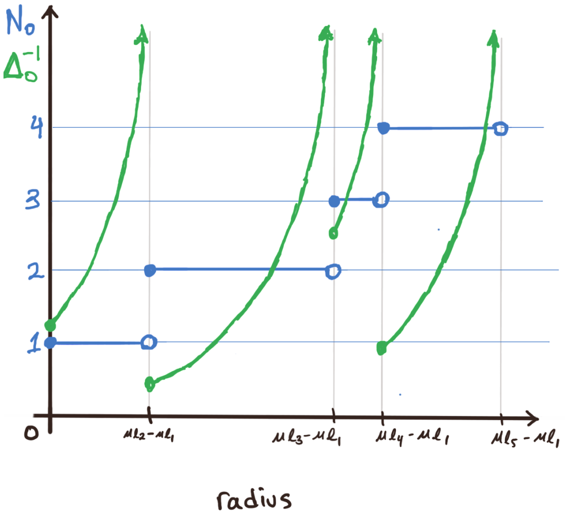
$\mathcal{D} = \mathrm{Ball}(\mu, \mathtt{radius})$ w/ $\mathbb{E}_\mu \ell_1 < \mathbb{E}_\mu \ell_2 < \dots$



- N_0 non-decreasing with radius
- N_0 increases discretely

Interpreting (N_0, Δ_0^{-1})

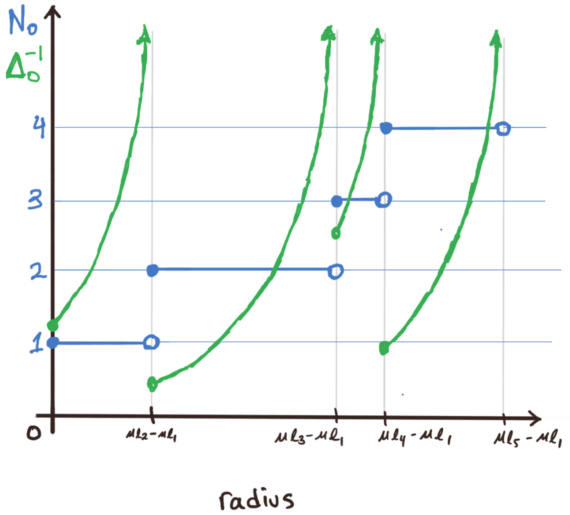
$\mathcal{D} = \text{Ball}(\mu, \texttt{radius})$ w/ $\mathbb{E}_\mu \ell_1 < \mathbb{E}_\mu \ell_2 < \dots$



- N_0 non-decreasing with radius
- N_0 increases discretely
- Δ_0^{-1} increases between N_0 jumps

Interpreting (N_0, Δ_0^{-1})

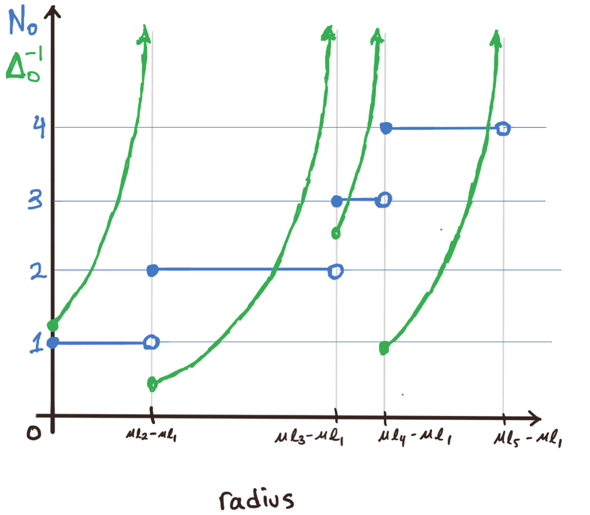
$\mathcal{D} = \text{Ball}(\mu, \texttt{radius})$ w/ $\mathbb{E}_\mu \ell_1 < \mathbb{E}_\mu \ell_2 < \dots$



- N_0 non-decreasing with radius
- N_0 increases discretely
- Δ_0^{-1} increases between N_0 jumps
- Δ_0^{-1} resets at each jump

Interpreting (N_0, Δ_0^{-1})

$\mathcal{D} = \mathrm{Ball}(\mu, \texttt{radius})$ w/ $\mathbb{E}_\mu \ell_1 < \mathbb{E}_\mu \ell_2 < \dots$

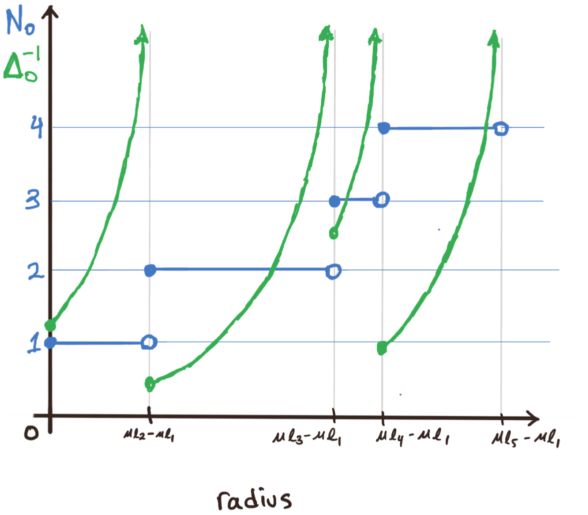


- N_0 non-decreasing with radius
- N_0 increases discretely
- Δ_0^{-1} increases between N_0 jumps
- Δ_0^{-1} resets at each jump

Lexicographical order on (N_0, Δ_0^{-1})

Interpreting (N_0, Δ_0^{-1})

$\mathcal{D} = \mathrm{Ball}(\mu, \mathtt{radius})$ w/ $\mathbb{E}_\mu \ell_1 < \mathbb{E}_\mu \ell_2 < \dots$



- N_0 non-decreasing with radius
- N_0 increases discretely
- Δ_0^{-1} increases between N_0 jumps
- Δ_0^{-1} resets at each jump

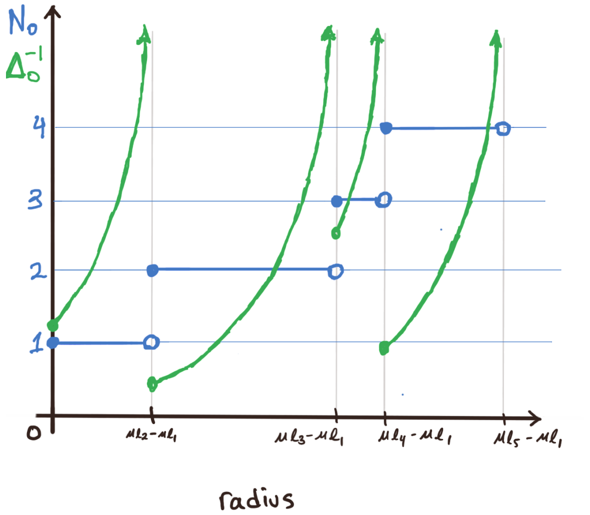
Lexicographical order on (N_0, Δ_0^{-1})

- For nested $\mathcal{D}$s,

  larger one is "harder".

Interpreting (N_0, Δ_0^{-1})

$\mathcal{D} = \mathrm{Ball}(\mu, \mathtt{radius})$ w/ $\mathbb{E}_\mu \ell_1 < \mathbb{E}_\mu \ell_2 < \dots$



- N_0 non-decreasing with radius
- N_0 increases discretely
- Δ_0^{-1} increases between N_0 jumps
- Δ_0^{-1} resets at each jump

Lexicographical order on (N_0, Δ_0^{-1})

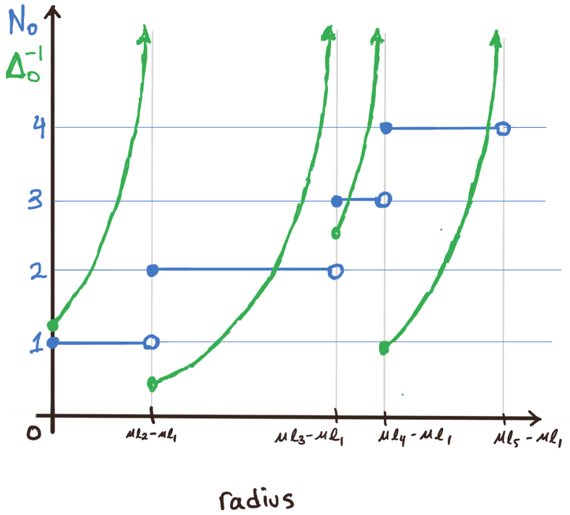
- For nested $\mathcal{D}$s,

  larger one is "harder".

- (N_0, Δ_0^{-1}) quantifies hardness.

Interpreting (N_0, Δ_0^{-1})

$\mathcal{D} = \text{Ball}(\mu, \texttt{radius})$ w/ $\mathbb{E}_\mu \ell_1 < \mathbb{E}_\mu \ell_2 < \dots$



- N_0 non-decreasing with radius
- N_0 increases discretely
- Δ_0^{-1} increases between N_0 jumps
- Δ_0^{-1} resets at each jump

Lexicographical order on (N_0, Δ_0^{-1})

- For nested $\mathcal{D}$s,

  larger one is "harder".

- (N_0, Δ_0^{-1}) quantifies hardness.

- $[\mathcal{D}, \subseteq] \mapsto [(N_0, \Delta_0^{-1}), \text{Lex}]$

  is order-preserving

A more refined bound for FTRL-CARL I

Theorem

For any time-homogeneous convex constraint $\mathcal{D}$, FTRL-CARL achieves:
For all T,

$$\mathbb{E}R_T \leq \sum_{t=1}^{T} \frac{1}{2\sqrt{t}} \sqrt{2 \log N_0^{(t)}} + \frac{20}{N\sqrt{\log N}} \sum_{i \in [N] \setminus \mathcal{I}_0} \frac{\mathbb{I}_{[T > T_i]}}{\Delta_i} + \sqrt{\log N},$$

where for each $i \in [N]$ and each $t \in \mathbb{N}$

$$\Delta_i = \inf_{\mu \in \mathcal{D}} \max_{i' \in [N]} \mu \left[\ell(i) - \ell(i') \right]$$

$$T_i = \left\lceil 8(\log N)/\Delta_i^2 \right\rceil$$

$$N_0^{(t)} = \left| \{ i \in [N] \ s.t. \ T_i > t \} \right|$$

A more refined bound for FTRL-CARL II

Theorem

For any time-homogeneous convex constraint $\mathcal{D}$, FTRL-CARL achieves:
For all T,

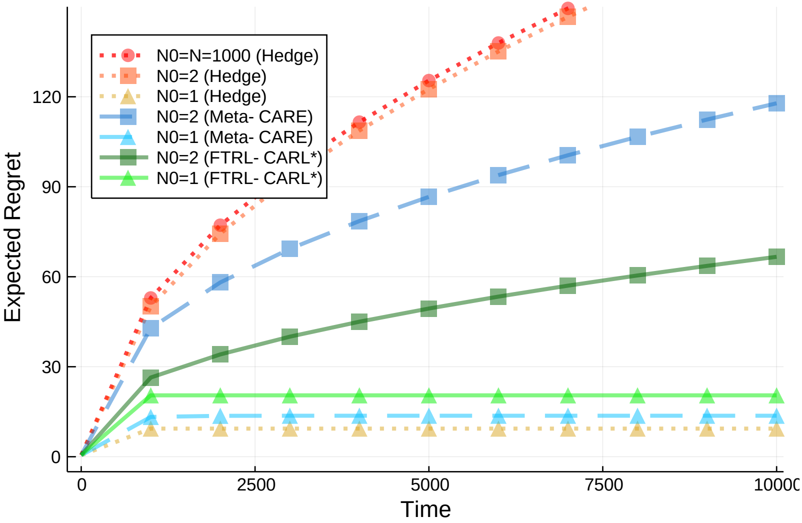
$$\mathbb{E}R_T \leq \sqrt{2T\log N},$$

and if $T > T_0$,

$$\mathbb{E}R_T \leq \sqrt{2T\log N_0} + 4(\log N) \sum_{j=0}^{N-N_0-1} W_{j,N,N_0} \frac{1}{\Delta_{(j)}} + \frac{20}{N\sqrt{\log N}} \sum_{i\in[N]\setminus\mathcal{I}_0} \frac{\mathbb{I}_{[T>T_i]}}{\Delta_i} + \sqrt{\log N},$$

where $W_{j,N,N_0} = \frac{1}{\sqrt{\log N}}\left(\sqrt{\log(N_0+j+1)} - \sqrt{\log(N_0+j)}\right)$.

Comparison of Methods



Optimality of Hedge for IID-with-a-Gap and Adversarial Cases