Relaxing the I.I.D. Assumption

Adaptively Minimax Optimal Regret via Root-Entropic Regularization

Blair Bilodeau*,1,2,3 and Jeffrey Negrea*,1,2,3
(Joint work with Daniel M. Roy1,2,3)

February 1, 2021
Presented to the RIKEN Center for Advanced Intelligence Project

* Equal Contribution
1 Department of Statistical Sciences, University of Toronto
2 Vector Institute
3 Institute for Advanced Study
Background
A Motivating Example

Stock Market Analogy

You need to invest your money into a stock portfolio.

You have access to several market experts that give you advice.

You regret not having always followed the post hoc best expert’s advice.

What assumptions should we make?

A simplifying assumption is that the data are I.I.D. (e.g., Black–Scholes–Merton).

In real life, market is driven in part by non-stochastic forces.

Is assuming adversarial data too pessimistic?

Is the departure from I.I.D.-ness benign? How can we quantify that?

Influence of non-stochastic forces “small” \(\Rightarrow\) maybe.

Meaning of “small” TBD.

Want to maximize profit without having to know what drives the market.
A Motivating Example

Stock Market Analogy

• You need to invest your money into a stock portfolio.
A Motivating Example

Stock Market Analogy

- You need to invest your money into a stock portfolio.
- You have access to several market experts that give you advice.
A Motivating Example

Stock Market Analogy

- You need to invest your money into a stock portfolio.
- You have access to several market experts that give you advice.
- You *regret* not having always followed the *post hoc* best expert’s advice.

What assumptions should we make?

A simplifying assumption is that the data are I.I.D. (e.g., Black–Scholes–Merton).

In real life, market is driven in part by non-stochastic forces.

Is assuming adversarial data too pessimistic?

Is the departure from I.I.D.-ness benign? How can we quantify that?

Influence of non-stochastic forces “small” ⇒ maybe.

Meaning of “small” TBD.

Want to maximize profit without having to know what drives the market.
A Motivating Example

Stock Market Analogy

- You need to invest your money into a stock portfolio.
- You have access to several market experts that give you advice.
- You *regret* not having always followed the *post hoc* best expert’s advice.

What assumptions should we make?

A simplifying assumption is that the data are I.I.D. (e.g., Black–Scholes–Merton).

In real life, market is driven in part by non-stochastic forces.

Is assuming adversarial data too pessimistic?

Is the departure from I.I.D.-ness benign? How can we quantify that?

Influence of non-stochastic forces “small” ⇒ maybe.

Meaning of “small” TBD.

Want to maximize profit without having to know what drives the market.
A Motivating Example

Stock Market Analogy

- You need to invest your money into a stock portfolio.
- You have access to several market experts that give you advice.
- You regret not having always followed the post hoc best expert’s advice.

What assumptions should we make?

A simplifying assumption is that the data are I.I.D. (e.g., Black–Scholes–Merton)
A Motivating Example

Stock Market Analogy

- You need to invest your money into a stock portfolio.
- You have access to several market experts that give you advice.
- You regret not having always followed the post hoc best expert’s advice

What assumptions should we make?

A simplifying assumption is that the data are I.I.D. (e.g., Black–Scholes–Merton)

In real life, market is driven in part by non-stochastic forces.
A Motivating Example

Stock Market Analogy

- You need to invest your money into a stock portfolio.
- You have access to several market experts that give you advice.
- You regret not having always followed the post hoc best expert’s advice

What assumptions should we make?

A simplifying assumption is that the data are I.I.D. (e.g., Black–Scholes–Merton)

In real life, market is driven in part by non-stochastic forces.

Is assuming adversarial data too pessimistic?
A Motivating Example

Stock Market Analogy

• You need to invest your money into a stock portfolio.
• You have access to several market experts that give you advice.
• You regret not having always followed the post hoc best expert’s advice

What assumptions should we make?

A simplifying assumption is that the data are I.I.D. (e.g., Black–Scholes–Merton)

In real life, market is driven in part by non-stochastic forces.

Is assuming adversarial data too pessimistic?

Is the departure from I.I.D.-ness benign? How can we quantify that?
A Motivating Example

Stock Market Analogy

• You need to invest your money into a stock portfolio.
• You have access to several market experts that give you advice.
• You regret not having always followed the post hoc best expert’s advice.

What assumptions should we make?

A simplifying assumption is that the data are I.I.D. (e.g., Black–Scholes–Merton).

In real life, market is driven in part by non-stochastic forces.

Is assuming adversarial data too pessimistic?

Is the departure from I.I.D.-ness benign? How can we quantify that?

Influence of non-stochastic forces “small” ⇒ maybe.
A Motivating Example

Stock Market Analogy

- You need to invest your money into a stock portfolio.
- You have access to several market experts that give you advice.
- You regret not having always followed the post hoc best expert’s advice

What assumptions should we make?

A simplifying assumption is that the data are I.I.D. (e.g., Black–Scholes–Merton)

In real life, market is driven in part by non-stochastic forces.

Is assuming adversarial data too pessimistic?

Is the departure from I.I.D.-ness benign? How can we quantify that?

  Influence of non-stochastic forces “small” \( \Rightarrow \) maybe.

  Meaning of "small" TBD.
A Motivating Example

Stock Market Analogy

- You need to invest your money into a stock portfolio.
- You have access to several market experts that give you advice.
- You regret not having always followed the post hoc best expert's advice

What assumptions should we make?

A simplifying assumption is that the data are I.I.D. (e.g., Black–Scholes–Merton)

In real life, market is driven in part by non-stochastic forces.

Is assuming adversarial data too pessimistic?

Is the departure from I.I.D.-ness benign? How can we quantify that?
  Influence of non-stochastic forces “small” ⇒ maybe.
  Meaning of "small" TBD.

Want to maximize profit without having to know what drives the market.
A Motivating Example

Stock Market Analogy

- You need to invest your money into a stock portfolio.
- You have access to several market experts that give you advice.
- You regret not having always followed the post hoc best expert’s advice

What assumptions should we make?

A simplifying assumption is that the data are I.I.D. (e.g., Black–Scholes–Merton)

In real life, market is driven in part by non-stochastic forces.

Is assuming adversarial data too pessimistic?

Is the departure from I.I.D.-ness benign? How can we quantify that?
  Influence of non-stochastic forces “small” ⇒ maybe.
  Meaning of "small" TBD.

Want to maximize profit without having to know what drives the market.
For rounds $t = 1, \ldots, T$:

- Predict $\hat{y}(t) \in \hat{Y}$ based on historical data before time $t$.
- Observe $y(t) \in Y$ from the environment.
- Incur loss $\ell(\hat{y}(t), y(t))$. 
Sequential Prediction a.k.a. Online Learning
For rounds $t = 1, \ldots, T$: 

• Predict $\hat{y}(t) \in \hat{Y}$ based on historical data before time $t$
• Observe $y(t) \in Y$ from the environment
• Incur loss $\ell(\hat{y}(t), y(t))$
Sequential Prediction a.k.a. Online Learning

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

- Predict $\hat{y}(t) \in \hat{Y}$ based on historical data before time $t$
Sequential Prediction a.k.a. Online Learning

For rounds \( t = 1, \ldots, T \):

- Predict \( \hat{y}(t) \in \hat{Y} \) based on historical data before time \( t \)
- Observe \( y(t) \in Y \) from the environment
Sequential Prediction a.k.a. Online Learning

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

- Predict $\hat{y}(t) \in \hat{Y}$ based on historical data before time $t$
- Observe $y(t) \in Y$ from the environment
- Incur loss $\ell(\hat{y}(t), y(t))$
Sequential Prediction with Expert Advice

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

- Receive $x(t) = (x_1(t), \ldots, x_N(t)) \subseteq \hat{Y}$ expert predictions
- Predict $\hat{y}(t) \in \hat{Y}$ based on historical data before time $t$ and expert predictions
- Observe $y(t) \in Y$ from the environment
- Incur loss $\ell(\hat{y}(t), y(t))$
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 $y(t) \in \mathcal{Y}$ from the environment
- Incur loss $\ell(\hat{y}(t), y(t))$
Sequential Prediction with Expert Advice

For rounds $t = 1, \ldots, T$:
- Receive $x(t) = (x_1(t), \ldots, x_N(t)) \subseteq \mathcal{Y}$ expert predictions
- Predict $\hat{y}(t) \in \mathcal{Y}$ based on historical data before time $t$ and expert predictions
- Observe $y(t) \in \mathcal{Y}$ from the environment
- Incur loss $\ell(\hat{y}(t), y(t))$
Sequential Prediction with Expert Advice

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

- Receive $x(t) = (x_1(t), \ldots, x_N(t)) \subseteq \hat{Y}$ expert predictions
- Predict $\hat{y}(t) \in \hat{Y}$ based on historical data before time $t$ and expert predictions
- Observe $y(t) \in Y$ from the environment
- Incur loss $\ell(\hat{y}(t), y(t))$
Sequential Prediction with Expert Advice

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

- Receive $x(t) = (x_1(t), \ldots, x_N(t)) \subseteq \hat{Y}$ expert predictions
- Predict $\hat{y}(t) \in \hat{Y}$ based on historical data before time $t$ and expert predictions
- Observe $y(t) \in Y$ from the environment
- Incur loss $\ell(\hat{y}(t), y(t))$
Sequential Prediction with Expert Advice

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

- Receive $x(t) = (x_1(t), \ldots, x_N(t)) \subseteq \hat{Y}$ expert predictions
- Predict $\hat{y}(t) \in \hat{Y}$ based on historical data before time $t$ and expert predictions
- Observe $y(t) \in Y$ from the environment
- Incur loss $\ell(\hat{y}(t), y(t))$
Sequential Prediction with Expert Advice

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

- Receive $x(t) = (x_1(t), \ldots, x_N(t)) \subseteq \hat{Y}$ expert predictions
- Predict $\hat{y}(t) \in \hat{Y}$ based on historical data before time $t$ and expert predictions
- Observe $y(t) \in Y$ from the environment
- Incur loss $\ell(\hat{y}(t), y(t))$
Measuring Performance

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

- Relative to the class of \( N \) reference experts;
- Given by the excess cumulative loss of the player over the best expert:

\[
\text{Regret: } R(T) = T \sum_{t=1}^{T} \ell(\hat{y}(t), y(t)) - \min_{i \in [N]} T \sum_{t=1}^{T} \ell(x_i(t), y(t))
\]

The prediction problem is online learnable if a player can incur sub-linear regret:

\[
E[R(T)] \in o(T).
\]

Where the \( E \) is taken with respect to the randomness in the player’s and expert’s predictions, and the data-generating mechanism for \((y(t))_{t \in N}\). (The \( E \) may be under a complicated, non-I.I.D. measure.)
Measuring Performance

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

- Relative to the class of $N$ reference experts;
Measuring Performance

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

- Relative to the class of $N$ reference experts;
- Given by the excess cumulative loss of the player over the best expert;

\[
\text{Regret: } R(T) = T \sum_{t=1}^{T} \ell(\hat{y}(t), y(t)) - \min_{i \in [N]} T \sum_{t=1}^{T} \ell(x_i(t), y(t))
\]

The prediction problem is online learnable if a player can incur sub-linear regret: $E[R(T)] \in o(T)$.

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

(The $E$ may be under a complicated, non-I.I.D. measure.)
Measuring Performance

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

- Relative to the class of $N$ reference experts;
- Given by the excess cumulative loss of the player over the best expert;

\[
\text{Regret: } R(T) = \sum_{t=1}^{T} \ell(\hat{y}(t), y(t)) - \min_{i \in [N]} \sum_{t=1}^{T} \ell(x_i(t), y(t))
\]
Measuring Performance

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

- Relative to the class of $N$ reference experts;
- Given by the excess cumulative loss of the player over the best expert;

\[
\text{Regret: } R(T) = \sum_{t=1}^{T} \ell(\hat{y}(t), y(t)) - \min_{i \in [N]} \sum_{t=1}^{T} \ell(x_i(t), y(t))
\]

The prediction problem is *online learnable* if a player can incur sub-linear regret:

\[
\mathbb{E}R(T) \in o(T).
\]
Measuring Performance

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

- Relative to the class of \( N \) reference experts;
- Given by the excess cumulative loss of the player over the best expert;

\[
\text{Regret: } R(T) = \sum_{t=1}^{T} \ell(\hat{y}(t), y(t)) - \min_{i \in [N]} \sum_{t=1}^{T} \ell(x_i(t), y(t))
\]

The prediction problem is online learnable if a player can incur sub-linear regret:

\[
\mathbb{E}R(T) \in o(T).
\]

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

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

- Relative to the class of $N$ reference experts;
- Given by the excess cumulative loss of the player over the best expert;

$$\text{Regret: } R(T) = \sum_{t=1}^{T} \ell(\hat{y}(t), y(t)) - \min_{i \in [N]} \sum_{t=1}^{T} \ell(x_i(t), y(t))$$

The prediction problem is online learnable if a player can incur sub-linear regret:

$$\mathbb{E}R(T) \in o(T).$$

Where the $\mathbb{E}$ is taken with respect to the randomness in the player’s and expert’s 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.)
Measuring Performance

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

- Relative to the class of $N$ reference experts;
- Given by the excess cumulative loss of the player over the best expert;

\[
\text{Regret: } R(T) = \sum_{t=1}^{T} \ell(\hat{y}(t), y(t)) - \min_{i \in [N]} \sum_{t=1}^{T} \ell(x_i(t), y(t))
\]

The prediction problem is online learnable if a player can incur sub-linear regret:

\[
\mathbb{E}R(T) \in o(T).
\]

Where the $\mathbb{E}$ is taken with respect to the randomness in the player’s and expert’s 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.)
Optimality in the Stochastic and Adversarial Regimes

Stochastic-with-a-Gap

• Expert predictions and data are I.I.D. over time from some distribution.
• There is an expert whose mean loss is $\Delta$ smaller than the others.

Theorem (Gaillard et al. 2014 + Mourtada and Gaïffas 2019)
A constructive algorithm achieves the minimax regret:
$$\mathbb{E} R(T) \approx \log N \Delta,$$
uniformly bounded in $T$.

Adversarial

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

Theorem (Vovk 1998, see also [FS97; CL06])
A constructive algorithm achieves the minimax regret:
$$\mathbb{E} R(T) \approx \sqrt{T \log N}$$
for all $T$.

Can a single algorithm be optimal in both settings simultaneously?
Optimality in the Stochastic and Adversarial Regimes

Stochastic-with-a-Gap

Expert predictions and data are I.I.D. over time from some distribution.

There is an expert whose mean loss is $\Delta$ smaller than the others.

Theorem (Gaillard et al. 2014 + Mourtada and Gaïffas 2019)

A constructive algorithm achieves the minimax regret:

$$E R(T) \approx \log N \Delta,$$

uniformly bounded in $T$.

Adversarial

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

Theorem (Vovk 1998, see also [FS97; CL06])

A constructive algorithm achieves the minimax regret:

$$E R(T) \approx \sqrt{T \log N},$$

for all $T$.

Can a single algorithm be optimal in both settings simultaneously?
Optimality in the Stochastic and Adversarial Regimes

**Stochastic-with-a-Gap**
- Expert predictions and data are I.I.D. over time from some distribution.
- There is an expert whose mean loss is $\Delta$ smaller than the others.

**Theorem (Gaillard et al. 2014 + Mourtada and Gaïffas 2019)**
A constructive algorithm achieves the minimax regret:
$$E[R(T)] \approx \log N \Delta,$$
uniformly bounded in $T$.

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

**Theorem (Vovk 1998, see also [FS97; CL06])**
A constructive algorithm achieves the minimax regret:
$$E[R(T)] \approx \sqrt{T \log N}$$
for all $T$. Can a single algorithm be optimal in both settings simultaneously?
Optimality in the Stochastic and Adversarial Regimes

Stochastic-with-a-Gap

- Expert predictions and data are i.i.d. over time from some distribution.
- There is an expert whose mean loss is $\Delta$ smaller than the others.

**Theorem (Gaillard et al. 2014 ± Mourtada and Gaïffas 2019)**

A constructive algorithm achieves the minimax regret:

$$\mathbb{E} R(T) \asymp \frac{\log N}{\Delta},$$

uniformly bounded in $T$. 

Adversarial

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

**Theorem (Vovk 1998, see also [FS97; CL06])**

A constructive algorithm achieves the minimax regret:

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

for all $T$. 

Can a single algorithm be optimal in both settings simultaneously?
Optimality in the Stochastic and Adversarial Regimes

**Stochastic-with-a-Gap**
- Expert predictions and data are i.i.d. over time from some distribution.
- There is an expert whose mean loss is $\Delta$ smaller than the others.

**Theorem (Gaillard et al. 2014 + Mourtada and Gaïffas 2019)**
A constructive algorithm achieves the minimax regret:
$$
\mathbb{E} R(T) \lesssim \frac{\log N}{\Delta},
$$
uniformly bounded in $T$.

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

Optimality in the Stochastic and Adversarial Regimes

**Stochastic-with-a-Gap**
- Expert predictions and data are I.I.D. over time from some distribution.
- There is an expert whose mean loss is $\Delta$ smaller than the others.

**Theorem (Gaillard et al. 2014 + Mourtada and Gaïffas 2019)**
A constructive algorithm achieves the minimax regret:
$$\mathbb{E} R(T) \asymp \frac{\log N}{\Delta},$$
uniformly bounded in $T$.

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

**Theorem (Vovk 1998, see also [FS97; CL06])**
A constructive algorithm achieves the minimax regret:
$$\mathbb{E} R(T) \asymp \sqrt{T \log N}$$
for all $T$. 

Can a single algorithm be optimal in both settings simultaneously?
Optimality in the Stochastic and Adversarial Regimes

Stochastic-with-a-Gap

- Expert predictions and data are I.I.D. over time from some distribution.
- There is an expert whose mean loss is $\Delta$ smaller than the others.

**Theorem (Gaillard et al. 2014 + Mourtada and Gaïffas 2019)**

A constructive algorithm achieves the minimax regret:

$$\mathbb{E} R(T) \asymp \frac{\log N}{\Delta}, \text{ uniformly bounded in } T.$$ 

Adversarial

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

**Theorem (Vovk 1998, see also [FS97; CL06])**

A constructive algorithm achieves the minimax regret:

$$\mathbb{E} R(T) \asymp \sqrt{T \log N} \text{ for all } T.$$ 

Can a single algorithm be optimal in both settings simultaneously?
Optimality in the Stochastic and Adversarial Regimes

**Stochastic-with-a-Gap**
- Expert predictions and data are i.i.d. over time from some distribution.
- There is an expert whose mean loss is $\Delta$ smaller than the others.

**Theorem (Gaillard et al. 2014 + Mourtada and Gaïffas 2019)**
A constructive algorithm achieves the minimax regret:
$$\mathbb{E} R(T) \asymp \frac{\log N}{\Delta},$$
uniformly bounded in $T$.

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

**Theorem (Vovk 1998, see also [FS97; CL06])**
A constructive algorithm achieves the minimax regret:
$$\mathbb{E} R(T) \asymp \sqrt{T \log N}$$
for all $T$.

Can a single algorithm be optimal in both settings simultaneously?
Simultaneous Optimality of Hedge

Can a single algorithm be optimal in both settings simultaneously? Yes! [MG19]
Simultaneous Optimality of Hedge

Can a single algorithm be optimal in both settings simultaneously? Yes! [MG19]

**Stochastic-with-a-gap:** $\mathbb{E} R(T) \asymp (\log N)/\Delta$ uniformly in $T$.

**Adversarial:** $\mathbb{E} R(T) \asymp \sqrt{T \log N}$
Simultaneous Optimality of Hedge

Can a single algorithm be optimal in both settings simultaneously? Yes! [MG19]

Stochastic-with-a-gap: $\mathbb{E} R(T) \asymp \frac{(\log N)}{\Delta}$ uniformly in $T$.

Adversarial: $\mathbb{E} R(T) \asymp \sqrt{T \log N}$

The same algorithm, Hedge, was used in both cases!
Simultaneous Optimality of Hedge

Can a single algorithm be optimal in both settings simultaneously? Yes! [MG19]

**Stochastic-with-a-gap:** \( E[R(T)] \asymp (\log N)/\Delta \) uniformly in \( T \).

**Adversarial:** \( E[R(T)] \asymp \sqrt{T \log N} \)

The same algorithm, Hedge, was used in both cases!
Real data $\not\equiv$ stochastic.
Real data $\not\equiv$ stochastic. ← Too optimistic.
Beyond Stochastic and Adversarial

Real data $\not\equiv$ stochastic. ← Too optimistic.

Real data $\not\equiv$ adversarial.
Beyond Stochastic and Adversarial

Real data \(\not\equiv\) stochastic. \(\leftarrow\) Too optimistic.

Real data \(\not\equiv\) adversarial. \(\leftarrow\) Too pessimistic.
Beyond Stochastic and Adversarial

Real data $\not= \text{stochastic.} \leftarrow \text{Too optimistic.}$

Real data $\not= \text{adversarial.} \leftarrow \text{Too pessimistic.}$

We provide a spectrum between stochastic and adversarial;
Beyond Stochastic and Adversarial

Real data $\not\equiv$ stochastic. ← Too optimistic.

Real data $\not\equiv$ adversarial. ← Too pessimistic.

We provide a spectrum between stochastic and adversarial;
   Intuitively, fix a “neighbourhood” of distributions;
Beyond Stochastic and Adversarial

Real data $\not\equiv$ stochastic. ← Too optimistic.

Real data $\not\equiv$ adversarial. ← Too pessimistic.

We provide a spectrum between stochastic and adversarial;

Intuitively, fix a “neighbourhood” of distributions;
Each data point drawn from an arbitrary distribution in “neighbourhood”.
Beyond Stochastic and Adversarial

Real data $\not\equiv$ stochastic. ← Too optimistic.

Real data $\not\equiv$ adversarial. ← Too pessimistic.

We provide a spectrum between stochastic and adversarial;
    Intuitively, fix a “neighbourhood” of distributions;
    Each data point drawn from an arbitrary distribution in “neighbourhood”.

Stochastic

Adversarial

This work
Beyond Stochastic and Adversarial

Real data \( \neq \) stochastic. ← Too optimistic.

Real data \( \neq \) adversarial. ← Too pessimistic.

We provide a spectrum between stochastic and adversarial;
Intuitively, fix a “neighbourhood” of distributions;
Each data point drawn from an arbitrary distribution in “neighbourhood”. 
Prediction algorithms should be robust to a range of data generating mechanisms.
Adaptively Minimax Optimal Algorithms

Prediction algorithms should be robust to a range of data generating mechanisms.

**Definition BNR20**

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

Prediction algorithms should be robust to a range of data generating mechanisms.

**Definition BNR20**

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.

How to formalize this?
Adaptively Minimax Optimal Algorithms

Prediction algorithms should be robust to a range of data generating mechanisms.

**Definition BNR20**

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.

**How to formalize this?**

- For an abstract range of settings $\Theta$...
Adaptively Minimax Optimal Algorithms

Prediction algorithms should be robust to a range of data generating mechanisms.

**Definition BNR20**

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.

**How to formalize this?**

- For an abstract range of settings $\Theta$...
- Parameterize the minimax regret in each setting: $(R^*_\theta(T))_{\theta \in \Theta}$. 
Prediction algorithms should be robust to a range of data generating mechanisms.

**Definition BNR20**

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.

**How to formalize this?**

- For an abstract range of settings $\Theta$...
- Parameterize the minimax regret in each setting: $(R^*_\theta(T))_{\theta \in \Theta}$.
- Algorithm satisfies $R_\theta(T) \leq C R^*_\theta(T)$ uniformly in $\theta$ for large enough $T$. 
Adaptively Minimax Optimal Algorithms

Prediction algorithms should be robust to a range of data generating mechanisms.

**Definition BNR20**

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.

**How to formalize this?**

- For an abstract range of settings $\Theta$...
- Parameterize the minimax regret in each setting: $(R^*_\theta(T))_{\theta \in \Theta}$.
- Algorithm satisfies $R_\theta(T) \leq C R^*_\theta(T)$ uniformly in $\theta$ for large enough $T$. 
Overview of Our Work

We show Hedge is suboptimal between Stochastic and Adversarial. This was surprising for us. We initially set out hoping to prove that Hedge was adaptive to all scenarios.

Theorem BNR20

Without oracle knowledge to tune the learning rate, Hedge is not simultaneously minimax optimal at all settings between stochastic-with-a-gap and adversarial.

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

Theorem BNR20

There is an adaptively minimax optimal algorithm: Meta-CARE.
Overview of Our Work

We show Hedge is suboptimal between Stochastic and Adversarial.

Theorem BNR20

Without oracle knowledge to tune the learning rate, Hedge is not simultaneously minimax optimal at all settings between stochastic-with-a-gap and adversarial.

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

...without knowledge of which setting prevails!

Theorem BNR20

There is an adaptively minimax optimal algorithm: Meta-CARE.
Overview of Our Work

We show Hedge is suboptimal between Stochastic and Adversarial. This was surprising for us.
Overview of Our Work

We show Hedge is suboptimal between Stochastic and Adversarial.

This was surprising for us.

We initially set out hoping to prove that Hedge was adaptive to all scenarios.
Overview of Our Work

We show Hedge is suboptimal between Stochastic and Adversarial.

This was surprising for us.

We initially set out hoping to prove that Hedge was adaptive to all scenarios.

**Theorem BNR20**

Without oracle knowledge to tune the learning rate, Hedge is not simultaneously minimax optimal at all settings between stochastic-with-a-gap and adversarial.
Overview of Our Work

We show Hedge is suboptimal between Stochastic and Adversarial. This was surprising for us. We initially set out hoping to prove that Hedge was adaptive to all scenarios.

**Theorem BNR20**

Without oracle knowledge to tune the learning rate, Hedge is not simultaneously minimax optimal at all settings between stochastic-with-a-gap and adversarial.

We provide a new algorithm that achieves the minimax rate in all settings...
Overview of Our Work

We show Hedge is suboptimal between Stochastic and Adversarial.

This was surprising for us.

We initially set out hoping to prove that Hedge was adaptive to all scenarios.

**Theorem BNR20**

Without oracle knowledge to tune the learning rate, Hedge is not simultaneously minimax optimal at all settings between stochastic-with-a-gap and adversarial.

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

...without knowledge of which setting prevails!
Overview of Our Work

We show Hedge is suboptimal between Stochastic and Adversarial.
This was surprising for us.
We initially set out hoping to prove that Hedge was adaptive to all scenarios.

**Theorem BNR20**
Without oracle knowledge to tune the learning rate, Hedge is not simultaneously minimax optimal at all settings between stochastic-with-a-gap and adversarial.

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

**Theorem BNR20**
There is an adaptively minimax optimal algorithm: Meta-CARE.
Overview of Our Work

We show Hedge is suboptimal between Stochastic and Adversarial.
This was surprising for us.
We initially set out hoping to prove that Hedge was adaptive to all scenarios.

**Theorem BNR20**
Without oracle knowledge to tune the learning rate, Hedge is not simultaneously minimax optimal at all settings between stochastic-with-a-gap and adversarial.

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

**Theorem BNR20**
There is an adaptively minimax optimal algorithm: Meta-CARE.
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 regret $\Theta(\sqrt{T \log N_0})$ in $(T, N_0)$.

If we know one expert is better than the rest by $\Delta_0$, but not which it is, then we are almost in the stochastic-with-a-gap case so we might hope for regret $\Theta((\log N_0) / \Delta_0)$.

Theorem BNR20

The adaptively minimax optimal rate of regret, which Meta-CARE achieves, is $\mathbb{E} R(T) \approx \begin{cases} \sqrt{T \log N_0} & N_0 \geq 2 \\ (\log N_0) / \Delta_0 & N_0 = 1 \end{cases}$. 
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, ...

Theorem BNR20

The adaptively minimax optimal rate of regret, which Meta-CARE achieves, is

\[
E R(T) \asymp \begin{cases} \\
\sqrt{T \log N_0} & N_0 \geq 2 \\
\frac{\log N}{\Delta_0} & N_0 = 1 
\end{cases}
\]
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 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 regret $\Theta(\sqrt{T \log N_0})$ in $(T, N_0)$

• If we know one expert is better than the rest by $\Delta_0$, but not which it is...
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 regret $\Theta(\sqrt{T \log N_0})$ in $(T, N_0)$

- If we know one expert is better than the rest by $\Delta_0$, but not which it is...
  - then we are *almost* in the stochastic-with-a-gap 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 regret $\Theta(\sqrt{T \log N_0})$ in $(T, N_0)$

- If we know one expert is better than the rest by $\Delta_0$, but not which it is...  
  - then we are almost in the stochastic-with-a-gap case  
  - so we might hope for regret $\Theta((\log N) / \Delta_0)$
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 regret $\Theta(\sqrt{T \log N_0})$ in $(T, N_0)$

- If we know one expert is better than the rest by $\Delta_0$, but not which it is...
  - then we are almost in the stochastic-with-a-gap case
  - so we might hope for regret $\Theta((\log N)/\Delta_0)$

Theorem BNR20

The adaptively minimax optimal rate of regret, which Meta-CARE achieves, is

$$\mathbb{E}R(T) \lesssim \begin{cases} 
\sqrt{T \log N_0} & N_0 \geq 2 \\
(\log N)/\Delta_0 & N_0 = 1.
\end{cases}$$
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 regret $\Theta(\sqrt{T\log N_0})$ in $(T, N_0)$

• If we know one expert is better than the rest by $\Delta_0$, but not which it is...
  • then we are almost in the stochastic-with-a-gap case
  • so we might hope for regret $\Theta((\log N)/\Delta_0)$

Theorem BNR20

The adaptively minimax optimal rate of regret, which Meta-CARE achieves, is

$$\mathbb{E} R(T) \asymp \begin{cases} \sqrt{T\log N_0} & N_0 \geq 2 \\ (\log N)/\Delta_0 & N_0 = 1. \end{cases}$$
Hedge Algorithm

We will consider only finite expert classes and bounded losses $\ell : \hat{Y} \times Y \to [0, 1]$. 

---

Hedge Algorithm

• Fix learning rate schedule $\eta : N \to R$; initialize the weights as uniform; define $\ell_i(t) = \ell(x_i(t), y(t))$, $L_i(t) = \sum_{s=1}^{t} \ell_i(s)$.

• Update weights for each $i \in [N]$ using $w_i(t) \propto \exp\{-\eta(t)L_i(t-1)\}$.
**Hedge Algorithm**

We will consider only finite expert classes and bounded losses $\ell : \hat{Y} \times Y \rightarrow [0, 1]$. All explicit algorithms we will consider are *proper*:

- Fix learning rate schedule $\eta : \mathbb{N} \rightarrow \mathbb{R}$; initialize the weights as uniform; define $\ell_i(t) = \ell(x_i(t), y(t))$, $L_i(t) = \sum_{s=1}^{t} \ell_i(s)$.
- Update weights for each $i \in [N]$ using $w_i(t) \propto \exp\{-\eta(t)L_i(t-1)\}$. 


We will consider only finite expert classes and bounded losses \( \ell : \hat{\mathcal{Y}} \times \mathcal{Y} \rightarrow [0, 1] \).

All explicit algorithms we will consider are *proper*:
- the player randomly selects an expert to emulate at each time.

**Hedge Algorithm**

- Fix learning rate schedule \( \eta : \mathbb{N} \rightarrow \mathbb{R} \); initialize the weights as uniform; define
  \[
  \ell_i(t) = \ell(x_i(t), y(t)),
  \]
  \[
  L_i(t) = t \sum_{s=1}^{t} \ell_i(s).
  \]

- Update weights for each \( i \in \mathbb{N} \) using
  \[
  w_i(t) \propto e^{-\eta(t)(L_i(t) - L_i(t-1))}.
  \]
Hedge Algorithm

We will consider only finite expert classes and bounded losses $\ell : \hat{Y} \times Y \to [0, 1]$.

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 $t$. 
Hedge Algorithm

We will consider only finite expert classes and bounded losses $\ell : \hat{Y} \times Y \to [0, 1]$.

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 $t$.

Hedge Algorithm
Hedge Algorithm

We will consider only finite expert classes and bounded losses $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, 1]$. 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 $t$.

**Hedge Algorithm**

- Fix learning rate schedule $\eta : \mathbb{N} \rightarrow \mathbb{R}$; initialize the weights as uniform; define
Hedge Algorithm

We will consider only finite expert classes and bounded losses $\ell : \hat{Y} \times Y \rightarrow [0, 1]$.

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 $t$.

Hedge Algorithm

- Fix learning rate schedule $\eta : \mathbb{N} \rightarrow \mathbb{R}$; initialize the weights as uniform; define

$$\ell_i(t) = \ell(x_i(t), y(t)), \quad L_i(t) = \sum_{s=1}^{t} \ell_i(s).$$
We will consider only finite expert classes and bounded losses $\ell : \hat{Y} \times Y \to [0, 1]$.

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 $t$.

**Hedge Algorithm**

- Fix learning rate schedule $\eta : \mathbb{N} \to \mathbb{R}$; initialize the weights as uniform; define

  $$\ell_i(t) = \ell(x_i(t), y(t)), \quad L_i(t) = \sum_{s=1}^{t} \ell_i(s).$$

- Update weights for each $i \in [N]$ using

  $$w_i(t) \propto \exp \left\{ -\eta(t)L_i(t - 1) \right\}.$$
**Hedge Algorithm**

We will consider only finite expert classes and bounded losses $\ell : \hat{Y} \times Y \rightarrow [0, 1]$. 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 $t$.

**Hedge Algorithm**

- Fix learning rate schedule $\eta : \mathbb{N} \rightarrow \mathbb{R}$; initialize the weights as uniform; define
  \[
  \ell_i(t) = \ell(x_i(t), y(t)), \quad L_i(t) = \sum_{s=1}^{t} \ell_i(s).
  \]
- Update weights for each $i \in [N]$ using
  \[
  w_i(t) \propto \exp \left\{ -\eta(t)L_i(t - 1) \right\}.
  \]
Hedge as Bayesian Inference

\[ w_i(t) \propto \exp \left\{ -\eta(t)L_i(t - 1) \right\} \, . \]
$w_i(t) \propto \exp \{-\eta(t)L_i(t-1)\}.$

For $\eta(t) = \eta$ constant in $t$, this looks like Bayes rule for a parameter in $[N]$. 
Hedge as Bayesian Inference

\[ w_i(t) \propto \exp \{-\eta(t)L_i(t - 1)\} . \]

For \( \eta(t) = \eta \) constant in \( t \), this looks like Bayes rule for a parameter in \( [N] \)
- with a flat prior, and
Hedge as Bayesian Inference

\[ w_i(t) \propto \exp \{-\eta(t)L_i(t-1)\}. \]

For \( \eta(t) = \eta \) constant in \( t \), this looks like Bayes rule for a parameter in \([N]\) with a flat prior, and model likelihood \( \exp\{-\eta \ell_i(t)\} \) for the \( t \)-th observation under parameter \( i \).
Hedge as Bayesian Inference

\[ w_i(t) \propto \exp \{-\eta(t)L_i(t - 1)\}. \]

For \( \eta(t) = \eta \) constant in \( t \), this looks like Bayes rule for a parameter in \([N]\)

- with a flat prior, and
- model likelihood \( \exp\{-\eta \ell_i(t)\} \) for the \( t \)-th observation under parameter \( i \).

**Variational Formulation**

\[
\begin{align*}
    w(t) &= \arg \min_{w \in \text{simp}([N])} \left( \langle w, L(t - 1) \rangle - \frac{1}{\eta(t)} H(w) \right) \\
    \text{where} \quad H(w) &= -\sum_{i \in [N]} w_i \log(w_i).
\end{align*}
\]
Hedge as Bayesian Inference

\[ w_i(t) \propto \exp \{ -\eta(t)L_i(t - 1) \}. \]

For \( \eta(t) = \eta \) constant in \( t \), this looks like Bayes rule for a parameter in \([N]\)
- with a flat prior, and
- model likelihood \( \exp \{ -\eta \ell_i(t) \} \) for the \( t \)-th observation under parameter \( i \).

**Variational Formulation**

\[
\begin{aligned}
w(t) = & \text{arg min}_{w \in \text{simp}([N])} \left( \langle w, L(t - 1) \rangle - \frac{1}{\eta(t)} H(w) + \frac{1}{\eta(t)} \sum_{i=1}^{N} w_i \log(N) \right) \\
\end{aligned}
\]

where

\[
H(w) = - \sum_{i \in [N]} w_i \log(w_i).
\]
Hedge as Bayesian Inference

\[ w_i(t) \propto \exp \{-\eta(t)L_i(t - 1)\}. \]

For \( \eta(t) = \eta \) constant in \( t \), this looks like Bayes rule for a parameter in \([N]\)

- with a flat prior, and
- model likelihood \( \exp\{-\eta\ell_i(t)\} \) for the \( t \)-th observation under parameter \( i \).

**Variational Formulation**

\[
w(t) = \arg\min_{w \in \text{simp}([N])} \left( \langle w, L(t - 1) \rangle + \frac{1}{\eta(t)} \text{KL}(w \| \text{Unif}([N])) \right)
\]

where

\[
\text{KL}(w \| p) = \sum_{i \in [N]} w_i \log\left(\frac{w_i}{p_i}\right).
\]
Hedge as Bayesian Inference

\[ w_i(t) \propto \exp \left\{ -\eta(t)L_i(t - 1) \right\}. \]

For \( \eta(t) = \eta \) constant in \( t \), this looks like Bayes rule for a parameter in \([N]\)

- with a flat prior, and
- model likelihood \( \exp\{-\eta \ell_i(t)\} \) for the \( t \)-th observation under parameter \( i \).

**Variational Formulation**

\[ w(t) = \arg \min_{w \in \text{simp}([N])} \left( \langle w, L(t - 1) \rangle + \frac{1}{\eta(t)} \text{KL}(w \| \text{Unif}([N])) \right) \]

**Gibbs Posterior**

\[ \hat{\pi}_t(\theta) = \arg \min_{\hat{\pi} \in \mathcal{M}(\Theta)} \left( \mathbb{E}_{\theta \sim \hat{\pi}} L_\theta(t - 1) + \frac{1}{\eta(t)} \text{KL}(\hat{\pi} \| \pi) \right) \]
Hedge as Bayesian Inference

\[ w_i(t) \propto \exp \{-\eta(t)L_i(t - 1)\} . \]

For \( \eta(t) = \eta \) constant in \( t \), this looks like Bayes rule for a parameter in \([N]\)

- with a flat prior, and
- model likelihood \( \exp \{-\eta\ell_i(t)\} \) for the \( t \)-th observation under parameter \( i \).

**Variational Formulation**

\[
w(t) = \arg\min_{w \in \text{simp}([N])} \left( \langle w, L(t - 1) \rangle + \frac{1}{\eta(t)} \text{KL} (w \| \text{Unif}([N])) \right)\]

**Gibbs Posterior**

\[
\hat{\pi}_t(\theta) = \arg\min_{\hat{\pi} \in \mathcal{M}(\Theta)} \left( \mathbb{E}_{\theta \sim \hat{\pi}} L_\theta(t - 1) + \frac{1}{\eta(t)} \text{KL} (\hat{\pi} \| \pi) \right)
\]

\[
\hat{\pi}_t(\theta) \propto \pi(\theta) \exp\{-\eta(t)L_\theta(t - 1)\}
\]
Relaxing the I.I.D. Assumption
Our Setting: Time-Homogeneous Convex Constraints

Intuition

Experts and observations may collude.
Our Setting: Time-Homogeneous Convex Constraints

Intuition
Experts and observations may collude.
Realizations \((x(t), y(t))\) are sampled from an adversarial conditional distribution.
Our Setting: Time-Homogeneous Convex Constraints

Intuition
Experts and observations may collude.
Realizations \((x(t), y(t))\) are sampled from an adversarial conditional distribution.

Formal Framework
- Fix a convex set of distributions \(\mathcal{D} \subseteq \mathcal{M}(\hat{\mathcal{Y}}^N \times \mathcal{Y})\).
Our Setting: Time-Homogeneous Convex Constraints

Intuition
Experts and observations may collude.
Realizations \((x(t), y(t))\) are sampled from an adversarial conditional distribution.

Formal Framework
- Fix a convex set of distributions \(D \subseteq \mathcal{M}(\hat{Y}^N \times Y)\).
- \((x(t), y(t))\) drawn from an element of \(D\) given the history prior to \(t\).
Our Setting: Time-Homogeneous Convex Constraints

Intuition
Experts and observations may collude.
Realizations \((x(t), y(t))\) are sampled from an adversarial conditional distribution.

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 \(t\).
  - Time-Homogeneous: \(\mathcal{D}\) does not depend on \(t\)
Our Setting: Time-Homogeneous Convex Constraints

Intuition
Experts and observations may collude.
Realizations \((x(t), y(t))\) are sampled from an adversarial conditional distribution.

Formal Framework
- Fix a convex set of distributions \(\mathcal{D} \subseteq \mathcal{M}(\hat{Y}^N \times Y)\).
- \((x(t), y(t))\) drawn from an element of \(\mathcal{D}\) given the history prior to \(t\).
  - Time-Homogeneous: \(\mathcal{D}\) does not depend on \(t\)
  - Convexity \(\iff\) environment can flip a coin to select between basic elements of \(\mathcal{D}\)
Our Setting: Time-Homogeneous Convex Constraints

Intuition
Experts and observations may collude. Realizations \((x(t), y(t))\) are sampled from an adversarial conditional distribution.

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 \(t\).
  - Time-Homogeneous: \(\mathcal{D}\) does not depend on \(t\)
  - Convexity \(\Leftrightarrow\) environment can flip a coin to select between basic elements of \(\mathcal{D}\)
  - Environment may aim to maximize regret subject to the constraint
**Our Setting: Time-Homogeneous Convex Constraints**

**Intuition**
Experts and observations may collude. Realizations \((x(t), y(t))\) are sampled from an adversarial conditional distribution.

**Formal Framework**
- Fix a convex set of distributions \(\mathcal{D} \subseteq \mathcal{M}(\hat{Y}^N \times Y)\).
- \((x(t), y(t))\) drawn from an element of \(\mathcal{D}\) given the history prior to \(t\).
  - **Time-Homogeneous**: \(\mathcal{D}\) does not depend on \(t\)
  - Convexity \(\Leftrightarrow\) environment can flip a coin to select between basic elements of \(\mathcal{D}\).
  - Environment may aim to maximize regret subject to the constraint.
- The choice of distribution is made based on outcomes of the previous rounds.
Our Setting: Time-Homogeneous Convex Constraints

Intuition
Experts and observations may collude.
Realizations \((x(t), y(t))\) are sampled from an adversarial conditional distribution.

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 \(t\).
  - Time-Homogeneous: \(\mathcal{D}\) does not depend on \(t\)
  - Convexity \(\iff\) environment can flip a coin to select between basic elements of \(\mathcal{D}\)
  - Environment may aim to maximize regret subject to the constraint
- The choice of distribution is made based on outcomes of the previous rounds.
Our Setting: Time-Homogeneous Convex Constraints

Intuition
Experts and observations may collude. Realizations \((x(t), y(t))\) are sampled from an adversarial conditional distribution.

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 \(t\).
  - Time-Homogeneous: \(\mathcal{D}\) does not depend on \(t\)
  - Convexity \(\iff\) environment can flip a coin to select between basic elements of \(\mathcal{D}\)
  - Environment may aim to maximize regret subject to the constraint
- The choice of distribution is made based on outcomes of the previous rounds.
Examples

Stochastic: \( D = \{ \mu_0 \} \),

Adversarial: \( D = \mathcal{M}(\hat{Y}_N \times Y) \)

Adversarial-with-an-E-gap (Mourtada and Gaïffas 2019)

• One expert has at least \( \Delta > 0 \) less \( E \) loss than the rest on every round.

Neighborhood-of-I.I.D.

• Fix a metric on the space of distributions over \( \hat{Y}_N \times Y \)
• Pick any \( \mu_0 \), and let \( D \) be a neighborhood of \( \mu_0 \), e.g. \( \text{Ball}(\mu_0, r) \) for \( r > 0 \)
• \( r \to 0 \) gives the stochastic case, specifically I.I.D. \( \mu_0 \).
• \( r \to \infty \) gives adversarial case.
• Smoothly transitions in between as \( r \) varies.

• A small neighborhood leads to a slight relaxation of I.I.D.-ness.
Examples

Stochastic:  $\mathcal{D} = \{\mu_0\}$,

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

**Stochastic:** \( \mathcal{D} = \{ \mu_0 \} \),

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

**Stochastic:**  \( \mathcal{D} = \{ \mu_0 \} \),

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

**Adversarial-with-an-\( \mathbb{E} \)-gap** (Mourtada and Gaïffas 2019)
Examples

Stochastic: \( \mathcal{D} = \{ \mu_0 \} \),

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

Adversarial-with-an-\( \mathbb{E} \)-gap (Mourtada and Gaïffas 2019)
- One expert has at least \( \Delta > 0 \) less \( \mathbb{E} \) loss than the rest on every round.
Examples

**Stochastic:** $\mathcal{D} = \{\mu_0\}$,

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

**Adversarial-with-an-\(\mathbb{E}\)-gap** (Mourtada and Gaïffas 2019)
- One expert has at least $\Delta > 0$ less $\mathbb{E}$ loss than the rest on every round.

**Neighborhood-of-I.I.D.**
Examples

Stochastic: \( \mathcal{D} = \{ \mu_0 \} \),

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

Adversarial-with-an-\( \mathbb{E} \)-gap (Mourtada and Gaïffas 2019)

- One expert has at least \( \Delta > 0 \) less \( \mathbb{E} \) loss than the rest on every round.

Neighborhood-of-I.I.D.

- Fix a metric on the space of distributions over \( \hat{Y}^N \times \mathcal{Y} \)
Examples

Stochastic: \( \mathcal{D} = \{\mu_0\} \),

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

Adversarial-with-an-\( \mathbb{E} \)-gap (Mourtada and Gaïffas 2019)

- One expert has at least \( \Delta > 0 \) less \( \mathbb{E} \) loss than the rest on every round.

Neighborhood-of-I.I.D.

- Fix a metric on the space of distributions over \( \hat{\mathcal{Y}}^N \times \mathcal{Y} \)
- Pick any \( \mu_0 \), and let \( \mathcal{D} \) be a neighborhood of \( \mu_0 \),
Examples

**Stochastic:** \( \mathcal{D} = \{ \mu_0 \} \),

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

**Adversarial-with-an-\( \mathbb{E} \)-gap** (Mourtada and Gaïffas 2019)

- One expert has at least \( \Delta > 0 \) less \( \mathbb{E} \) loss than the rest on every round.

**Neighborhood-of-I.I.D.**

- Fix a metric on the space of distributions over \( \hat{Y}^N \times \mathcal{Y} \)
- Pick any \( \mu_0 \), and let \( \mathcal{D} \) be a neighborhood of \( \mu_0 \), e.g. \( \text{Ball}(\mu_0, r) \) for \( r > 0 \)
Examples

**Stochastic:** \( D = \{\mu_0\} \),

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

**Adversarial-with-an-\( \mathbb{E} \)-gap** (Mourtada and Gaïffas 2019)
- One expert has at least \( \Delta > 0 \) less \( \mathbb{E} \) loss than the rest on every round.

**Neighborhood-of-I.I.D.**
- Fix a metric on the space of distributions over \( \hat{Y}^N \times \mathcal{Y} \)
- Pick any \( \mu_0 \), and let \( D \) be a neighborhood of \( \mu_0 \), e.g. Ball(\( \mu_0 \), \( r \)) for \( r > 0 \)
- \( r \to 0 \) gives the stochastic case, specifically I.I.D. \( \mu_0 \).
Examples

**Stochastic:** \( \mathcal{D} = \{ \mu_0 \} \),

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

**Adversarial-with-an-\( \mathbb{E} \)-gap** (Mourtada and Gaïffas 2019)
- One expert has at least \( \Delta > 0 \) less \( \mathbb{E} \) loss than the rest on every round.

**Neighborhood-of-I.I.D.**
- Fix a metric on the space of distributions over \( \hat{Y}^N \times \mathcal{Y} \)
- Pick any \( \mu_0 \), and let \( \mathcal{D} \) be a neighborhood of \( \mu_0 \), e.g. \( \text{Ball}(\mu_0, r) \) for \( r > 0 \)
- \( r \rightarrow 0 \) gives the stochastic case, specifically I.I.D. \( \mu_0 \).
- \( r \rightarrow \infty \) gives adversarial case.
Examples

**Stochastic:** $\mathcal{D} = \{\mu_0\},$

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

**Adversarial-with-an-$\mathbb{E}$-gap** (Mourtada and Gaïffas 2019)
- One expert has at least $\Delta > 0$ less $\mathbb{E}$ loss than the rest on every round.

**Neighborhood-of-I.I.D.**
- Fix a metric on the space of distributions over $\mathcal{Y}^N \times \mathcal{Y}$
- Pick any $\mu_0$, and let $\mathcal{D}$ be a neighborhood of $\mu_0$, e.g. $\text{Ball}(\mu_0, r)$ for $r > 0$
- $r \to 0$ gives the stochastic case, specifically I.I.D. $\mu_0$.
- $r \to \infty$ gives adversarial case. Smoothly transitions in between as $r$ varies.
Examples

Stochastic: \( \mathcal{D} = \{ \mu_0 \} \),

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

Adversarial-with-an-\( \mathbb{E} \)-gap (Mourtada and Gaïffas 2019)
- One expert has at least \( \Delta > 0 \) less \( \mathbb{E} \) loss than the rest on every round.

Neighborhood-of-I.I.D.
- Fix a metric on the space of distributions over \( \hat{Y}^N \times \mathcal{Y} \)
- Pick any \( \mu_0 \), and let \( \mathcal{D} \) be a neighborhood of \( \mu_0 \), e.g. \( \text{Ball}(\mu_0, r) \) for \( r > 0 \)
- \( r \to 0 \) gives the stochastic case, specifically I.I.D. \( \mu_0 \).
- \( r \to \infty \) gives adversarial case. Smoothly transitions in between as \( r \) varies.
- A small neighborhood leads to a slight relaxation of I.I.D.-ness.
Examples

**Stochastic:** \( D = \{ \mu_0 \} \),

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

**Adversarial-with-an-\( \mathbb{E} \)-gap** (Mourtada and Gaïffas 2019)

- One expert has at least \( \Delta > 0 \) less \( \mathbb{E} \) loss than the rest on every round.

**Neighborhood-of-I.I.D.**

- Fix a metric on the space of distributions over \( \hat{Y}^N \times \mathcal{Y} \)
- Pick any \( \mu_0 \), and let \( D \) be a neighborhood of \( \mu_0 \), e.g. \( \text{Ball}(\mu_0, r) \) for \( r > 0 \)
- \( r \to 0 \) gives the stochastic case, specifically I.I.D. \( \mu_0 \).
- \( r \to \infty \) gives adversarial case. Smoothly transitions in between as \( r \) varies.
- A small neighborhood leads to a slight relaxation of I.I.D.-ness.
Constraint-Characterizing Quantities

We use quantities to characterize the constraint that:

Effective Experts

$I_0 = \{ \text{experts that are optimal in } E \text{ for some } \mu \in D \}$

$N_0 = |I_0|$  

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

Effective Stochastic Gap

$\Delta_0 = \inf_{\mu \in D} \{ \mu - \text{expected difference in loss of best expert and best expert not in } I_0 \}$

Analogous to the gap in the stochastic-with-a-gap setting.
Constraint-Characterizing Quantities

We use quantities to characterize the constraint that:

- are representative of whether the data is “easy” or not;
Constraint-Characterizing Quantities

We use quantities to characterize the constraint that:

- are representative of whether the data is “easy” or not;
- yield matching lower and upper bounds on regret.
We use quantities to characterize the constraint that:

- are representative of whether the data is “easy” or not;
- yield matching lower and upper bounds on regret.

Effective Experts
Constraint-Characterizing Quantities

We use quantities to characterize the constraint that:

- are representative of whether the data is “easy” or not;
- yield matching lower and upper bounds on regret.

**Effective Experts**

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

$$N_0 = |\mathcal{I}_0|$$
Constraint-Characterizing Quantities

We use quantities to characterize the constraint that:

- are representative of whether the data is “easy” or not;
- yield matching lower and upper bounds on regret.

Effective Experts

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

Analogous to the single best expert in the stochastic-with-a-gap setting.
We use quantities to characterize the constraint that:

- are representative of whether the data is “easy” or not;
- yield matching lower and upper bounds on regret.

**Effective Experts**

\[ I_0 = \{ \text{experts that are optimal in } E \text{ for some } \mu \in D \} \]

\[ N_0 = |I_0| \]

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

**Effective Stochastic Gap**

\[ \Delta_0 = \inf_{\mu \in D} \{ \mu - \text{expected difference in loss of best expert and best expert not in } I_0 \} \]

Analogous to the gap in the stochastic-with-a-gap setting.
Constraint-Characterizing Quantities

We use quantities to characterize the constraint that:

- are representative of whether the data is “easy” or not;
- yield matching lower and upper bounds on regret.

**Effective Experts**

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

\[ N_0 = |\mathcal{I}_0| \]

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

**Effective Stochastic Gap**

\[ \Delta_0 = \inf_{\mu \in \mathcal{D}} \{ \mu\text{-expected difference in loss of best expert and best expert not in } \mathcal{I}_0 \} \]
Constraint-Characterizing Quantities

We use quantities to characterize the constraint that:

- are representative of whether the data is “easy” or not;
- yield matching lower and upper bounds on regret.

Effective Experts

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

\[
N_0 = |\mathcal{I}_0|
\]

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

Effective Stochastic Gap

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

Analogous to the gap in the stochastic-with-a-gap setting.
Constraint-Characterizing Quantities

We use quantities to characterize the constraint that:

- are representative of whether the data is “easy” or not;
- yield matching lower and upper bounds on regret.

Effective Experts

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

\[ N_0 = |\mathcal{I}_0| \]

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

Effective Stochastic Gap

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

Analogous to the gap in the stochastic-with-a-gap setting.
Constraint Example

\[ \mathcal{I}_0 = \{ \text{experts that are optimal for some } \mu \in \mathcal{D} \} \quad N_0 = | \mathcal{I}_0 | \]

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

**Setting:** the means for each expert are jointly defined by a parameter \( \alpha \),

\( N = 5, \quad \mathcal{I}_0 = \{ 1, 3, 5 \}, \quad N_0 = 3. \)
Constraint Example

\[ I_0 = \{ \text{experts that are optimal for some } \mu \in D \} \quad N_0 = |I_0| \]

\[ \Delta_0 = \inf_{\mu \in D} \{ \mu \text{-expected difference in loss of best expert and best expert not in } I_0 \} \]

**Setting:** the means for each expert are jointly defined by a parameter \( \alpha \),

\( N = 5, I_0 = \{1, 3, 5\}, N_0 = 3 \).
\[ I_0 = \{ \text{experts that are optimal for some } \mu \in D \} \quad N_0 = |I_0| \]

\[ \Delta_0 = \inf_{\mu \in D} \{ \mu \text{-expected difference in loss of best expert and best expert not in } I_0 \} \]

**Setting:** the means for each expert are jointly defined by a parameter \( \alpha \),

\( N = 5, \ I_0 = \{1, 3, 5\}, \ N_0 = 3 \).
**Constraint Example**

\[ I_0 = \{ \text{experts that are optimal for some } \mu \in \mathcal{D} \} \quad N_0 = |I_0| \]

\[ \Delta_0 = \inf_{\mu \in \mathcal{D}} \{ \mu\text{-expected difference in loss of best expert and best expert not in } I_0 \} \]

**Setting:** the means for each expert are jointly defined by a parameter \( \alpha \),

\( N = 5, I_0 = \{1, 3, 5\}, N_0 = 3. \)
Constraint Example

\( \mathcal{I}_0 = \{ \text{experts that are optimal for some } \mu \in \mathcal{D} \} \) \quad \text{\( N_0 = |\mathcal{I}_0| \)}

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

**Setting:** the means for each expert are jointly defined by a parameter \( \alpha \),

\( N = 5, \mathcal{I}_0 = \{1, 3, 5\}, N_0 = 3. \)
\[ I_0 = \{ \text{experts that are optimal for some } \mu \in \mathcal{D} \} \quad N_0 = |I_0| \]

\[ \Delta_0 = \inf_{\mu \in \mathcal{D}} \{ \mu \text{-expected difference in loss of best expert and best expert not in } I_0 \} \]

**Setting:** the means for each expert are jointly defined by a parameter \( \alpha \),

\( N = 5, \ I_0 = \{1, 3, 5\}, \ N_0 = 3. \)
Constraint Example

\[ \mathcal{I}_0 = \{ \text{experts that are optimal for some } \mu \in \mathcal{D} \} \quad N_0 = |\mathcal{I}_0| \]

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

**Setting:** the means for each expert are jointly defined by a parameter \( \alpha \),

\( N = 5, \mathcal{I}_0 = \{1, 3, 5\}, N_0 = 3. \)
\[ I_0 = \{ \text{experts that are optimal for some } \mu \in D \} \quad N_0 = |I_0| \]

\[ \Delta_0 = \inf_{\mu \in D} \{ \mu \text{-expected difference in loss of best expert and best expert not in } I_0 \} \]

**Setting:** the means for each expert are jointly defined by a parameter \( \alpha \),
\[ N = 5, \ I_0 = \{1, 3, 5\}, \ N_0 = 3. \]
$(N_0, \Delta_0)$ for Examples

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

- $\mathcal{N}_0 = 1$,
Stochastic-with-a-gap:  $\mathcal{D} = \{\mu_0\}$,

- $N_0 = 1$,  $\mathcal{I}_0 = \left\{ i^* = \arg \min_{i \in [N]} E_{\mu_0}[\ell_i] \right\}$,
(N₀, Δ₀) for Examples

**Stochastic-with-a-gap:** \( \mathcal{D} = \{\mu_0\} \),

- \( N_0 = 1 \), \( \mathcal{I}_0 = \left\{ i^* = \arg \min_{i \in [N]} \mathbb{E}_{\mu_0}[\ell_i] \right\} \), \( \Delta_0 = \min_{i \neq i^*} \mathbb{E}_{\mu_0}[\ell_i - \ell_{i^*}] \)
\((N_0, \Delta_0)\) for Examples

**Stochastic-with-a-gap:** \(\mathcal{D} = \{\mu_0\}\),
- \(N_0 = 1\), \(\mathcal{I}_0 = \left\{ i^* = \arg \min_{i \in [N]} \mathbb{E}_{\mu_0}[\ell_i] \right\}\), \(\Delta_0 = \min_{i \neq i^*} \mathbb{E}_{\mu_0}[\ell_i - \ell_{i^*}]\)

**Adversarial:** \(\mathcal{D} = \mathcal{M}(\hat{\mathcal{Y}}^N \times \mathcal{Y})\)
\((N_0, \Delta_0)\) for Examples

**Stochastic-with-a-gap:**  \(D = \{\mu_0\},\)

- \(N_0 = 1, \ I_0 = \left\{ i^* = \arg\min_{i \in [N]} E_{\mu_0}[\ell_i] \right\}, \ \Delta_0 = \min_{i \neq i^*} E_{\mu_0}[\ell_i - \ell_{i^*}]\)

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

- \(N_0 = N,\)
\((N_0, \Delta_0)\) for Examples

**Stochastic-with-a-gap:** \( \mathcal{D} = \{\mu_0\} \),
- \( N_0 = 1 \), \( I_0 = \left\{ i^* = \arg\min_{i \in [N]} \mathbb{E}_{\mu_0}[\ell_i] \right\} \), \( \Delta_0 = \min_{i \neq i^*} \mathbb{E}_{\mu_0}[\ell_i - \ell_{i^*}] \)

**Adversarial:** \( \mathcal{D} = \mathcal{M}(\hat{\mathcal{Y}}^N \times \mathcal{Y}) \)
- \( N_0 = N \), \( \Delta_0 = +\infty \)
(\mathcal{N}_0, \Delta_0) \text{ for Examples}

**Stochastic-with-a-gap:** \( \mathcal{D} = \{ \mu_0 \} \),

- \( \mathcal{N}_0 = 1, \quad \mathcal{I}_0 = \left\{ i^* = \arg \min_{i \in [\mathcal{N}]} \mathbb{E}_{\mu_0}[\ell_i] \right\} \), \( \Delta_0 = \min_{i \neq i^*} \mathbb{E}_{\mu_0}[\ell_i - \ell_{i^*}] \)

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

- \( \mathcal{N}_0 = \mathcal{N}, \quad \Delta_0 = +\infty \)

**Adversarial-with-an-\( \mathbb{E} \)-gap**
Stochastic-with-a-gap: \( D = \{ \mu_0 \} \),
- \( N_0 = 1 \), \( I_0 = \left\{ i^* = \arg \min_{i \in [N]} \mathbb{E}_{\mu_0} [ \ell_i ] \right\} \), \( \Delta_0 = \min_{i \neq i^*} \mathbb{E}_{\mu_0} [ \ell_i - \ell_{i^*} ] \)

Adversarial: \( D = \mathcal{M}(\hat{Y}^N \times Y) \)
- \( N_0 = N \), \( \Delta_0 = +\infty \)

Adversarial-with-an-\( \mathbb{E} \)-gap
- All measures where a common expert is better than others in \( \mathbb{E} \) by \( \Delta > 0 \).
\((N_0, \Delta_0)\) for Examples

**Stochastic-with-a-gap:** \(\mathcal{D} = \{\mu_0\},\)
- \(N_0 = 1, \ I_0 = \{i^* = \arg \min_{i \in [N]} \mathbb{E}_{\mu_0}[\ell_i]\}, \ \Delta_0 = \min_{i \neq i^*} \mathbb{E}_{\mu_0}[\ell_i - \ell_{i^*}]\)

**Adversarial:** \(\mathcal{D} = \mathcal{M}(\hat{Y}^N \times \mathcal{Y})\)
- \(N_0 = N, \ \Delta_0 = +\infty\)

**Adversarial-with-an-\(\mathbb{E}\)-gap**
- All measures where a common expert is better than others in \(\mathbb{E}\) by \(\Delta > 0\).
- By design, \(N_0 = 1\) and \(\Delta_0 = \Delta\).
Examples

Stochastic-with-a-gap: \( D = \{ \mu_0 \} \),
- \( N_0 = 1 \), \( I_0 = \left\{ i^* = \text{arg min}_{i \in [N]} \mathbb{E}_{\mu_0}[\ell_i] \right\} \), \( \Delta_0 = \min_{i \neq i^*} \mathbb{E}_{\mu_0}[\ell_i - \ell_{i^*}] \)

Adversarial: \( D = \mathcal{M}(\hat{\mathcal{Y}}^N \times \mathcal{Y}) \)
- \( N_0 = N \), \( \Delta_0 = +\infty \)

Adversarial-with-an-\( \mathbb{E} \)-gap
- All measures where a common expert is better than others in \( \mathbb{E} \) by \( \Delta > 0 \).
- By design, \( N_0 = 1 \) and \( \Delta_0 = \Delta \).

Neighborhood-of-I.I.D.
\((N_0, \Delta_0)\) for Examples

**Stochastic-with-a-gap:** \(\mathcal{D} = \{\mu_0\}\),
- \(N_0 = 1, \mathcal{I}_0 = \left\{ i^* = \arg\min_{i \in [N]} E_{\mu_0}[\ell_i] \right\}, \quad \Delta_0 = \min_{i \neq i^*} E_{\mu_0}[\ell_i - \ell_{i^*}]\)

**Adversarial:** \(\mathcal{D} = \mathcal{M}(\hat{\mathcal{Y}}^N \times \mathcal{Y})\)
- \(N_0 = N, \quad \Delta_0 = +\infty\)

**Adversarial-with-an-\(E\)-gap**
- All measures where a common expert is better than others in \(E\) by \(\Delta > 0\).
- By design, \(N_0 = 1\) and \(\Delta_0 = \Delta\).

**Neighborhood-of-I.I.D.**
- Pick any distribution \(\mu_0\), and any radius, \(r \geq 0\). \(\mathcal{D} = \text{Ball}(\mu_0, r)\)
Stochastic-with-a-gap: $\mathcal{D} = \{\mu_0\}$,
- $N_0 = 1$, $\mathcal{I}_0 = \left\{ i^* = \arg\min_{i \in [N]} \mathbb{E}_{\mu_0}[\ell_i] \right\}$, $\Delta_0 = \min_{i \neq i^*} \mathbb{E}_{\mu_0}[\ell_i - \ell_{i^*}]$

Adversarial: $\mathcal{D} = \mathcal{M}(\hat{\mathcal{Y}}^N \times \mathcal{Y})$
- $N_0 = N$, $\Delta_0 = +\infty$

Adversarial-with-an-$\mathbb{E}$-gap
- All measures where a common expert is better than others in $\mathbb{E}$ by $\Delta > 0$.
- By design, $N_0 = 1$ and $\Delta_0 = \Delta$.

Neighborhood-of-I.I.D.
- Pick any distribution $\mu_0$, and any radius, $r \geq 0$. $\mathcal{D} = \text{Ball}(\mu_0, r)$
- Suppose that $\mu_0$ has a gaps between all the mean losses.
\((N_0, \Delta_0)\) for Examples

**Stochastic-with-a-gap:** \(\mathcal{D} = \{\mu_0\},\)

- \(N_0 = 1, \quad I_0 = \left\{ i^* = \arg\min_{i \in [N]} \mathbb{E}_{\mu_0}[\ell_i] \right\}, \quad \Delta_0 = \min_{i \neq i^*} \mathbb{E}_{\mu_0}[\ell_i - \ell_{i^*}]\)

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

- \(N_0 = N, \quad \Delta_0 = +\infty\)

**Adversarial-with-an-\(E\)-gap**

- All measures where a common expert is better than others in \(\mathbb{E}\) by \(\Delta > 0\).
- By design, \(N_0 = 1\) and \(\Delta_0 = \Delta\).

**Neighborhood-of-I.I.D.**

- Pick any distribution \(\mu_0\), and any radius, \(r \geq 0\). \(\mathcal{D} = \text{Ball}(\mu_0, r)\)
- Suppose that \(\mu_0\) has a gaps between all the mean losses.
- \(N_0, \Delta_0\) depend on the radius of the ball...
\((\mathcal{N}_0, \Delta_0)\) for Examples

**Stochastic-with-a-gap:** \(\mathcal{D} = \{\mu_0\},\)
- \(\mathcal{N}_0 = 1, \ I_0 = \{i^* = \arg\min_{i \in [N]} \mathbb{E}_{\mu_0}[\ell_i]\}, \ \Delta_0 = \min_{i \neq i^*} \mathbb{E}_{\mu_0}[\ell_i - \ell_{i^*}]\)

**Adversarial:** \(\mathcal{D} = \mathcal{M}((\hat{Y}^N \times Y))\)
- \(\mathcal{N}_0 = N, \ \Delta_0 = +\infty\)

**Adversarial-with-an-\(\mathbb{E}\)-gap**
- All measures where a common expert is better than others in \(\mathbb{E}\) by \(\Delta > 0\).
- By design, \(\mathcal{N}_0 = 1\) and \(\Delta_0 = \Delta\).

**Neighborhood-of-I.I.D.**
- Pick any distribution \(\mu_0\), and any radius, \(r \geq 0\). \(\mathcal{D} = \text{Ball}(\mu_0, r)\)
- Suppose that \(\mu_0\) has a gaps between all the mean losses.
- \(\mathcal{N}_0, \ \Delta_0\) depend on the radius of the ball...
Interpreting \((N_0, \Delta_0^{-1})\)

Minimax Regret

\[
\mathbb{E}R(T) \lesssim \begin{cases} 
\sqrt{T \log N_0} & : N_0 \geq 2 \\
(\log N)/\Delta_0 & : N_0 = 1.
\end{cases}
\]
Interpreting \((N_0, \Delta_0^{-1})\)

Minimax Regret

\[
\mathbb{E}R(T) \preceq \begin{cases} 
\sqrt{T \log N_0} & : N_0 \geq 2 \\
\frac{(\log N)}{\Delta_0} & : N_0 = 1.
\end{cases}
\]

\(\mathcal{D} = \text{Ball}(\mu, \text{radius}) \) w/ \(\mathbb{E}_\mu \ell_1 < \mathbb{E}_\mu \ell_2 < \ldots\)
Interpreting \((N_0, \Delta_0^{-1})\)

Minimax Regret

\[
\mathbb{E} R(T) \preceq \begin{cases} 
\sqrt{T \log N_0} & : N_0 \geq 2 \\
\frac{(\log N)}{\Delta_0} & : N_0 = 1.
\end{cases}
\]

\[D = \text{Ball}(\mu, \text{radius}) \text{ w/ } \mathbb{E}_\mu \ell_1 < \mathbb{E}_\mu \ell_2 < \ldots\]
Interpreting \((N_0, \Delta_0^{-1})\)

Minimax Regret

\[
\mathbb{E} R(T) \lesssim \begin{cases} 
\sqrt{T \log N_0} & : N_0 \geq 2 \\
(\log N)/\Delta_0 & : N_0 = 1.
\end{cases}
\]

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

- \(N_0\) non-decreasing with radius
Minimax Regret

\[ \mathbb{E}R(T) \begin{cases} \sqrt{T \log N_0} & : N_0 \geq 2 \\ \frac{(\log N)}{\Delta_0} & : N_0 = 1. \end{cases} \]

\[ \mathcal{D} = \text{Ball}(\mu, \text{radius}) \text{ w/ } \mathbb{E}_\mu \ell_1 < \mathbb{E}_\mu \ell_2 < \ldots \]

- \( N_0 \) non-decreasing with radius
- \( N_0 \) increases discretely
Interpreting \((N_0, \Delta_0^{-1})\)

Minimax Regret

\[
\mathbb{E} R(T) \sim \begin{cases} \sqrt{T \log N_0} & : N_0 \geq 2 \\ (\log N)/\Delta_0 & : N_0 = 1. \end{cases}
\]

\[D = \text{Ball}(\mu, \text{radius}) \text{ w/ } \mathbb{E}_\mu \ell_1 < \mathbb{E}_\mu \ell_2 < \ldots\]

- \(N_0\) non-decreasing with radius
- \(N_0\) increases discretely
Interpreting \((N_0, \Delta_0^{-1})\)

Minimax Regret

\[
\mathbb{E}R(T) \preceq \begin{cases} 
\sqrt{T \log N_0} &: N_0 \geq 2 \\
(\log N)/\Delta_0 &: N_0 = 1.
\end{cases}
\]

\[D = \text{Ball}(\mu, \text{radius}) \text{ w/ } \mathbb{E}_\mu \ell_1 < \mathbb{E}_\mu \ell_2 < \ldots\]

- \(N_0\) non-decreasing with radius
- \(N_0\) increases discretely
- \(\Delta_0^{-1}\) increases between jumps in \(N_0\)
Interpreting \((N_0, \Delta_0^{-1})\)

Minimax Regret

\[
\mathbb{E}R(T) \approx \begin{cases} 
\sqrt{T \log N_0} & : N_0 \geq 2 \\
(\log N)/\Delta_0 & : N_0 = 1.
\end{cases}
\]

\[\mathcal{D} = \text{Ball}(\mu, \text{radius}) \text{ w/ } \mathbb{E}_\mu \ell_1 < \mathbb{E}_\mu \ell_2 < \ldots\]

- \(N_0\) non-decreasing with radius
- \(N_0\) increases discretely
- \(\Delta_0^{-1}\) increases between jumps in \(N_0\)
- \(\Delta_0^{-1}\) resets each time \(N_0\) increases
Interpreting \((N_0, \Delta_0^{-1})\)

Minimax Regret

\[
\mathbb{E} R(T) \simeq \begin{cases} \sqrt{T \log N_0} : N_0 \geq 2 \\ \frac{(\log N)}{\Delta_0} : N_0 = 1. \end{cases}
\]

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

- \(N_0\) non-decreasing with radius
- \(N_0\) increases discretely
- \(\Delta_0^{-1}\) increases between jumps in \(N_0\)
- \(\Delta_0^{-1}\) resets each time \(N_0\) increases

Lexicographical order on \((N_0, \Delta_0^{-1})\) respects “\(\subseteq\)” for nested \(D\)s.
Interpreting \((N_0, \Delta_0^{-1})\)

Minimax Regret

\[
\mathbb{E} R(T) \approx \begin{cases} 
\sqrt{T \log N_0} & : N_0 \geq 2 \\
(log N) / \Delta_0 & : N_0 = 1.
\end{cases}
\]

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

- \(N_0\) non-decreasing with radius
- \(N_0\) increases discretely
- \(\Delta_0^{-1}\) increases between jumps in \(N_0\)
- \(\Delta_0^{-1}\) resets each time \(N_0\) increases

Lexicographical order on \((N_0, \Delta_0^{-1})\) respects “\(\subseteq\)” for nested \(\mathcal{D}\)s.

- For nested \(\mathcal{D}\)s, larger one is harder.
Interpreting \((N_0, \Delta_0^{-1})\)

Minimax Regret

\[
\mathbb{E} R(T) \leq \begin{cases} 
\sqrt{T \log N_0} & : N_0 \geq 2 \\
(\log N) / \Delta_0 & : N_0 = 1.
\end{cases}
\]

\[D = \text{Ball}(\mu, \text{radius}) \text{ w/ } \mathbb{E}_\mu \ell_1 < \mathbb{E}_\mu \ell_2 < \ldots\]

- \(N_0\) non-decreasing with radius
- \(N_0\) increases discretely
- \(\Delta_0^{-1}\) increases between jumps in \(N_0\)
- \(\Delta_0^{-1}\) resets each time \(N_0\) increases

Lexicographical order on \((N_0, \Delta_0^{-1})\) respects “\(\subseteq\)” for nested \(D\)s.

- For nested \(D\)s, larger one is harder.
- \((N_0, \Delta_0^{-1})\) quantifies the difficulty of \(D\)
Performance of Hedge
Hedge Regret Bounds

Consider playing Hedge with $\eta(t) = c/\sqrt{t}$ for any convex $D$.

Recall:
- $N_0$ is the number of effective experts,
- $\Delta_0$ is the effective stochastic gap.
Hedge Regret Bounds

Consider playing Hedge with \( \eta(t) = c / \sqrt{t} \) for any convex \( D \).

Recall:
- \( N_0 \) is the number of effective experts,
- \( \Delta_0 \) is the effective stochastic gap.

**Theorem BNR20**

Taking \( c \propto \sqrt{\log N} \) gives

\[
\mathbb{E} R(T) \lesssim \begin{cases} 
\sqrt{T \log N} + \frac{\log N}{\Delta_0} & : N_0 \geq 2 \\
(\log N)/\Delta_0 & : N_0 = 1.
\end{cases}
\]

Taking \( c \propto 1 \) gives

\[
\mathbb{E} R(T) \lesssim (\log N_0) \sqrt{T} + \frac{(\log N)^2}{\Delta_0}
\]

We also prove matching lower bounds!
Hedge Regret Bounds

Consider playing Hedge with $\eta(t) = c/\sqrt{t}$ for any convex $D$.

Recall:
- $N_0$ is the number of effective experts,
- $\Delta_0$ is the effective stochastic gap.

**Theorem BNR20**

If the player has oracle knowledge of $N_0 > 1$, taking $c \propto \sqrt{\log (N_0)}$ gives

$$\mathbb{E}R(T) \lesssim \sqrt{T \log N_0} + \frac{(\log N)^2}{(\log N_0)\Delta_0}.$$

- $T$: long run regret accumulation after adapting
- $(\log N, \Delta_0)$: adversarial regret over adaptation period of duration $O((\log N)^2 c^2 \Delta_0^2)$
Hedge Regret Bounds

Consider playing Hedge with $\eta(t) = c/\sqrt{t}$ for any convex $\mathcal{D}$.

Recall:
- $N_0$ is the number of effective experts,
- $\Delta_0$ is the effective stochastic gap.

**Theorem BNR20**

If the player has oracle knowledge of $N_0 > 1$, taking $c \propto \sqrt{\log(N_0)}$ gives

$$\mathbb{E}R(T) \lesssim \sqrt{T \log N_0} + \frac{(\log N)^2}{(\log N_0)\Delta_0}.$$ 

In all three cases, we interpret terms involving...
Consider playing Hedge with $\eta(t) = c/\sqrt{t}$ for any convex $D$.

Recall:
- $N_0$ is the number of effective experts,
- $\Delta_0$ is the effective stochastic gap.

**Theorem BNR20**

If the player has oracle knowledge of $N_0 > 1$, taking $c \propto \sqrt{\log(N_0)}$ gives

$$\mathbb{E}R(T) \lesssim \sqrt{T \log N_0} + \frac{(\log N)^2}{(\log N_0)\Delta_0}.$$ 

In all three cases, we interpret terms involving...
- $T$: long run regret accumulation after adapting
Hedge Regret Bounds

Consider playing Hedge with $\eta(t) = c/\sqrt{t}$ for any convex $D$.

Recall:
- $N_0$ is the number of effective experts,
- $\Delta_0$ is the effective stochastic gap.

**Theorem BNR20**

If the player has oracle knowledge of $N_0 > 1$, taking $c \propto \sqrt{\log(N_0)}$ gives

$$\mathbb{E}R(T) \lesssim \sqrt{T \log N_0} + \frac{(\log N)^2}{(\log N_0)\Delta_0}.$$

In all three cases, we interpret terms involving...
- $T$: long run regret accumulation after adapting
- $(\log N, \Delta_0)$: adversarial regret over adaptation period of duration $\mathcal{O}\left(\frac{(\log N)^2}{c^2 \Delta_0^2}\right)$
Question: If we don’t know \( N_0 \), can we learn adaptively and minimax optimally?

Answer: Yes! We introduce two new algorithms in order to do this...

- FTRL-CARE, accomplished 1 and 3, but not 2.
- Meta-CARE, accomplished all 3 by boosting FTRL-CARE with Hedge.
Can we do better than Hedge?

**Question:** If we don’t know $N_0$, can we learn adaptively and minimax optimally? In particular, is there an algorithm for which...

- $(T, N_0)$-dependence matches Hedge with oracle knowledge of $N_0$,
- $(\log N, \Delta_0)$-dependence optimal for the stochastic case when $N_0 = 1$, and
- no information is needed about the true setting?

**Answer:** Yes! We introduce two new algorithms in order to do this...

- **FTRL-CARE**, accomplished 1 and 3, but not 2.
- slightly worse dependence on $N$.
- **Meta-CARE**, accomplished all 3 by boosting FTRL-CARE with Hedge.
Can we do better than Hedge?

Question: If we don’t know \( N_0 \), can we learn adaptively and minimax optimally? In particular, is there an algorithm for which...

1. \( (T, N_0) \)-dependence matches Hedge with oracle knowledge of \( N_0 \),

Answer: Yes!

We introduce two new algorithms in order to do this...

- FTRL-CARE, accomplished 1 and 3, but not 2.
- Slightly worse dependence on \( N_0 \).
- Meta-CARE, accomplished all 3 by boosting FTRL-CARE with Hedge.
Can we do better than Hedge?

**Question:** If we don’t know \( N_0 \), can we learn adaptively and minimax optimally?

In particular, is there an algorithm for which...

1. \((T, N_0)\)-dependence matches Hedge with oracle knowledge of \( N_0 \),
2. \((\log N, \Delta_0)\)-dependence optimal for the stochastic case when \( N_0 = 1 \), and

Answer: Yes! We introduce two new algorithms in order to do this...

- **FTRL-CARE**, accomplished 1 and 3, but not 2.
- **Meta-CARE**, accomplished all 3 by boosting FTRL-CARE with Hedge.
Can we do better than Hedge?

**Question:** If we don’t know $N_0$, can we learn adaptively and minimax optimally?

In particular, is there an algorithm for which...

1. $(T, N_0)$-dependence matches Hedge with oracle knowledge of $N_0$,
2. $(\log N, \Delta_0)$-dependence optimal for the stochastic case when $N_0 = 1$, and
3. no information is needed about the true setting?

Answer: Yes!

We introduce two new algorithms in order to do this...

- **FTRL-CARE**, accomplished 1 and 3, but not 2.
- Slightly worse dependence on $N$.
- **Meta-CARE**, accomplished all 3 by **boosting** FTRL-CARE with Hedge.
Can we do better than Hedge?

**Question:** If we don’t know \( N_0 \), can we learn adaptively and minimax optimally? In particular, is there an algorithm for which...

1. \((T, N_0)\)-dependence matches Hedge with oracle knowledge of \( N_0 \),
2. \((\log N, \Delta_0)\)-dependence optimal for the stochastic case when \( N_0 = 1 \), and
3. no information is needed about the true setting?

**Answer:** Yes!

We introduce two new algorithms in order to do this...

- FTRL-CARE, accomplished 1 and 3, but not 2.
- Meta-CARE, accomplished all 3 by boosting FTRL-CARE with Hedge.
Can we do better than Hedge?

**Question:** If we don’t know $N_0$, can we learn adaptively and minimax optimally? In particular, is there an algorithm for which...

1. $(T, N_0)$-dependence matches Hedge with oracle knowledge of $N_0$,
2. $(\log N, \Delta_0)$-dependence optimal for the stochastic case when $N_0 = 1$, and
3. no information is needed about the true setting?

**Answer:** Yes!

We introduce two new algorithms in order to do this...
**Can we do better than Hedge?**

**Question:** If we don’t know $N_0$, can we learn adaptively and minimax optimally?

In particular, is there an algorithm for which...

1. $(T, N_0)$-dependence matches Hedge with oracle knowledge of $N_0$,
2. $(\log N, \Delta_0)$-dependence optimal for the stochastic case when $N_0 = 1$, and
3. no information is needed about the true setting?

**Answer:** Yes!

We introduce two new algorithms in order to do this...

- FTRL-CARE, accomplished 1 and 3, but not 2.

- Meta-CARE, accomplished all 3 by boosting FTRL-CARE with Hedge.
Can we do better than Hedge?

**Question:** If we don’t know $N_0$, can we learn adaptively and minimax optimally?

In particular, is there an algorithm for which...

1. $(T, N_0)$-dependence matches Hedge with oracle knowledge of $N_0$,
2. $(\log N, \Delta_0)$-dependence optimal for the stochastic case when $N_0 = 1$, and
3. no information is needed about the true setting?

**Answer:** Yes!

We introduce two new algorithms in order to do this...

- **FTRL-CARE**, accomplished 1 and 3, but not 2.
  - slightly worse dependence on $N$. 
- **Meta-CARE**, accomplished all 3 by boosting FTRL-CARE with Hedge.
Can we do better than Hedge?

**Question:** If we don’t know $N_0$, can we learn adaptively and minimax optimally? In particular, is there an algorithm for which...

1. $(T, N_0)$-dependence matches Hedge with oracle knowledge of $N_0$,
2. $(\log N, \Delta_0)$-dependence optimal for the stochastic case when $N_0 = 1$, and
3. no information is needed about the true setting?

**Answer:** Yes!

We introduce two new algorithms in order to do this...

- **FTRL-CARE**, accomplished 1 and 3, but not 2.
  - slightly worse dependence on $N$.
- **Meta-CARE**, accomplished all 3 by *boosting* FTRL-CARE with Hedge.
Question: If we don’t know $N_0$, can we learn adaptively and minimax optimally?

In particular, is there an algorithm for which...

1. $(T, N_0)$-dependence matches Hedge with oracle knowledge of $N_0$,

2. $(\log N, \Delta_0)$-dependence optimal for the stochastic case when $N_0 = 1$, and

3. no information is needed about the true setting?

Answer: Yes!

We introduce two new algorithms in order to do this...

- FTRL-CARE, accomplished 1 and 3, but not 2.
  - slightly worse dependence on $N$.

- Meta-CARE, accomplished all 3 by boosting FTRL-CARE with Hedge.
Improved Algorithms and Bounds
Intuition for Improving on Hedge

Three Key Insights:

1. From our oracle Hedge bound, we want a learning rate $\alpha \propto \sqrt{\log N_0/t}$.

2. The regret of Hedge closely depends on the entropy of the weights:
   $$H(w) = -\sum_{i \in [N]} w_i \log(w_i).$$

3. Worst-case adversary forces weights to concentrate to $\text{Unif}(I_0)$, so $H(w) \approx \log N_0$. 
Intuition for Improving on Hedge

Three Key Insights:

1. From our oracle Hedge bound, we want a learning rate $\propto \sqrt{\log N_0}/t$. 
Intuition for Improving on Hedge

Three Key Insights:

1. From our oracle Hedge bound, we want a learning rate $\propto \sqrt{\frac{\log N_0}{t}}$.

2. The regret of Hedge closely depends on the entropy of the weights:

$$H(w) = -\sum_{i \in [N]} w_i \log(w_i).$$
Intuition for Improving on Hedge

Three Key Insights:

1. From our oracle Hedge bound, we want a learning rate $\propto \sqrt{\frac{\log N_0}{t}}$.

2. The regret of Hedge closely depends on the entropy of the weights:

   $$H(w) = - \sum_{i \in [N]} w_i \log(w_i).$$

3. Worst-case adversary forces weights to concentrate to $\text{Unif}(I_0)$, so

   $$H(w) \approx \log N_0.$$
Intuition for Improving on Hedge

Three Key Insights:

1. From our oracle Hedge bound, we want a learning rate $\propto \sqrt{\frac{\log N_0}{t}}$.
2. The regret of Hedge closely depends on the *entropy* of the weights:

   $$
   H(w) = -\sum_{i \in [N]} w_i \log(w_i).
   $$

3. Worst-case adversary forces weights to concentrate to $\text{Unif}(I_0)$, so

   $$
   H(w) \approx \log N_0.
   $$

What if we could have our learning rate at time $t$, $\eta(t)$, look like

$$
\eta(t) = \sqrt{\frac{H(w(t))}{t}}.
$$
Follow the Regularized Leader

FTRL is a fundamental online linear optimization algorithm.
Follow the Regularized Leader

FTRL is a fundamental online linear optimization algorithm.
Parametrized by a sequence of regularizers \( (\psi_t)_{t \in \mathbb{N}} \subseteq \text{simp}([N]) \to \mathbb{R} \),
FTRL is a fundamental online linear optimization algorithm. Parametrized by a sequence of regularizers \( (\psi_t)_{t \in \mathbb{N}} \subseteq \text{simp}([N]) \to \mathbb{R} \),

\[
w(t + 1) = \arg\min_{w \in \text{simp}([N])} (\langle w, L(t) \rangle + \psi_{t+1}(w)).
\]
Follow the Regularized Leader

FTRL is a fundamental online linear optimization algorithm. Parametrized by a sequence of regularizers \( (\psi_t)_{t \in \mathbb{N}} \subseteq \text{simp}([N]) \rightarrow \mathbb{R} \),

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

Hedge corresponds to \( \psi_{t+1}(w) = -\frac{H(w)}{\eta(t+1)} \).
Follow the Regularized Leader

FTRL is a fundamental online linear optimization algorithm. Parametrized by a sequence of regularizers \((\psi_t)_{t \in \mathbb{N}} \subseteq \text{simp}([\mathbb{N}]) \to \mathbb{R}\),

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

Hedge corresponds to \(\psi_{t+1}(w) = -\frac{H(w)}{\eta(t+1)}\). That is,

\[
    \frac{\exp\{-\eta(t+1)L(t)\}}{\sum_{i \in [\mathbb{N}] \exp\{-\eta(t+1)L_i(t)\}}} = \arg\min_{w \in \text{simp}([\mathbb{N}])} \left( \langle w, L(t) \rangle - \frac{H(w)}{\eta(t+1)} \right)
\]
Follow the Regularized Leader

FTRL is a fundamental online linear optimization algorithm. Parametrized by a sequence of regularizers \((\psi_t)_{t \in \mathbb{N}} \subseteq \text{simp}([N]) \rightarrow \mathbb{R}\),

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

Hedge corresponds to \(\psi_{t+1}(w) = -\frac{H(w)}{\eta(t+1)}\). That is,

\[
\frac{\exp \left\{ -\eta(t + 1)L(t) \right\}}{\sum_{i \in [N]} \exp \left\{ -\eta(t + 1)L_i(t) \right\}} = \arg \min_{w \in \text{simp}([N])} \left( \langle w, L(t) \rangle - \frac{H(w)}{\eta(t + 1)} \right).
\]
Introducing FTRL-CARE

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

\[
\begin{align*}
\hat{w}(t+1) & \in \arg\min_{w} \left\{ \langle w, L(t) \rangle - \sqrt{t+1} c_1 \sqrt{H(w)} + c_2 \right\}, \\
\eta(t+1) &= c_1 \sqrt{H(w(t+1))} + c_2 t+1, \\
w(t+1) &= \exp\{-\eta(t+1)L(t)\} \sum_{i \in [N]} \exp\{-\eta(t+1)L_i(t)\}.
\end{align*}
\]

Theorem BNR20

For any convex $D$, FTRL-CARE achieves

\[
E_R(T) \lesssim \sqrt{T \log N_0 + (\log N)^{3/2}} \Delta_0.
\]
Follow the Regularized Leader with CARE

Introducing FTRL-CARE
Follow the Regularized Leader with Constraint-Adaptive Root-Entropic regularization

\[ w(t + 1) \in \arg \min_{w \in \text{simp}([N])} \left( \langle w, L(t) \rangle - \frac{\sqrt{t+1}}{c_1} \sqrt{H(w) + c_2} \right), \]
Follow the Regularized Leader with CARE

Introducing **FTRL-CARE**

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

\[
\begin{align*}
    w(t + 1) &\in \arg\min_{w \in \text{simp}([N])} \left( \langle w, L(t) \rangle - \frac{\sqrt{t+1}}{c_1} \sqrt{H(w) + c_2} \right),
\end{align*}
\]

which is equivalent to solving the system of equations...

\[
\begin{align*}
    \eta(t + 1) &= c_1 \sqrt{\frac{H(w(t + 1)) + c_2}{t + 1}} \quad \text{and} \quad w(t + 1) = \frac{\exp \left\{ -\eta(t + 1)L(t) \right\}}{\sum_{i \in [N]} \exp \left\{ -\eta(t + 1)L_i(t) \right\}}.
\end{align*}
\]

**Theorem BNR20**

For any convex \( D \), FTRL-CARE achieves

\[
E_R(T) \lesssim \sqrt{T \log N_0 + \left(\log N\right)^{3/2}} \Delta_0.
\]
Follow the Regularized Leader with CARE

Introducing FTRL-CARE

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

\[
    w(t + 1) \in \arg \min_{w \in \text{simp}([N])} \left( \langle w, L(t) \rangle - \frac{\sqrt{t+1}}{c_1} \sqrt{H(w) + c_2} \right),
\]

which is equivalent to solving the system of equations...

\[
    \eta(t + 1) = c_1 \sqrt{\frac{H(w(t + 1)) + c_2}{t + 1}} \quad \text{and} \quad w(t + 1) = \frac{\exp \left\{ -\eta(t + 1)L(t) \right\}}{\sum_{i \in [N]} \exp \left\{ -\eta(t + 1)L_i(t) \right\}}.
\]

Theorem BNR20

For any convex \( \mathcal{D} \), FTRL-CARE achieves

\[
    \mathbb{E} R(T) \lesssim \sqrt{T \log N_0} + \frac{(\log N)^{3/2}}{\Delta_0}.
\]
Introducing \textbf{FTRL-CARE}

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

\[ w(t+1) \in \arg\min_{w \in \text{simp}([N])} \left( \langle w, L(t) \rangle - \frac{\sqrt{t+1}}{c_1} \sqrt{H(w) + c_2} \right), \]

which is equivalent to solving the system of equations...

\[ \eta(t+1) = c_1 \sqrt{\frac{H(w(t+1)) + c_2}{t+1}} \quad \text{and} \quad w(t+1) = \frac{\exp \{ -\eta(t+1) L(t) \}}{\sum_{i \in [N]} \exp \{ -\eta(t+1) L_i(t) \}}. \]

\textbf{Theorem BNR20}

For any convex \( \mathcal{D} \), FTRL-CARE achieves

\[ \mathbb{E} R(T) \lesssim \sqrt{T \log N_0} + \frac{(\log N)^{3/2}}{\Delta_0}. \]
CARE if you can, Hedge if you must; or, Meta-CARE for All

FTRL-CARE has adaptively minimax optimal dependence on $T, N_0$... but when $N_0 = 1$, it incurs total regret of order $(\log N)^{3/2}$ instead of $(\log N)\Delta_0$.

To be minimax optimal even when $N_0 = 1$, combine Hedge and FTRL-CARE.

Meta-CARE

• Treat the predictions of Hedge and FTRL-CARE as meta-experts...
• Use Hedge on these two meta-experts.
• Incur best regret of the two, plus some excess from meta-learning.
• Excess regret from meta-learning does not affect the order.

Theorem BNR20

For any convex $D$, Meta-CARE achieves $\mathbb{E}R(T) \lesssim \sqrt{T} \log N_0 + I[N_0 = 1] \log N \Delta_0 + I[N_0 \geq 2] (\log N)^{3/2} \Delta_0$. 


FTRL-CARE has adaptively minimax optimal dependence on $(T, N_0)$...
CARE if you can, Hedge if you must; or, Meta-CARE for All

FTRL-CARE has adaptively minimax optimal dependence on $(T, N_0)$...

... but when $N_0 = 1$, it incurs total regret of order $\frac{(\log N)^{3/2}}{\Delta_0}$ instead of $\frac{(\log N)}{\Delta_0}$.

Theorem BNR20

For any convex $D$, Meta-CARE achieves $\mathbb{E}R(T) \lesssim \sqrt{T \log N_0 + I[N_0 = 1] \log N \Delta_0 + I[N_0 \geq 2] (\log N)^{3/2} \Delta_0}$. 
FTRL-CARE has adaptively minimax optimal dependence on \((T, N_0)\)...

... but when \(N_0 = 1\), it incurs total regret of order \(\frac{(\log N)^{3/2}}{\Delta_0}\) instead of \(\frac{(\log N)}{\Delta_0}\).

To be minimax optimal even when \(N_0 = 1\), combine Hedge and FTRL-CARE.

Theorem BNR20

For any convex \(D\), Meta-CARE achieves

\[
E_R(T) \lessapprox \sqrt{T \log N_0} + \text{I}[N_0 = 1] \log N \Delta_0 + \text{I}[N_0 \geq 2] \left(\frac{\log N}{\Delta_0}\right)^{3/2}.
\]
FTRL-CARE has adaptively minimax optimal dependence on \((T, N_0)\)... but when \(N_0 = 1\), it incurs total regret of order \(\frac{(\log N)^{3/2}}{\Delta_0}\) instead of \(\frac{(\log N)}{\Delta_0}\).

To be minimax optimal even when \(N_0 = 1\), combine Hedge and FTRL-CARE.

**Meta-CARE**
- Treat the predictions of Hedge and FTRL-CARE as *meta-experts*...
CARE if you can, Hedge if you must; or, Meta-CARE for All

FTRL-CARE has adaptively minimax optimal dependence on \((T, N_0)\)...

... but when \(N_0 = 1\), it incurs total regret of order \(\frac{(\log N)^{3/2}}{\Delta_0}\) instead of \(\frac{(\log N)}{\Delta_0}\).

To be minimax optimal even when \(N_0 = 1\), combine Hedge and FTRL-CARE.

Meta-CARE

- Treat the predictions of Hedge and FTRL-CARE as *meta-experts*...
- Use Hedge on these two meta-experts.
CARE if you can, Hedge if you must; or, Meta-CARE for All

FTRL-CARE has adaptively minimax optimal dependence on \((T, N_0)\)...

... but when \(N_0 = 1\), it incurs total regret of order \(\frac{(\log N)^{3/2}}{\Delta_0}\) instead of \(\frac{(\log N)}{\Delta_0}\).

To be minimax optimal even when \(N_0 = 1\), combine Hedge and FTRL-CARE.

Meta-CARE

- Treat the predictions of Hedge and FTRL-CARE as meta-experts...
- Use Hedge on these two meta-experts.
- Incur best regret of the two, plus some excess from meta-learning.
FTRL-CARE has adaptively minimax optimal dependence on \((T, N_0)\)...

... but when \(N_0 = 1\), it incurs total regret of order \(\frac{(\log N)^{3/2}}{\Delta_0}\) instead of \(\frac{(\log N)}{\Delta_0}\).

To be minimax optimal even when \(N_0 = 1\), combine Hedge and FTRL-CARE.

**Meta-CARE**

- Treat the predictions of Hedge and FTRL-CARE as *meta-experts*...
- Use Hedge on these two meta-experts.
- Incur best regret of the two, plus some excess from meta-learning.
- Excess regret from meta-learning does not affect the order.
FTRL-CARE has adaptively minimax optimal dependence on \((T, N_0)\) ...
... but when \(N_0 = 1\), it incurs total regret of order \(\frac{(\log N)^{3/2}}{\Delta_0}\) instead of \(\frac{(\log N)}{\Delta_0}\).

To be minimax optimal even when \(N_0 = 1\), combine Hedge and FTRL-CARE.

Meta-CARE

- Treat the predictions of Hedge and FTRL-CARE as *meta-experts*...
- Use Hedge on these two meta-experts.
- Incur best regret of the two, plus some excess from meta-learning.
- Excess regret from meta-learning does not affect the order.

**Theorem BNR20**

For any convex \(\mathcal{D}\), Meta-CARE achieves

\[
\mathbb{E} R(T) \lesssim \sqrt{T \log N_0} + \mathbb{I}_{[N_0=1]} \frac{\log N}{\Delta_0} + \mathbb{I}_{[N_0 \geq 2]} \frac{(\log N)^{3/2}}{\Delta_0}.
\]
FTRL-CARE has adaptively minimax optimal dependence on \((T, N_0)\)...

... but when \(N_0 = 1\), it incurs total regret of order \(\frac{(\log N)^{3/2}}{\Delta_0}\) instead of \(\frac{(\log N)}{\Delta_0}\).

To be minimax optimal even when \(N_0 = 1\), combine Hedge and FTRL-CARE.

**Meta-CARE**

- Treat the predictions of Hedge and FTRL-CARE as *meta-experts*...
- Use Hedge on these two meta-experts.
- Incur best regret of the two, plus some excess from meta-learning.
- Excess regret from meta-learning does not affect the order.

**Theorem BNR20**

For any convex \(\mathcal{D}\), Meta-CARE achieves

\[
\mathbb{E} R(T) \lesssim \sqrt{T \log N_0} + \mathbb{I}_{[N_0=1]} \frac{\log N}{\Delta_0} + \mathbb{I}_{[N_0 \geq 2]} \frac{(\log N)^{3/2}}{\Delta_0}.
\]
Summary
Our Contributions

1. Introduced a spectrum of relaxations of the I.I.D. assumption.
   • Indexed by time-homogeneous convex constraints on the environment.
   • Interpolate between the pure stochastic and adversarial settings.
   • Data that we want to predict won’t be purely adversarial or stochastic.

2. Characterized the difficulty of learning along the spectrum using $N_0$ and $\Delta_0$.
   • Defined what it means to be adaptively minimax optimal along the spectrum.

3. Derived regret bounds for Hedge along spectrum from I.I.D. to adversarial.
   • In terms of the constraint $D$ via explicit dependence on $(N_0, \Delta_0)$.
   • Requires oracle knowledge to get minimax optimal dependence on $T$ and $N_0$.
   • Therefore Hedge is not adaptively minimax optimal.

4. Provided a new algorithm, Meta-CARE, and corresponding regret bounds.
   • Adapts optimally to our full spectrum of relaxations of the I.I.D. assumption.
   • ...without using oracle knowledge of $N_0$. 
Our Contributions

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

- Indexed by time-homogeneous convex constraints on the environment.
- Interpolate between the pure stochastic and adversarial settings.
- Data that we want to predict won’t be purely adversarial or stochastic.
- We want to know that we do well in intermediate scenarios as well.

2. Characterized the difficulty of learning along the spectrum using $N_0$ and $\Delta_0$.

- Defined what it means to be adaptively minimax optimal along the spectrum.

3. Derived regret bounds for Hedge along spectrum from I.I.D. to adversarial.

- In terms of the constraint $D$ via explicit dependence on $(N_0, \Delta_0)$.
- Requires oracle knowledge to get minimax optimal dependence on $T$ and $N_0$.
- Therefore Hedge is not adaptively minimax optimal.

4. Provided a new algorithm, Meta-CARE, and corresponding regret bounds.

- Adapts optimally to our full spectrum of relaxations of the I.I.D. assumption.
- ...without using oracle knowledge of $N_0$. 
Our Contributions

1. Introduced a spectrum of relaxations of the I.I.D. assumption.
   - Indexed by time-homogeneous convex constraints on the environment.
Our Contributions

1. Introduced a spectrum of relaxations of the I.I.D. assumption.
   - Indexed by time-homogeneous convex constraints on the environment.
   - Interpolate between the pure stochastic and adversarial settings.

2. Characterized the difficulty of learning along the spectrum using $N_0$ and $\Delta_0$.
   - Defined what it means to be adaptively minimax optimal along the spectrum.

3. Derived regret bounds for Hedge along spectrum from I.I.D. to adversarial.
   - In terms of the constraint $D$ via explicit dependence on $(N_0, \Delta_0)$.
   - Requires oracle knowledge to get minimax optimal dependence on $T$ and $N_0$.
   - Therefore Hedge is not adaptively minimax optimal.

4. Provided a new algorithm, Meta-CARE, and corresponding regret bounds.
   - Adapts optimally to our full spectrum of relaxations of the I.I.D. assumption.
   - ...without using oracle knowledge of $N_0$. 
Our Contributions

1. Introduced a spectrum of relaxations of the I.I.D. assumption.
   - Indexed by time-homogeneous convex constraints on the environment.
   - Interpolate between the pure stochastic and adversarial settings.
   - Data that we want to predict won’t be purely adversarial or stochastic.

2. Characterized the difficulty of learning along the spectrum using $N_0$ and $\Delta_0$.
   - Defined what it means to be adaptively minimax optimal along the spectrum.

3. Derived regret bounds for Hedge along spectrum from I.I.D. to adversarial.
   - In terms of the constraint $D$ via explicit dependence on $(N_0, \Delta_0)$.
   - Requires oracle knowledge to get minimax optimal dependence on $T$ and $N_0$.
   - Therefore Hedge is not adaptively minimax optimal.

4. Provided a new algorithm, Meta-CARE, and corresponding regret bounds.
   - Adapts optimally to our full spectrum of relaxations of the I.I.D. assumption.
   - Without using oracle knowledge of $N_0$. 


Our Contributions

1. Introduced a spectrum of relaxations of the I.I.D. assumption.
   - Indexed by time-homogeneous convex constraints on the environment.
   - Interpolate between the pure stochastic and adversarial settings.
   - Data that we want to predict won’t be purely adversarial or stochastic.
   - We want to know that we do well in intermediate scenarios as well.

2. Characterized the difficulty of learning along the spectrum using $N_0$ and $\Delta_0$.
   - Defined what it means to be adaptively minimax optimal along the spectrum.

3. Derived regret bounds for Hedge along spectrum from I.I.D. to adversarial.
   - In terms of the constraint $D$ via explicit dependence on $(N_0, \Delta_0)$.
   - Requires oracle knowledge to get minimax optimal dependence on $T$ and $N_0$.
   - Therefore Hedge is not adaptively minimax optimal.

4. Provided a new algorithm, Meta-CARE, and corresponding regret bounds.
   - Adapts optimally to our full spectrum of relaxations of the I.I.D. assumption.
   - ...without using oracle knowledge of $N_0$. 
Our Contributions

1. Introduced a spectrum of relaxations of the I.I.D. assumption.
   - Indexed by time-homogeneous convex constraints on the environment.
   - Interpolate between the pure stochastic and adversarial settings.
   - Data that we want to predict won’t be purely adversarial or stochastic.
   - We want to know that we do well in intermediate scenarios as well.

2. Characterized the difficulty of learning along the spectrum using $N_0$ and $\Delta_0$. 

3. Derived regret bounds for Hedge along spectrum from I.I.D. to adversarial.
   - In terms of the constraint $D$ via explicit dependence on $(N_0, \Delta_0)$.
   - Requires oracle knowledge to get minimax optimal dependence on $T$ and $N_0$.
   - Therefore Hedge is not adaptively minimax optimal.

4. Provided a new algorithm, Meta-CARE, and corresponding regret bounds.
   - Adapts optimally to our full spectrum of relaxations of the I.I.D. assumption.
   - ...without using oracle knowledge of $N_0$. 

Our Contributions

1. Introduced a spectrum of relaxations of the I.I.D. assumption.
   - Indexed by time-homogeneous convex constraints on the environment.
   - Interpolate between the pure stochastic and adversarial settings.
   - Data that we want to predict won’t be purely adversarial or stochastic.
   - We want to know that we do well in intermediate scenarios as well.

2. Characterized the difficulty of learning along the spectrum using \( N_0 \) and \( \Delta_0 \).
   - Defined what it means to be adaptively minimax optimal along the spectrum.
Our Contributions

1. Introduced a spectrum of relaxations of the I.I.D. assumption.
   - Indexed by time-homogeneous convex constraints on the environment.
   - Interpolate between the pure stochastic and adversarial settings.
   - Data that we want to predict won’t be purely adversarial or stochastic.
   - We want to know that we do well in intermediate scenarios as well.

2. Characterized the difficulty of learning along the spectrum using $N_0$ and $\Delta_0$.
   - Defined what it means to be adaptively minimax optimal along the spectrum.

3. Derived regret bounds for Hedge along spectrum from I.I.D. to adversarial.
Our Contributions

1. Introduced a spectrum of relaxations of the I.I.D. assumption.
   - Indexed by time-homogeneous convex constraints on the environment.
   - Interpolate between the pure stochastic and adversarial settings.
   - Data that we want to predict won’t be purely adversarial or stochastic.
   - We want to know that we do well in intermediate scenarios as well.

2. Characterized the difficulty of learning along the spectrum using $N_0$ and $\Delta_0$.
   - Defined what it means to be adaptively minimax optimal along the spectrum.

3. Derived regret bounds for Hedge along spectrum from I.I.D. to adversarial.
   - In terms of the constraint $\mathcal{D}$ via explicit dependence on $(N_0, \Delta_0)$. 
Our Contributions

1. Introduced a spectrum of relaxations of the I.I.D. assumption.
   - Indexed by time-homogeneous convex constraints on the environment.
   - Interpolate between the pure stochastic and adversarial settings.
   - Data that we want to predict won’t be purely adversarial or stochastic.
   - We want to know that we do well in intermediate scenarios as well.

2. Characterized the difficulty of learning along the spectrum using $N_0$ and $\Delta_0$.
   - Defined what it means to be adaptively minimax optimal along the spectrum.

3. Derived regret bounds for Hedge along spectrum from I.I.D. to adversarial.
   - In terms of the constraint $\mathcal{D}$ via explicit dependence on $(N_0, \Delta_0)$.
   - Requires oracle knowledge to get minimax optimal dependence on $T$ and $N_0$. 
Our Contributions

1. Introduced a spectrum of relaxations of the I.I.D. assumption.
   - Indexed by time-homogeneous convex constraints on the environment.
   - Interpolate between the pure stochastic and adversarial settings.
   - Data that we want to predict won’t be purely adversarial or stochastic.
   - We want to know that we do well in intermediate scenarios as well.

2. Characterized the difficulty of learning along the spectrum using $N_0$ and $\Delta_0$.
   - Defined what it means to be adaptively minimax optimal along the spectrum.

3. Derived regret bounds for Hedge along spectrum from I.I.D. to adversarial.
   - In terms of the constraint $D$ via explicit dependence on $(N_0, \Delta_0)$.
   - Requires oracle knowledge to get minimax optimal dependence on $T$ and $N_0$.
   - Therefore Hedge is not adaptively minimax optimal.

4. Provided a new algorithm, Meta-CARE, and corresponding regret bounds.
   - Adapts optimally to our full spectrum of relaxations of the I.I.D. assumption.
   - ...without using oracle knowledge of $N_0$. 
Our Contributions

1. Introduced a spectrum of relaxations of the I.I.D. assumption.
   - Indexed by time-homogeneous convex constraints on the environment.
   - Interpolate between the pure stochastic and adversarial settings.
   - Data that we want to predict won’t be purely adversarial or stochastic.
   - We want to know that we do well in intermediate scenarios as well.

2. Characterized the difficulty of learning along the spectrum using $N_0$ and $\Delta_0$.
   - Defined what it means to be adaptively minimax optimal along the spectrum.

3. Derived regret bounds for Hedge along spectrum from I.I.D. to adversarial.
   - In terms of the constraint $\mathcal{D}$ via explicit dependence on $(N_0, \Delta_0)$.
   - Requires oracle knowledge to get minimax optimal dependence on $T$ and $N_0$.
   - Therefore Hedge is not adaptively minimax optimal.

4. Provided a new algorithm, Meta-CARE, and corresponding regret bounds.
Our Contributions

1. Introduced a spectrum of relaxations of the I.I.D. assumption.
   - Indexed by time-homogeneous convex constraints on the environment.
   - Interpolate between the pure stochastic and adversarial settings.
   - Data that we want to predict won't be purely adversarial or stochastic.
   - We want to know that we do well in intermediate scenarios as well.

2. Characterized the difficulty of learning along the spectrum using $N_0$ and $\Delta_0$.
   - Defined what it means to be adaptively minimax optimal along the spectrum.

3. Derived regret bounds for Hedge along spectrum from I.I.D. to adversarial.
   - In terms of the constraint $\mathcal{D}$ via explicit dependence on $(N_0, \Delta_0)$.
   - Requires oracle knowledge to get minimax optimal dependence on $T$ and $N_0$.
   - Therefore Hedge is not adaptively minimax optimal.

4. Provided a new algorithm, Meta-CARE, and corresponding regret bounds.
   - Adapts optimally to our full spectrum of relaxations of the I.I.D. assumption.
Our Contributions

1. Introduced a spectrum of relaxations of the I.I.D. assumption.
   - Indexed by time-homogeneous convex constraints on the environment.
   - Interpolate between the pure stochastic and adversarial settings.
   - Data that we want to predict won’t be purely adversarial or stochastic.
   - We want to know that we do well in intermediate scenarios as well.

2. Characterized the difficulty of learning along the spectrum using $N_0$ and $\Delta_0$.
   - Defined what it means to be adaptively minimax optimal along the spectrum.

3. Derived regret bounds for Hedge along spectrum from I.I.D. to adversarial.
   - In terms of the constraint $\mathcal{D}$ via explicit dependence on $(N_0, \Delta_0)$.
   - Requires oracle knowledge to get minimax optimal dependence on $T$ and $N_0$.
   - Therefore Hedge is not adaptively minimax optimal.

4. Provided a new algorithm, Meta-CARE, and corresponding regret bounds.
   - Adapts optimally to our full spectrum of relaxations of the I.I.D. assumption.
   - ...without using oracle knowledge of $N_0$. 
Our Contributions

1. Introduced a spectrum of relaxations of the I.I.D. assumption.
   - Indexed by time-homogeneous convex constraints on the environment.
   - Interpolate between the pure stochastic and adversarial settings.
   - Data that we want to predict won’t be purely adversarial or stochastic.
   - We want to know that we do well in intermediate scenarios as well.

2. Characterized the difficulty of learning along the spectrum using $N_0$ and $\Delta_0$.
   - Defined what it means to be adaptively minimax optimal along the spectrum.

3. Derived regret bounds for Hedge along spectrum from I.I.D. to adversarial.
   - In terms of the constraint $\mathcal{D}$ via explicit dependence on $(N_0, \Delta_0)$.
   - Requires oracle knowledge to get minimax optimal dependence on $T$ and $N_0$.
   - Therefore Hedge is not adaptively minimax optimal.

4. Provided a new algorithm, Meta-CARE, and corresponding regret bounds.
   - Adapts optimally to our full spectrum of relaxations of the I.I.D. assumption.
   - …without using oracle knowledge of $N_0$. 
Our Contributions

1. Introduced a spectrum of relaxations of the I.I.D. assumption.
   - Indexed by time-homogeneous convex constraints on the environment.
   - Interpolate between the pure stochastic and adversarial settings.
   - Data that we want to predict won’t be purely adversarial or stochastic.
   - We want to know that we do well in intermediate scenarios as well.

2. Characterized the difficulty of learning along the spectrum using $N_0$ and $\Delta_0$.
   - Defined what it means to be adaptively minimax optimal along the spectrum.

3. Derived regret bounds for Hedge along spectrum from I.I.D. to adversarial.
   - In terms of the constraint $\mathcal{D}$ via explicit dependence on $(N_0, \Delta_0)$.
   - Requires oracle knowledge to get minimax optimal dependence on $T$ and $N_0$.
   - Therefore Hedge is not adaptively minimax optimal.

4. Provided a new algorithm, Meta-CARE, and corresponding regret bounds.
   - Adapts optimally to our full spectrum of relaxations of the I.I.D. assumption.
   - ...without using oracle knowledge of $N_0$. 
Our Contributions

1. Introduced a spectrum of relaxations of the I.I.D. assumption.
   - Indexed by time-homogeneous convex constraints on the environment.
   - Interpolate between the pure stochastic and adversarial settings.
   - Data that we want to predict won’t be purely adversarial or stochastic.
   - We want to know that we do well in intermediate scenarios as well.

2. Characterized the difficulty of learning along the spectrum using $N_0$ and $\Delta_0$.
   - Defined what it means to be adaptively minimax optimal along the spectrum.

3. Derived regret bounds for Hedge along spectrum from I.I.D. to adversarial.
   - In terms of the constraint $\mathcal{D}$ via explicit dependence on $(N_0, \Delta_0)$.
   - Requires oracle knowledge to get minimax optimal dependence on $T$ and $N_0$.
   - Therefore Hedge is not adaptively minimax optimal.

4. Provided a new algorithm, Meta-CARE, and corresponding regret bounds.
   - Adapts optimally to our full spectrum of relaxations of the I.I.D. assumption.
   - ...without using oracle knowledge of $N_0$. 
References


