# Some Elements of Learning Theory

Nicolò Cesa-Bianchi

Università degli Studi di Milano

# Contents

# Contents

- A brief introduction to statistical learning
- From statistical learning to sequential decision making

# Contents

- A brief introduction to statistical learning
- From statistical learning to sequential decision making
- Prediction with expert advice and multiarmed bandits

# Contents

# Contents

# Contents

- A brief introduction to statistical learning
- From statistical learning to sequential decision making
- Prediction with expert advice and multiarmed bandits
- Online convex optimization
- Contextual bandits

- We do some (short) proofs

# Statistical learning

- One of the most important mathematical frameworks for the analysis of learning algorithms (mainly supervised learning)

# Statistical learning

- One of the most important mathematical frameworks for the analysis of learning algorithms (mainly supervised learning)
- Pioneered by Vladimir Vapnik in the Seventies

# Statistical learning

- One of the most important mathematical frameworks for the analysis of learning algorithms (mainly supervised learning)
- Pioneered by Vladimir Vapnik in the Seventies
- Later —and independently— Leslie Valiant introduces computational learning theory (A theory of the learnable, 1984)

# Statistical learning

- One of the most important mathematical frameworks for the analysis of learning algorithms (mainly supervised learning)
- Pioneered by Vladimir Vapnik in the Seventies
- Later —and independently— Leslie Valiant introduces computational learning theory (A theory of the learnable, 1984)

## Main contributions:

# Statistical learning

- One of the most important mathematical frameworks for the analysis of learning algorithms (mainly supervised learning)
- Pioneered by Vladimir Vapnik in the Seventies
- Later —and independently— Leslie Valiant introduces computational learning theory (A theory of the learnable, 1984)

## Main contributions:

- Mathematical model of learning and conditions characterizing what can be learned

# Statistical learning

- One of the most important mathematical frameworks for the analysis of learning algorithms (mainly supervised learning)
- Pioneered by Vladimir Vapnik in the Seventies
- Later —and independently— Leslie Valiant introduces computational learning theory (A theory of the learnable, 1984)

## Main contributions:

- Mathematical model of learning and conditions characterizing what can be learned
- Guidelines to practitioners (e.g., choice of learning bias, control of overfitting)

## Statistical learning

- One of the most important mathematical frameworks for the analysis of learning algorithms (mainly supervised learning)
- Pioneered by Vladimir Vapnik in the Seventies
- Later —and independently— Leslie Valiant introduces computational learning theory (A theory of the learnable, 1984)

### Main contributions:

- Mathematical model of learning and conditions characterizing what can be learned
- Guidelines to practitioners (e.g., choice of learning bias, control of overfitting)
- Principled and successful algorithms (SVM, Boosting)

# Ingredients

- Data space $\mathcal{X}$ (often $\mathcal{X} = \mathbb{R}^d$)

# Ingredients

- Data space $\mathcal{X}$ (often $\mathcal{X} = \mathbb{R}^d$)
- Label space $\mathcal{Y}$
  - $\mathcal{Y} = \mathbb{R}$ for regression
  - $\mathcal{Y} = \{-1, 1\}$ for binary classification

# Ingredients

- Data space $\mathcal{X}$ (often $\mathcal{X} = \mathbb{R}^d$)
- Label space $\mathcal{Y}$
  - $\mathcal{Y} = \mathbb{R}$ for regression
  - $\mathcal{Y} = \{-1, 1\}$ for binary classification
- Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$
  - Quadratic $\ell(y, \widehat{y}) = (\widehat{y} - y)^2$ for regression
  - Zero-one $\ell(y, \widehat{y}) = \mathbb{I}\{\widehat{y} \neq y\}$ for binary classification
  - Hinge $\ell(y, \widehat{y}) = \big[1 - y\,\widehat{y}\big]_+$ convex proxy for binary classification

# Ingredients

- Data space $\mathcal{X}$ (often $\mathcal{X} = \mathbb{R}^d$)
- Label space $\mathcal{Y}$
  - $\mathcal{Y} = \mathbb{R}$ for regression
  - $\mathcal{Y} = \{-1, 1\}$ for binary classification
- Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$
  - Quadratic $\ell(y, \widehat{y}) = (\widehat{y} - y)^2$ for regression
  - Zero-one $\ell(y, \widehat{y}) = \mathbb{I}\{\widehat{y} \neq y\}$ for binary classification
  - Hinge $\ell(y, \widehat{y}) = \big[1 - y\,\widehat{y}\big]_+$ convex proxy for binary classification

- Predictor $f : \mathcal{X} \to \mathcal{Y}$ maps data points to labels

# Ingredients

- Data space $\mathcal{X}$ (often $\mathcal{X} = \mathbb{R}^d$)
- Label space $\mathcal{Y}$
  - $\mathcal{Y} = \mathbb{R}$ for regression
  - $\mathcal{Y} = \{-1, 1\}$ for binary classification
- Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$
  - Quadratic $\ell(y, \widehat{y}) = (\widehat{y} - y)^2$ for regression
  - Zero-one $\ell(y, \widehat{y}) = \mathbb{I}\{\widehat{y} \neq y\}$ for binary classification
  - Hinge $\ell(y, \widehat{y}) = \big[1 - y\,\widehat{y}\big]_+$ convex proxy for binary classification

- Predictor $f : \mathcal{X} \to \mathcal{Y}$ maps data points to labels
- Training set $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_m, y_m)$ a (multi)set $S$ of labeled data points

# Ingredients

- Data space $\mathcal{X}$ (often $\mathcal{X} = \mathbb{R}^d$)
- Label space $\mathcal{Y}$
  - $\mathcal{Y} = \mathbb{R}$ for regression
  - $\mathcal{Y} = \{-1, 1\}$ for binary classification
- Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$
  - Quadratic $\ell(y, \widehat{y}) = (\widehat{y} - y)^2$ for regression
  - Zero-one $\ell(y, \widehat{y}) = \mathbb{I}\{\widehat{y} \neq y\}$ for binary classification
  - Hinge $\ell(y, \widehat{y}) = \left[1 - y\,\widehat{y}\right]_+$ convex proxy for binary classification

- Predictor $f : \mathcal{X} \to \mathcal{Y}$ maps data points to labels
- Training set $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_m, y_m)$ a (multi)set $S$ of labeled data points
- Learning algorithm: given a loss function, maps finite training sets to predictors

# Statistical learning

- A learning problem is defined by an unknown distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$

# Statistical learning

- A learning problem is defined by an unknown distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$
- Any data point $(\boldsymbol{x}, y)$ is the realization of an indipendent random draw $(\boldsymbol{X}, Y)$ from $\mathcal{D}$

# Statistical learning

- A learning problem is defined by an unknown distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$
- Any data point $(\boldsymbol{x}, y)$ is the realization of an indipendent random draw $(\boldsymbol{X}, Y)$ from $\mathcal{D}$
- Therefore, the training set $S$ is a random sample from $\mathcal{D}$

# Statistical learning

- A learning problem is defined by an unknown distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$
- Any data point $(\boldsymbol{x}, y)$ is the realization of an indipendent random draw $(\boldsymbol{X}, Y)$ from $\mathcal{D}$
- Therefore, the training set $S$ is a random sample from $\mathcal{D}$
- Given a loss, the statistical risk of predictor $f$ is $\ell_{\mathcal{D}}(f) = \mathbb{E}\big[\ell(Y, f(\boldsymbol{X}))\big]$

# Statistical learning

- A learning problem is defined by an unknown distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$
- Any data point $(\boldsymbol{x}, y)$ is the realization of an indipendent random draw $(\boldsymbol{X}, Y)$ from $\mathcal{D}$
- Therefore, the training set $S$ is a random sample from $\mathcal{D}$
- Given a loss, the statistical risk of predictor $f$ is $\boxed{\ell_{\mathcal{D}}(f) = \mathbb{E}\big[\ell(Y, f(\boldsymbol{X}))\big]}$
- Bayes optimal predictor $f^* : \mathcal{X} \to \mathcal{Y}$ is $\quad f^*(\boldsymbol{x}) = \operatorname*{argmin}_{\widehat{y} \in \mathcal{Y}} \mathbb{E}\big[\ell(Y, \widehat{y}) \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big]$

# Statistical learning

- A learning problem is defined by an unknown distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$
- Any data point $(\boldsymbol{x}, y)$ is the realization of an indipendent random draw $(\boldsymbol{X}, Y)$ from $\mathcal{D}$
- Therefore, the training set $S$ is a random sample from $\mathcal{D}$
- Given a loss, the statistical risk of predictor $f$ is $\boxed{\ell_{\mathcal{D}}(f) = \mathbb{E}\big[\ell(Y, f(\boldsymbol{X}))\big]}$
- Bayes optimal predictor $f^* : \mathcal{X} \to \mathcal{Y}$ is $\quad f^*(\boldsymbol{x}) = \operatorname*{argmin}_{\widehat{y} \in \mathcal{Y}} \mathbb{E}\big[\ell(Y, \widehat{y}) \mid \boldsymbol{X} = \boldsymbol{x}\big]$
- Bayes risk $\ell_{\mathcal{D}}(f^*)$

# Statistical learning

- A learning problem is defined by an unknown distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$
- Any data point $(\boldsymbol{x}, y)$ is the realization of an indipendent random draw $(\boldsymbol{X}, Y)$ from $\mathcal{D}$
- Therefore, the training set $S$ is a random sample from $\mathcal{D}$
- Given a loss, the statistical risk of predictor $f$ is $\boxed{\ell_{\mathcal{D}}(f) = \mathbb{E}\big[\ell(Y, f(\boldsymbol{X}))\big]}$
- Bayes optimal predictor $f^* : \mathcal{X} \to \mathcal{Y}$ is $\quad f^*(\boldsymbol{x}) = \operatorname*{argmin}_{\widehat{y} \in \mathcal{Y}} \mathbb{E}\big[\ell(Y, \widehat{y}) \mid \boldsymbol{X} = \boldsymbol{x}\big]$
- Bayes risk $\ell_{\mathcal{D}}(f^*)$
- Square loss: $f^*(\boldsymbol{x}) = \mathbb{E}\big[Y \mid \boldsymbol{X} = \boldsymbol{x}\big]$ and $\ell_{\mathcal{D}}(f^*) = \mathbb{E}\big[\mathrm{Var}[Y \mid \boldsymbol{X}]\big]$

# Statistical learning

- A learning problem is defined by an unknown distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$
- Any data point $(\boldsymbol{x}, y)$ is the realization of an indipendent random draw $(\boldsymbol{X}, Y)$ from $\mathcal{D}$
- Therefore, the training set $S$ is a random sample from $\mathcal{D}$
- Given a loss, the statistical risk of predictor $f$ is $\ell_{\mathcal{D}}(f) = \mathbb{E}\big[\ell(Y, f(\boldsymbol{X}))\big]$
- Bayes optimal predictor $f^* : \mathcal{X} \to \mathcal{Y}$ is $\quad f^*(\boldsymbol{x}) = \operatorname*{argmin}_{\widehat{y} \in \mathcal{Y}} \mathbb{E}\big[\ell(Y, \widehat{y}) \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big]$
- Bayes risk $\ell_{\mathcal{D}}(f^*)$
- Square loss: $f^*(\boldsymbol{x}) = \mathbb{E}\big[Y \mid \boldsymbol{X} = \boldsymbol{x}\big]$ and $\ell_{\mathcal{D}}(f^*) = \mathbb{E}\big[\mathrm{Var}[Y \mid \boldsymbol{X}]\big]$
- Zero-one loss: $f^*(\boldsymbol{x}) = 2\mathbb{I}\{\eta(\boldsymbol{x}) \geq 1/2\} - 1$ and $\ell_{\mathcal{D}}(f^*) = \mathbb{E}\Big[\min\{\eta(\boldsymbol{X}), 1 - \eta(\boldsymbol{X})\}\Big]$
  where $\eta(\boldsymbol{x}) = \mathbb{P}(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x})$

# The bias-variance decomposition

Suppose $h_S \in \mathcal{H}$ is the predictor output by a learning algorithm $A$ with training set $S$ ($h_S \in \mathcal{H}$ is a random variable)

# The bias-variance decomposition

Suppose $h_S \in \mathcal{H}$ is the predictor output by a learning algorithm $A$ with training set $S$ ($h_S \in \mathcal{H}$ is a random variable)

$$
\begin{aligned}
\ell_{\mathcal{D}}(h_S) = \ell_{\mathcal{D}}(h_S) - \inf_{h \in \mathcal{H}} \ell_{\mathcal{D}}(h) & \qquad \text{(estimation error} \rightarrow \text{overfitting)} \\
+ \inf_{h \in \mathcal{H}} \ell_{\mathcal{D}}(h) - \ell_{\mathcal{D}}(f^*) & \qquad \text{(approximation error} \rightarrow \text{underfitting)} \\
+ \ell_{\mathcal{D}}(f^*) & \qquad \text{(Bayes risk)}
\end{aligned}
$$

## Trade-offs

# The bias-variance decomposition

Suppose $h_S \in \mathcal{H}$ is the predictor output by a learning algorithm $A$ with training set $S$ ($h_S \in \mathcal{H}$ is a random variable)

$$
\begin{aligned}
\ell_{\mathcal{D}}(h_S) = \ell_{\mathcal{D}}(h_S) - \inf_{h \in \mathcal{H}} \ell_{\mathcal{D}}(h) & \qquad \text{(estimation error} \to \text{overfitting)} \\
+ \inf_{h \in \mathcal{H}} \ell_{\mathcal{D}}(h) - \ell_{\mathcal{D}}(f^*) & \qquad \text{(approximation error} \to \text{underfitting)} \\
+ \ell_{\mathcal{D}}(f^*) & \qquad \text{(Bayes risk)}
\end{aligned}
$$

## Trade-offs

- Underfitting control: Let $\mathcal{H}$ be as large as possible

# The bias-variance decomposition

Suppose $h_S \in \mathcal{H}$ is the predictor output by a learning algorithm $A$ with training set $S$ ($h_S \in \mathcal{H}$ is a random variable)

$$
\begin{aligned}
\ell_{\mathcal{D}}(h_S) = \ell_{\mathcal{D}}(h_S) - \inf_{h \in \mathcal{H}} \ell_{\mathcal{D}}(h) & \qquad \text{(estimation error} \rightarrow \text{overfitting)} \\
+ \inf_{h \in \mathcal{H}} \ell_{\mathcal{D}}(h) - \ell_{\mathcal{D}}(f^*) & \qquad \text{(approximation error} \rightarrow \text{underfitting)} \\
+ \ell_{\mathcal{D}}(f^*) & \qquad \text{(Bayes risk)}
\end{aligned}
$$

## Trade-offs

- Underfitting control: Let $\mathcal{H}$ be as large as possible
- Overfitting control:

# The bias-variance decomposition

Suppose $h_S \in \mathcal{H}$ is the predictor output by a learning algorithm $A$ with training set $S$ ($h_S \in \mathcal{H}$ is a random variable)

$$
\begin{aligned}
\ell_{\mathcal{D}}(h_S) = \ell_{\mathcal{D}}(h_S) - \inf_{h\in\mathcal{H}} \ell_{\mathcal{D}}(h) & \qquad \text{(estimation error} \rightarrow \text{overfitting)} \\
+ \inf_{h\in\mathcal{H}} \ell_{\mathcal{D}}(h) - \ell_{\mathcal{D}}(f^*) & \qquad \text{(approximation error} \rightarrow \text{underfitting)} \\
+ \ell_{\mathcal{D}}(f^*) & \qquad \text{(Bayes risk)}
\end{aligned}
$$

## Trade-offs

- Underfitting control: Let $\mathcal{H}$ be as large as possible
- Overfitting control:
  - Ensure that training error of $h$ is close to $\ell_{\mathcal{D}}(h)$ for all $h \in \mathcal{H}$ (uniform convergence)

# The bias-variance decomposition

Suppose $h_S \in \mathcal{H}$ is the predictor output by a learning algorithm $A$ with training set $S$ ($h_S \in \mathcal{H}$ is a random variable)

$$
\begin{aligned}
\ell_{\mathcal{D}}(h_S) = \ell_{\mathcal{D}}(h_S) - \inf_{h \in \mathcal{H}} \ell_{\mathcal{D}}(h) & \qquad \text{(estimation error} \rightarrow \text{overfitting)} \\
+ \inf_{h \in \mathcal{H}} \ell_{\mathcal{D}}(h) - \ell_{\mathcal{D}}(f^*) & \qquad \text{(approximation error} \rightarrow \text{underfitting)} \\
+ \ell_{\mathcal{D}}(f^*) & \qquad \text{(Bayes risk)}
\end{aligned}
$$

## Trade-offs

- Underfitting control: Let $\mathcal{H}$ be as large as possible
- Overfitting control:
  - Ensure that training error of $h$ is close to $\ell_{\mathcal{D}}(h)$ for all $h \in \mathcal{H}$ (uniform convergence)
  - Minimize regularized training error (stability)

# The bias-variance decomposition

Suppose $h_S \in \mathcal{H}$ is the predictor output by a learning algorithm $A$ with training set $S$ ($h_S \in \mathcal{H}$ is a random variable)

$$\begin{aligned}
\ell_{\mathcal{D}}(h_S) &= \ell_{\mathcal{D}}(h_S) - \inf_{h\in\mathcal{H}} \ell_{\mathcal{D}}(h) && \text{(estimation error} \to \text{overfitting)}\\
&\quad + \inf_{h\in\mathcal{H}} \ell_{\mathcal{D}}(h) - \ell_{\mathcal{D}}(f^*) && \text{(approximation error} \to \text{underfitting)}\\
&\quad + \ell_{\mathcal{D}}(f^*) && \text{(Bayes risk)}
\end{aligned}$$

## Trade-offs

- Underfitting control: Let $\mathcal{H}$ be as large as possible
- Overfitting control:
  - Ensure that training error of $h$ is close to $\ell_{\mathcal{D}}(h)$ for all $h \in \mathcal{H}$ (uniform convergence)
  - Minimize regularized training error (stability)
  - Show that $A$ can compress the training set (compression implies learning)

# Success stories: Characterization of sample complexity

What is the training set size $m_{\mathcal{H}}$ necessary and sufficient to ensure

$$\ell_{\mathcal{D}}(h_S) - \inf_{h \in \mathcal{H}} \ell_{\mathcal{D}}(h) \leq \varepsilon$$

with probability at least $1 - \delta$ w.r.t. the random draw of $S$ and irrespective to $\mathcal{D}$?

# Success stories: Characterization of sample complexity

What is the training set size $m_{\mathcal{H}}$ necessary and sufficient to ensure

$$\ell_{\mathcal{D}}(h_S) - \inf_{h \in \mathcal{H}} \ell_{\mathcal{D}}(h) \leq \varepsilon$$

with probability at least $1 - \delta$ w.r.t. the random draw of $S$ and irrespective to $\mathcal{D}$?

Binary classification with zero-one loss

# Success stories: Characterization of sample complexity

What is the training set size $m_{\mathcal{H}}$ necessary and sufficient to ensure

$$\ell_{\mathcal{D}}(h_S) - \inf_{h \in \mathcal{H}} \ell_{\mathcal{D}}(h) \leq \varepsilon$$

with probability at least $1 - \delta$ w.r.t. the random draw of $S$ and irrespective to $\mathcal{D}$?

## Binary classification with zero-one loss

- $m_{\mathcal{H}}$ is determined by a simple combinatorial parameter, the VC-dimension $d_{\mathcal{H}}$

# Success stories: Characterization of sample complexity

What is the training set size $m_{\mathcal{H}}$ necessary and sufficient to ensure

$$\ell_{\mathcal{D}}(h_S) - \inf_{h \in \mathcal{H}} \ell_{\mathcal{D}}(h) \le \varepsilon$$

with probability at least $1 - \delta$ w.r.t. the random draw of $S$ and irrespective to $\mathcal{D}$?

## Binary classification with zero-one loss

- $m_{\mathcal{H}}$ is determined by a simple combinatorial parameter, the VC-dimension $d_{\mathcal{H}}$
- Agnostic case: $m_{\mathcal{H}} = \Theta\left(\frac{d_{\mathcal{H}} + \ln(1/\delta)}{\varepsilon^2}\right)$

## Success stories: Characterization of sample complexity

What is the training set size $m_{\mathcal{H}}$ necessary and sufficient to ensure

$$\ell_{\mathcal{D}}(h_S) - \inf_{h \in \mathcal{H}} \ell_{\mathcal{D}}(h) \leq \varepsilon$$

with probability at least $1 - \delta$ w.r.t. the random draw of $S$ and irrespective to $\mathcal{D}$?

### Binary classification with zero-one loss

- $m_{\mathcal{H}}$ is determined by a simple combinatorial parameter, the VC-dimension $d_{\mathcal{H}}$
- Agnostic case: $m_{\mathcal{H}} = \Theta\left(\dfrac{d_{\mathcal{H}} + \ln(1/\delta)}{\varepsilon^2}\right)$
- Realizable case: ($f^* \in \mathcal{H}$ and $\ell_{\mathcal{D}}(f^*) = 0$) $m_{\mathcal{H}} = \Theta\left(\dfrac{d_{\mathcal{H}} + \ln(1/\delta)}{\varepsilon}\right)$

## Success stories: Characterization of sample complexity

What is the training set size $m_{\mathcal{H}}$ necessary and sufficient to ensure

$$\ell_{\mathcal{D}}(h_S) - \inf_{h \in \mathcal{H}} \ell_{\mathcal{D}}(h) \leq \varepsilon$$

with probability at least $1 - \delta$ w.r.t. the random draw of $S$ and irrespective to $\mathcal{D}$?

### Binary classification with zero-one loss

- $m_{\mathcal{H}}$ is determined by a simple combinatorial parameter, the VC-dimension $d_{\mathcal{H}}$
- Agnostic case: $m_{\mathcal{H}} = \Theta\left(\frac{d_{\mathcal{H}} + \ln(1/\delta)}{\varepsilon^2}\right)$
- Realizable case: $(f^* \in \mathcal{H}$ and $\ell_{\mathcal{D}}(f^*) = 0)$ $m_{\mathcal{H}} = \Theta\left(\frac{d_{\mathcal{H}} + \ln(1/\delta)}{\varepsilon}\right)$
- $d_{\mathcal{H}}$ can be infinite, implying $\mathcal{H}$ is not learnable

# Success stories: Characterization of sample complexity

What is the training set size $m_{\mathcal{H}}$ necessary and sufficient to ensure

$$\ell_{\mathcal{D}}(h_S) - \inf_{h \in \mathcal{H}} \ell_{\mathcal{D}}(h) \leq \varepsilon$$

with probability at least $1 - \delta$ w.r.t. the random draw of $S$ and irrespective to $\mathcal{D}$?

## Binary classification with zero-one loss

- $m_{\mathcal{H}}$ is determined by a simple combinatorial parameter, the VC-dimension $d_{\mathcal{H}}$
- Agnostic case: $m_{\mathcal{H}} = \Theta\left(\frac{d_{\mathcal{H}} + \ln(1/\delta)}{\varepsilon^2}\right)$
- Realizable case: ($f^* \in \mathcal{H}$ and $\ell_{\mathcal{D}}(f^*) = 0$) $m_{\mathcal{H}} = \Theta\left(\frac{d_{\mathcal{H}} + \ln(1/\delta)}{\varepsilon}\right)$
- $d_{\mathcal{H}}$ can be infinite, implying $\mathcal{H}$ is not learnable
- Minimizing training error in $\mathcal{H}$ achieves upper bound in the agnostic case

# Success stories: Characterization of sample complexity

What is the training set size $m_{\mathcal{H}}$ necessary and sufficient to ensure

$$\ell_{\mathcal{D}}(h_S) - \inf_{h \in \mathcal{H}} \ell_{\mathcal{D}}(h) \leq \varepsilon$$

with probability at least $1 - \delta$ w.r.t. the random draw of $S$ and irrespective to $\mathcal{D}$?

## Binary classification with zero-one loss

- $m_{\mathcal{H}}$ is determined by a simple combinatorial parameter, the VC-dimension $d_{\mathcal{H}}$
- Agnostic case: $m_{\mathcal{H}} = \Theta\left(\frac{d_{\mathcal{H}} + \ln(1/\delta)}{\varepsilon^2}\right)$
- Realizable case: ($f^* \in \mathcal{H}$ and $\ell_{\mathcal{D}}(f^*) = 0$) $m_{\mathcal{H}} = \Theta\left(\frac{d_{\mathcal{H}} + \ln(1/\delta)}{\varepsilon}\right)$
- $d_{\mathcal{H}}$ can be infinite, implying $\mathcal{H}$ is not learnable
- Minimizing training error in $\mathcal{H}$ achieves upper bound in the agnostic case
- Majority vote over a set of consistent predictors achieves upper bound in the realizable case

# Statistical consistency

- $A$ is statistically consistent if $\lim_{m\to\infty} \mathbb{E}\Big[\ell_{\mathcal{D}}(A(S_m))\Big] = \ell_{\mathcal{D}}(f^*)$

# Statistical consistency

- $A$ is statistically consistent if $\displaystyle\lim_{m\to\infty} \mathbb{E}\Big[\ell_{\mathcal{D}}(A(S_m))\Big] = \ell_{\mathcal{D}}(f^*)$
- In order to achieve distribution-free consistency, $A$ has to be nonparametric (e.g., $k$-NN, tree classifiers, SVMs with Gaussian kernels)

# Statistical consistency

- $A$ is statistically consistent if $\lim_{m \to \infty} \mathbb{E}\Big[\ell_{\mathcal{D}}\big(A(S_m)\big)\Big] = \ell_{\mathcal{D}}(f^*)$
- In order to achieve distribution-free consistency, $A$ has to be nonparametric (e.g., $k$-NN, tree classifiers, SVMs with Gaussian kernels)

## No Free Lunch Theorem

Let $a_1, a_2, \cdots > 0$ be any sequence of numbers slowly converging to zero.
For all binary classification algorithms $A$ there exists $\mathcal{D}$ such that the Bayes risk is zero and, simultaneously, $\mathbb{E}\big[\ell_{\mathcal{D}}\big(A(S_m)\big)\big] \geq a_m$ for all $m \geq 1$.

# Statistical consistency

- $A$ is statistically consistent if $\lim_{m\to\infty} \mathbb{E}\Big[\ell_{\mathcal{D}}(A(S_m))\Big] = \ell_{\mathcal{D}}(f^*)$
- In order to achieve distribution-free consistency, $A$ has to be nonparametric (e.g., $k$-NN, tree classifiers, SVMs with Gaussian kernels)

## No Free Lunch Theorem

Let $a_1, a_2, \dots > 0$ be any sequence of numbers slowly converging to zero.
For all binary classification algorithms $A$ there exists $\mathcal{D}$ such that the Bayes risk is zero and, simultaneously, $\mathbb{E}[\ell_{\mathcal{D}}(A(S_m))] \geq a_m$ for all $m \geq 1$.

## Curse of dimensionality

- Typical parametric rates for convergence to $\ell_{\mathcal{D}}(h^*)$: $m^{-1/2}$
- Typical nonparametric rates for convergence to Bayes risk: $m^{-1/d}$ for $d \geq 2$ (under assumptions on $\mathcal{D}$)

# Online learning


- Data streams are ubiquitous: sensors, markets, user interactions

# Online learning


- Data streams are ubiquitous: sensors, markets, user interactions
- New data is being generated all the time

# Online learning


- Data streams are ubiquitous: sensors, markets, user interactions
- New data is being generated all the time
- The train-test model of statistical learning is ill-suited for learning on data streams

# Online learning


- Data streams are ubiquitous: sensors, markets, user interactions
- New data is being generated all the time
- The train-test model of statistical learning is ill-suited for learning on data streams
- After observing a new data point, predictors should be incrementally adjusted at a constant cost

# History bits

- Online learning model formalized by Nick Littlestone and Manfred Warmuth (Mistake bounds and logarithmic linear-threshold learning algorithms, 1989)

# History bits


- Online learning model formalized by Nick Littlestone and Manfred Warmuth (Mistake bounds and logarithmic linear-threshold learning algorithms, 1989)
- Volodya Vovk independently develops a related framework (Aggregating strategies, 1990)

# History bits

- Online learning model formalized by Nick Littlestone and Manfred Warmuth (Mistake bounds and logarithmic linear-threshold learning algorithms, 1989)
- Volodya Vovk independently develops a related framework (Aggregating strategies, 1990)
- Similar ideas also independently emerged in game theory and information theory

# The online learning protocol

The algorithm starts with a default model $h_1 \in \mathcal{H}$

For $t = 1, 2, \ldots$

# The online learning protocol

The algorithm starts with a default model $h_1 \in \mathcal{H}$

For $t = 1, 2, \dots$

1. The current model $h_t \in \mathcal{H}$ is tested on the next data point $(\boldsymbol{x}_t, y_t)$ in the stream

# The online learning protocol

The algorithm starts with a default model $h_1 \in \mathcal{H}$

For $t = 1, 2, \ldots$

1. The current model $h_t \in \mathcal{H}$ is tested on the next data point $(\boldsymbol{x}_t, y_t)$ in the stream
2. $A$ is charged with loss $\ell(y_t, h_t(\boldsymbol{x}_t))$

# The online learning protocol

The algorithm starts with a default model $h_1 \in \mathcal{H}$

For $t = 1, 2, \dots$

1. The current model $h_t \in \mathcal{H}$ is tested on the next data point $(\boldsymbol{x}_t, y_t)$ in the stream
2. $A$ is charged with loss $\ell(y_t, h_t(\boldsymbol{x}_t))$
3. $h_{t+1} \in \mathcal{H}$ is computed based on $h_t$ and $(\boldsymbol{x}_t, y_t)$

# The online learning protocol

The algorithm starts with a default model $h_1 \in \mathcal{H}$

For $t = 1, 2, \ldots$

1. The current model $h_t \in \mathcal{H}$ is tested on the next data point $(\boldsymbol{x}_t, y_t)$ in the stream
2. $A$ is charged with loss $\ell(y_t, h_t(\boldsymbol{x}_t))$
3. $h_{t+1} \in \mathcal{H}$ is computed based on $h_t$ and $(\boldsymbol{x}_t, y_t)$

- Computation of $h_{t+1}$ relies on local information

# The online learning protocol

The algorithm starts with a default model $h_1 \in \mathcal{H}$

For $t = 1, 2, \ldots$

1. The current model $h_t \in \mathcal{H}$ is tested on the next data point $(\boldsymbol{x}_t, y_t)$ in the stream
2. $A$ is charged with loss $\ell(y_t, h_t(\boldsymbol{x}_t))$
3. $h_{t+1} \in \mathcal{H}$ is computed based on $h_t$ and $(\boldsymbol{x}_t, y_t)$

- Computation of $h_{t+1}$ relies on local information
- No stochastic assumptions on the stream

# Regret

## Sequential risk

Given a convex loss $\ell$ and a stream $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots$, the sequential risk of $A$ is

$$\boxed{\sum_{t=1}^{T} \ell(y_t, h_t(\boldsymbol{x}_t))}$$

# Regret

## Sequential risk

Given a convex loss $\ell$ and a stream $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots$, the sequential risk of $A$ is

$$\boxed{\sum_{t=1}^{T} \ell(y_t, h_t(\boldsymbol{x}_t))}$$

## Regret

$$R_T = \sum_{t=1}^{T} \ell(y_t, h_t(\boldsymbol{x}_t)) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(y_t, h(\boldsymbol{x}_t))$$

# Regret

## Sequential risk

Given a convex loss $\ell$ and a stream $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots$, the sequential risk of $A$ is

$$\boxed{\sum_{t=1}^{T} \ell(y_t, h_t(\boldsymbol{x}_t))}$$

## Regret

$$R_T = \sum_{t=1}^{T} \ell(y_t, h_t(\boldsymbol{x}_t)) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(y_t, h(\boldsymbol{x}_t))$$

- A sequential counterpart to the variance error in statistical learning

# Regret

## Sequential risk

Given a convex loss $\ell$ and a stream $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots$, the sequential risk of $A$ is

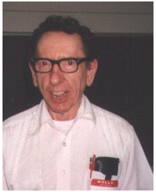
$$\boxed{\sum_{t=1}^{T} \ell(y_t, h_t(\boldsymbol{x}_t))}$$

## Regret

$$R_T = \sum_{t=1}^{T} \ell(y_t, h_t(\boldsymbol{x}_t)) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(y_t, h(\boldsymbol{x}_t))$$

- A sequential counterpart to the variance error in statistical learning
- Can we ensure $\frac{R_T}{T} \to 0$ as $T \to \infty$ for all streams?

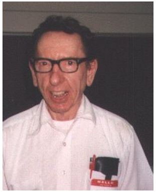
# Online learning as a repeated game



## Learning to play a game (1956)

- Theory of repeated games pioneered by James Hannan and David Blackwell

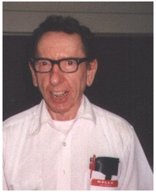
# Online learning as a repeated game



## Learning to play a game (1956)

- Theory of repeated games pioneered by James Hannan and David Blackwell
- Play a game repeatedly against a possibly suboptimal opponent (a.k.a. the data stream)

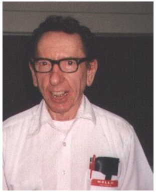
# Online learning as a repeated game



## Learning to play a game (1956)

- Theory of repeated games pioneered by James Hannan and David Blackwell
- Play a game repeatedly against a possibly suboptimal opponent (a.k.a. the data stream)
- Replace data stream with sequence of loss functions, e.g., $\ell_t(h_t) = \ell(y_t, h_t(\boldsymbol{x}_t))$

# Online learning as a repeated game



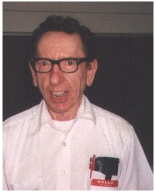
## Learning to play a game (1956)

- Theory of repeated games pioneered by James Hannan and David Blackwell
- Play a game repeatedly against a possibly suboptimal opponent (a.k.a. the data stream)
- Replace data stream with sequence of loss functions, e.g., $\ell_t(h_t) = \ell(y_t, h_t(\boldsymbol{x}_t))$

## Online learning in the simplex

- Let $\mathcal{H}$ be the $d$-dimensional simplex $\Delta_d$

# Online learning as a repeated game



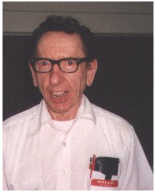
## Learning to play a game (1956)

- Theory of repeated games pioneered by James Hannan and David Blackwell
- Play a game repeatedly against a possibly suboptimal opponent (a.k.a. the data stream)
- Replace data stream with sequence of loss functions, e.g., $\ell_t(h_t) = \ell(y_t, h_t(\boldsymbol{x}_t))$

## Online learning in the simplex

- Let $\mathcal{H}$ be the $d$-dimensional simplex $\Delta_d$
- The loss at time $t$ of $\boldsymbol{p}_t \in \Delta_d$ is $\boldsymbol{\ell}_t^\top \boldsymbol{p}_t = \mathbb{E}[I_t]$ for $I_t \sim \boldsymbol{p}_t$

# Online learning as a repeated game



## Learning to play a game (1956)

- Theory of repeated games pioneered by James Hannan and David Blackwell
- Play a game repeatedly against a possibly suboptimal opponent (a.k.a. the data stream)
- Replace data stream with sequence of loss functions, e.g., $\ell_t(h_t) = \ell(y_t, h_t(\boldsymbol{x}_t))$

## Online learning in the simplex

- Let $\mathcal{H}$ be the $d$-dimensional simplex $\Delta_d$
- The loss at time $t$ of $\boldsymbol{p}_t \in \Delta_d$ is $\boldsymbol{\ell}_t^\top \boldsymbol{p}_t = \mathbb{E}[I_t]$ for $I_t \sim \boldsymbol{p}_t$
- This is a linear loss with bounded coefficients $\ell_t(i) \in [0, 1]$

# Prediction with expert advice

## A sequential decision problem

- $d$ actions
- Unknown deterministic assignment of losses to actions $\boldsymbol{\ell}_t = (\ell_t(1), \ldots, \ell_t(d)) \in [0,1]^d$ for each time step $t$

? ? ? ? ? ? ? ? ? ?

For $t = 1, 2, \ldots$

# Prediction with expert advice

## A sequential decision problem

- $d$ actions
- Unknown deterministic assignment of losses to actions $\boldsymbol{\ell}_t = (\ell_t(1), \ldots, \ell_t(d)) \in [0,1]^d$ for each time step $t$


## For $t = 1, 2, \ldots$

1. Player picks an action $I_t$ (possibly using randomization) and incurs loss $\ell_t(I_t)$

# Prediction with expert advice

## A sequential decision problem

- $d$ actions
- Unknown deterministic assignment of losses to actions $\boldsymbol{\ell}_t = (\ell_t(1), \ldots, \ell_t(d)) \in [0,1]^d$ for each time step $t$


For $t = 1, 2, \ldots$

1. Player picks an action $I_t$ (possibly using randomization) and incurs loss $\ell_t(I_t)$
2. Player gets feedback information: $\ell_t(1), \ldots, \ell_t(d)$

# Regret

$$R_T = \sum_{t=1}^{T} \boldsymbol{\ell}_t^\top \boldsymbol{p}_t - \min_{\boldsymbol{p} \in \Delta_d} \sum_{t=1}^{T} \boldsymbol{\ell}_t^\top \boldsymbol{p}$$

# Regret

$$R_T = \sum_{t=1}^{T} \boldsymbol{\ell}_t^\top \boldsymbol{p}_t - \min_{\boldsymbol{p} \in \Delta_d} \sum_{t=1}^{T} \boldsymbol{\ell}_t^\top \boldsymbol{p} = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \min_{i=1,\ldots,d} \sum_{t=1}^{T} \ell_t(i)$$

# Regret

$$R_T = \sum_{t=1}^{T} \boldsymbol{\ell}_t^\top \boldsymbol{p}_t - \min_{\boldsymbol{p} \in \Delta_d} \sum_{t=1}^{T} \boldsymbol{\ell}_t^\top \boldsymbol{p} = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \min_{i=1,\dots,d} \sum_{t=1}^{T} \ell_t(i)$$

Lower bound using a statistical learning argument

# Regret

$$R_T = \sum_{t=1}^{T} \boldsymbol{\ell}_t^\top \boldsymbol{p}_t - \min_{\boldsymbol{p}\in\Delta_d} \sum_{t=1}^{T} \boldsymbol{\ell}_t^\top \boldsymbol{p} = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \min_{i=1,\dots,d} \sum_{t=1}^{T} \ell_t(i)$$

Lower bound using a statistical learning argument

- $\ell_t(i) \to L_t(i) \in \{0,1\}$ independent random coin flip

# Regret

$$R_T = \sum_{t=1}^T \boldsymbol{\ell}_t^\top \boldsymbol{p}_t - \min_{\boldsymbol{p}\in\Delta_d} \sum_{t=1}^T \boldsymbol{\ell}_t^\top \boldsymbol{p} = \mathbb{E}\left[\sum_{t=1}^T \ell_t(I_t)\right] - \min_{i=1,\dots,d} \sum_{t=1}^T \ell_t(i)$$

## Lower bound using a statistical learning argument

- $\ell_t(i) \to L_t(i) \in \{0,1\}$ independent random coin flip
- For any player strategy $\quad \mathbb{E}\left[\sum_{t=1}^T L_t(I_t)\right] = \frac{T}{2}$

# Regret

$$R_T = \sum_{t=1}^{T} \boldsymbol{\ell}_t^\top \boldsymbol{p}_t - \min_{\boldsymbol{p}\in\Delta_d} \sum_{t=1}^{T} \boldsymbol{\ell}_t^\top \boldsymbol{p} = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \min_{i=1,\dots,d} \sum_{t=1}^{T} \ell_t(i)$$

## Lower bound using a statistical learning argument

- $\ell_t(i) \to L_t(i) \in \{0,1\}$ independent random coin flip
- For any player strategy $\mathbb{E}\left[\sum_{t=1}^{T} L_t(I_t)\right] = \frac{T}{2}$
- Then the expected regret is

$$\mathbb{E}\left[\max_{i=1,\dots,d} \sum_{t=1}^{T} \left(\frac{1}{2} - L_t(i)\right)\right] = (1-o(1))\sqrt{\frac{T \ln d}{2}}$$

for $d, T \to \infty$

# Exponentially weighted forecaster (Hedge)

At time $t$ pick action $I_t = i$ with probability proportional to

$$\exp\left(-\eta \sum_{s=1}^{t-1} \ell_s(i)\right)$$

the sum at the exponent is the total loss of action $i$ up to the previous time step

# Exponentially weighted forecaster (Hedge)

At time $t$ pick action $I_t = i$ with probability proportional to

$$\exp\left(-\eta \sum_{s=1}^{t-1} \ell_s(i)\right)$$

the sum at the exponent is the total loss of action $i$ up to the previous time step

## Regret bound

# Exponentially weighted forecaster (Hedge)

At time $t$ pick action $I_t = i$ with probability proportional to

$$\exp\left(-\eta \sum_{s=1}^{t-1} \ell_s(i)\right)$$

the sum at the exponent is the total loss of action $i$ up to the previous time step

## Regret bound

- If $\eta = \sqrt{\frac{\ln d}{8T}}$ then $\boxed{R_T \le \sqrt{\frac{T \ln d}{2}}}$

# Exponentially weighted forecaster (Hedge)

At time $t$ pick action $I_t = i$ with probability proportional to

$$\exp\left(-\eta \sum_{s=1}^{t-1} \ell_s(i)\right)$$

the sum at the exponent is the total loss of action $i$ up to the previous time step

## Regret bound

- If $\eta = \sqrt{\frac{\ln d}{8T}}$ then $\boxed{R_T \le \sqrt{\frac{T \ln d}{2}}}$
- This matches the asymptotic lower bound, including constants

# Exponentially weighted forecaster (Hedge)

At time $t$ pick action $I_t = i$ with probability proportional to

$$\exp\left(-\eta \sum_{s=1}^{t-1} \ell_s(i)\right)$$

the sum at the exponent is the total loss of action $i$ up to the previous time step

## Regret bound

- If $\eta = \sqrt{\frac{\ln d}{8T}}$ then $\boxed{R_T \le \sqrt{\frac{T \ln d}{2}}}$
- This matches the asymptotic lower bound, including constants
- We prove this later in a more general setting

# The bandit problem: playing an unknown game

- $d$ actions
- Unknown deterministic assignment of losses to actions $\boldsymbol{\ell}_t = (\ell_t(1), \dots, \ell_t(d)) \in [0,1]^d$ for each time step $t$

? ? ? ? ? ? ? ? ? ?

For $t = 1, 2, \dots$

# The bandit problem: playing an unknown game

- $d$ actions
- Unknown deterministic assignment of losses to actions $\boldsymbol{\ell}_t = (\ell_t(1), \dots, \ell_t(d)) \in [0,1]^d$ for each time step $t$

? ? ? ? ? ? ? ? ? ?

For $t = 1, 2, \dots$

1. Player picks an action $I_t$ (possibly using randomization) and incurs loss $\ell_t(I_t)$

# The bandit problem: playing an unknown game

- $d$ actions
- Unknown deterministic assignment of losses to actions $\boldsymbol{\ell}_t = (\ell_t(1), \dots, \ell_t(d)) \in [0,1]^d$ for each time step $t$

?   .3   ?   ?   ?   ?   ?   ?   ?   ?

For $t = 1, 2, \dots$

1. Player picks an action $I_t$ (possibly using randomization) and incurs loss $\ell_t(I_t)$
2. Player gets feedback information: Only $\ell_t(I_t)$ is revealed

# A growing range of applications

- Ad placement

# A growing range of applications

- Ad placement
- Dynamic content/layout optimization

# A growing range of applications

- Ad placement
- Dynamic content/layout optimization
- Real time bidding

# A growing range of applications

- Ad placement
- Dynamic content/layout optimization
- Real time bidding
- Recommender systems

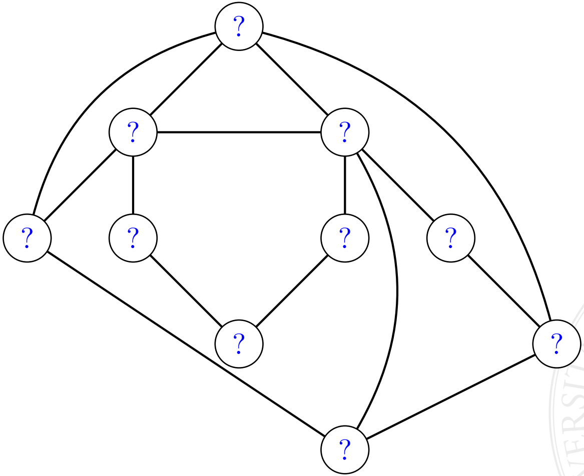
# A growing range of applications

- Ad placement
- Dynamic content/layout optimization
- Real time bidding
- Recommender systems
- Clinical trials

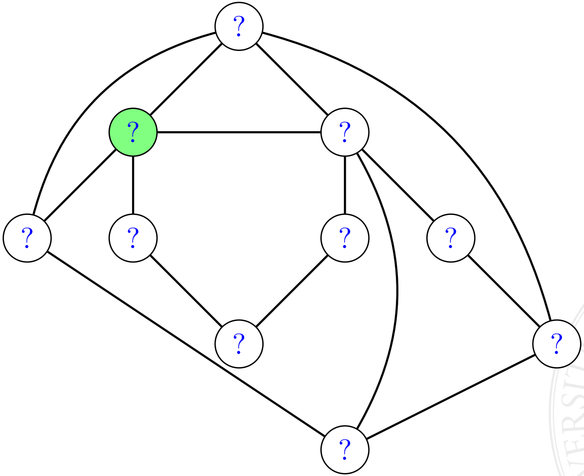
# A growing range of applications

- Ad placement
- Dynamic content/layout optimization
- Real time bidding
- Recommender systems
- Clinical trials
- Network protocol optimization

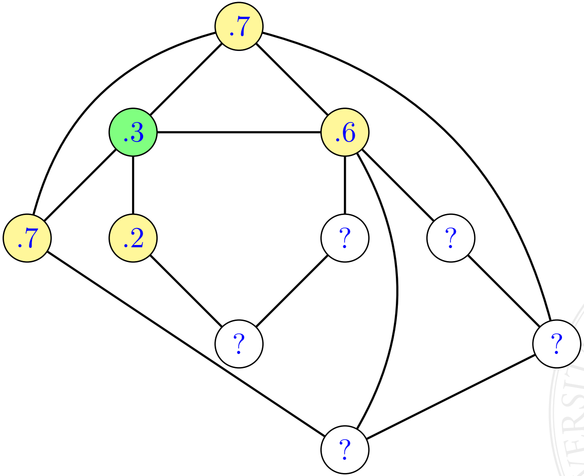
# An observability graph over actions



# An observability graph over actions



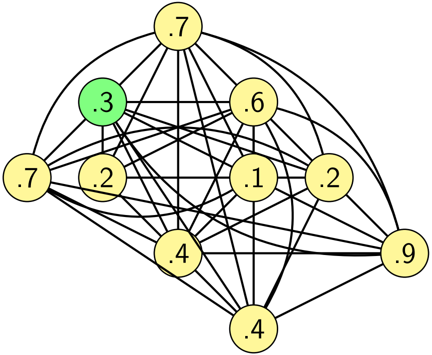
# An observability graph over actions



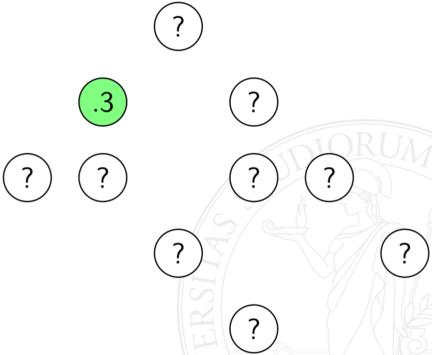

$\ell_t(i)$ is observed iff $I_t \in \{i\} \cup \mathcal{N}_G(i)$

# Recovering expert and bandit settings

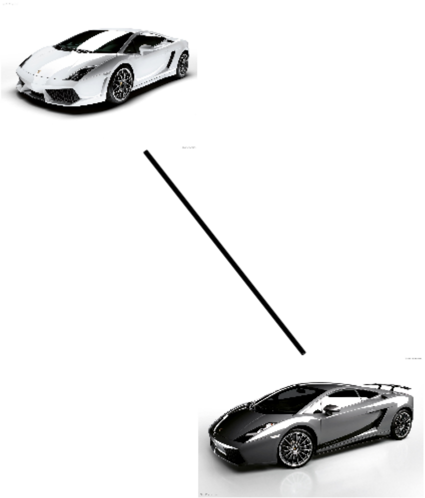
Experts: clique

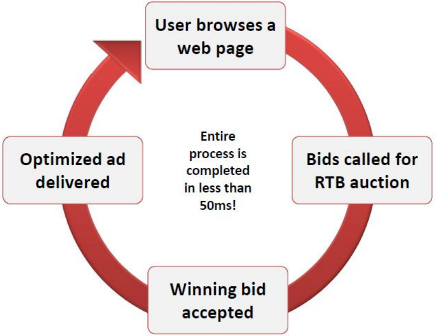



Bandits: edgeless graph



# Relationships between actions





# Hedge revisited on an observability graph $G$

## Player's strategy must use loss estimates

- $$p_t(i) \quad \propto \quad \exp\left(-\eta \sum_{s=1}^{t-1} \widehat{\ell}_s(i)\right) \qquad i = 1, \dots, d$$

# Hedge revisited on an observability graph $G$

Player's strategy must use loss estimates

- $p_t(i) \quad \propto \quad \exp\left(-\eta \sum_{s=1}^{t-1} \widehat{\ell}_s(i)\right) \qquad i = 1, \dots, d$
- $\widehat{\ell}_t(i) = \begin{cases} \dfrac{\ell_t(i)}{\mathbb{P}_t\big(\ell_t(i) \text{ observed}\big)} & \text{if } \ell_t(i) \text{ is observed because } I_t \in \{i\} \cup \mathcal{N}_G(i) \\ 0 & \text{otherwise} \end{cases}$

# Hedge revisited on an observability graph $G$

Player's strategy must use loss estimates

- $p_t(i) \quad \propto \quad \exp\left(-\eta \sum_{s=1}^{t-1} \widehat{\ell}_s(i)\right) \qquad i = 1, \dots, d$
- $\widehat{\ell}_t(i) = \begin{cases} \dfrac{\ell_t(i)}{\mathbb{P}_t(\ell_t(i) \text{ observed})} & \text{if } \ell_t(i) \text{ is observed because } I_t \in \{i\} \cup \mathcal{N}_G(i) \\ 0 & \text{otherwise} \end{cases}$

Importance sampling estimator

# Hedge revisited on an observability graph $G$

## Player's strategy must use loss estimates

- $p_t(i) \quad \propto \quad \exp\left(-\eta \sum_{s=1}^{t-1} \widehat{\ell}_s(i)\right) \qquad i = 1, \dots, d$
- $\widehat{\ell}_t(i) = \begin{cases} \dfrac{\ell_t(i)}{\mathbb{P}_t(\ell_t(i) \text{ observed})} & \text{if } \ell_t(i) \text{ is observed because } I_t \in \{i\} \cup \mathcal{N}_G(i) \\ 0 & \text{otherwise} \end{cases}$

## Importance sampling estimator

$$\mathbb{E}_t\left[\widehat{\ell}_t(i)\right] = \frac{\ell_t(i)}{\mathbb{P}_t(\ell_t(i) \text{ observed})} \times \mathbb{P}_t(\ell_t(i) \text{ observed}) + 0 = \ell_t(i)$$

$$\mathbb{E}_t\left[\widehat{\ell}_t(i)^2\right] = \frac{\ell_t(i)^2}{\mathbb{P}_t(\ell_t(i) \text{ observed})^2} \times \mathbb{P}_t(\ell_t(i) \text{ observed}) + 0 = \frac{\ell_t(i)^2}{\mathbb{P}_t(\ell_t(i) \text{ observed})}$$

# Hedge revisited on an observability graph $G$

## Player's strategy must use loss estimates

- $p_t(i) \quad \propto \quad \exp\left(-\eta \sum_{s=1}^{t-1} \widehat{\ell}_s(i)\right) \qquad i = 1, \dots, d$
- $\widehat{\ell}_t(i) = \begin{cases} \dfrac{\ell_t(i)}{\mathbb{P}_t(\ell_t(i) \text{ observed})} & \text{if } \ell_t(i) \text{ is observed because } I_t \in \{i\} \cup \mathcal{N}_G(i) \\ 0 & \text{otherwise} \end{cases}$

## Importance sampling estimator

$$\mathbb{E}_t\left[\widehat{\ell}_t(i)\right] = \frac{\ell_t(i)}{\mathbb{P}_t(\ell_t(i) \text{ observed})} \times \mathbb{P}_t(\ell_t(i) \text{ observed}) + 0 = \ell_t(i)$$

$$\mathbb{E}_t\left[\widehat{\ell}_t(i)^2\right] = \frac{\ell_t(i)^2}{\mathbb{P}_t(\ell_t(i) \text{ observed})^2} \times \mathbb{P}_t(\ell_t(i) \text{ observed}) + 0 \le \frac{1}{\mathbb{P}_t(\ell_t(i) \text{ observed})}$$

# Regret analysis

$$\frac{W_{t+1}}{W_t} = \sum_{i=1}^{d} \frac{w_{t+1}(i)}{W_t} \qquad p_t(i) = \frac{1}{W_t} \exp\left(-\eta \sum_{s=1}^{t-1} \widehat{\ell}_s(i)\right) = \frac{w_t(i)}{W_t} \quad \text{is a r.v.!}$$

# Regret analysis

$$\frac{W_{t+1}}{W_t} = \sum_{i=1}^{d} \frac{w_{t+1}(i)}{W_t} \qquad p_t(i) = \frac{1}{W_t} \exp\left(-\eta \sum_{s=1}^{t-1} \widehat{\ell}_s(i)\right) = \frac{w_t(i)}{W_t} \quad \text{is a r.v.!}$$

$$= \sum_{i=1}^{d} \frac{w_t(i)}{W_t} \exp\left(-\eta\, \widehat{\ell}_t(i)\right) \qquad \text{(because } w_{t+1}(i) = e^{-\eta \sum_{s=1}^{t-1} \hat{\ell}_s(i) - \eta \hat{\ell}_t(i)}\text{)}$$

# Regret analysis

$$\frac{W_{t+1}}{W_t} = \sum_{i=1}^{d} \frac{w_{t+1}(i)}{W_t} \qquad p_t(i) = \frac{1}{W_t} \exp\left(-\eta \sum_{s=1}^{t-1} \widehat{\ell}_s(i)\right) = \frac{w_t(i)}{W_t} \quad \text{is a r.v.!}$$

$$= \sum_{i=1}^{d} \frac{w_t(i)}{W_t} \exp\bigl(-\eta\, \widehat{\ell}_t(i)\bigr) \qquad \text{(because } w_{t+1}(i) = e^{-\eta \sum_{s=1}^{t-1} \hat{\ell}_s(i) - \eta \hat{\ell}_t(i)}\text{)}$$

$$= \sum_{i=1}^{d} p_t(i) \exp\bigl(-\eta\, \widehat{\ell}_t(i)\bigr)$$

# Regret analysis

$$\frac{W_{t+1}}{W_t} = \sum_{i=1}^{d} \frac{w_{t+1}(i)}{W_t} \qquad p_t(i) = \frac{1}{W_t} \exp\left(-\eta \sum_{s=1}^{t-1} \widehat{\ell}_s(i)\right) = \frac{w_t(i)}{W_t} \quad \text{is a r.v.!}$$

$$= \sum_{i=1}^{d} \frac{w_t(i)}{W_t} \exp\big(-\eta\, \widehat{\ell}_t(i)\big) \qquad \text{(because } w_{t+1}(i) = e^{-\eta \sum_{s=1}^{t-1} \hat{\ell}_s(i) - \eta \hat{\ell}_t(i)}\text{)}$$

$$= \sum_{i=1}^{d} p_t(i) \exp\big(-\eta\, \widehat{\ell}_t(i)\big)$$

$$\leq \sum_{i=0}^{d} p_t(i) \left(1 - \eta\, \widehat{\ell}_t(i) + \frac{\big(\eta\, \widehat{\ell}_t(i)\big)^2}{2}\right) \qquad \text{(using } e^{-x} \leq 1 - x + x^2/2 \text{ for all } x \geq 0\text{)}$$

# Regret analysis

$$\frac{W_{t+1}}{W_t} = \sum_{i=1}^{d} \frac{w_{t+1}(i)}{W_t} \qquad p_t(i) = \frac{1}{W_t} \exp\left(-\eta \sum_{s=1}^{t-1} \widehat{\ell}_s(i)\right) = \frac{w_t(i)}{W_t} \quad \text{is a r.v.!}$$

$$= \sum_{i=1}^{d} \frac{w_t(i)}{W_t} \exp(-\eta\, \widehat{\ell}_t(i)) \qquad \text{(because } w_{t+1}(i) = e^{-\eta \sum_{s=1}^{t-1} \hat{\ell}_s(i) - \eta \hat{\ell}_t(i)}\text{)}$$

$$= \sum_{i=1}^{d} p_t(i) \exp(-\eta\, \widehat{\ell}_t(i))$$

$$\leq \sum_{i=0}^{d} p_t(i) \left(1 - \eta\, \widehat{\ell}_t(i) + \frac{(\eta\, \widehat{\ell}_t(i))^2}{2}\right) \qquad \text{(using } e^{-x} \leq 1 - x + x^2/2 \text{ for all } x \geq 0\text{)}$$

$$\leq 1 - \eta \sum_{i=1}^{d} p_t(i) \widehat{\ell}_t(i) + \frac{\eta^2}{2} \sum_{i=1}^{d} p_t(i) \widehat{\ell}_t(i)^2$$

# Regret analysis (cont.)

Taking logs, using $\ln(1+x) \leq x$, and summing over $t = 1, \dots, T$ yields

$$\ln \frac{W_{T+1}}{W_1} \leq -\eta \sum_{t=1}^{T} \sum_{i=1}^{d} p_t(i) \widehat{\ell}_t(i) + \frac{\eta^2}{2} \sum_{t=1}^{T} \sum_{i=1}^{d} p_t(i) \widehat{\ell}_t(i)^2$$

## Regret analysis (cont.)

Taking logs, using $\ln(1+x) \le x$, and summing over $t = 1, \dots, T$ yields

$$\ln \frac{W_{T+1}}{W_1} \le -\eta \sum_{t=1}^{T} \sum_{i=1}^{d} p_t(i) \widehat{\ell}_t(i) + \frac{\eta^2}{2} \sum_{t=1}^{T} \sum_{i=1}^{d} p_t(i) \widehat{\ell}_t(i)^2$$

Moreover, for any fixed action $k$, we also have

$$\ln \frac{W_{T+1}}{W_1} \ge \ln \frac{w_{T+1}(k)}{W_1} = -\eta \sum_{t=1}^{T} \widehat{\ell}_t(k) - \ln d$$

# Regret analysis (cont.)

Taking logs, using $\ln(1+x) \leq x$, and summing over $t = 1, \ldots, T$ yields

$$\ln \frac{W_{T+1}}{W_1} \leq -\eta \sum_{t=1}^{T} \sum_{i=1}^{d} p_t(i) \widehat{\ell}_t(i) + \frac{\eta^2}{2} \sum_{t=1}^{T} \sum_{i=1}^{d} p_t(i) \widehat{\ell}_t(i)^2$$

Moreover, for any fixed action $k$, we also have

$$\ln \frac{W_{T+1}}{W_1} \geq \ln \frac{w_{T+1}(k)}{W_1} = -\eta \sum_{t=1}^{T} \widehat{\ell}_t(k) - \ln d$$

Putting together and dividing both sides by $\eta > 0$ gives

$$\sum_{t=1}^{T} \sum_{i=1}^{d} p_t(i) \widehat{\ell}_t(i) - \sum_{t=1}^{T} \widehat{\ell}_t(k) \leq \frac{\ln d}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{i=1}^{d} p_t(i) \widehat{\ell}_t(i)^2$$

# Regret analysis (cont.)

Recall where we were:

$$\sum_{t=1}^{T}\sum_{i=1}^{d} p_t(i)\widehat{\ell}_t(i) - \sum_{t=1}^{T}\widehat{\ell}_t(k) \le \frac{\ln d}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{i=1}^{d} p_t(i)\widehat{\ell}_t(i)^2$$

# Regret analysis (cont.)

Recall where we were:

$$\sum_{t=1}^{T}\sum_{i=1}^{d} p_t(i)\widehat{\ell}_t(i) - \sum_{t=1}^{T}\widehat{\ell}_t(k) \le \frac{\ln d}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{i=1}^{d} p_t(i)\widehat{\ell}_t(i)^2$$

Take expectation w.r.t. $I_1, \ldots, I_T$

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{d} p_t(i)\mathbb{E}_t\left[\widehat{\ell}_t(i)\right] - \sum_{t=1}^{T}\mathbb{E}_t\left[\widehat{\ell}_t(k)\right]\right] \le \frac{\ln d}{\eta} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{d} p_t(i)\mathbb{E}_t\left[\widehat{\ell}_t(i)^2\right]\right]$$

# Regret analysis (cont.)

Recall where we were:

$$\sum_{t=1}^{T}\sum_{i=1}^{d} p_t(i)\widehat{\ell}_t(i) - \sum_{t=1}^{T}\widehat{\ell}_t(k) \le \frac{\ln d}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{i=1}^{d} p_t(i)\widehat{\ell}_t(i)^2$$

Take expectation w.r.t. $I_1, \dots, I_T$

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{d} p_t(i)\mathbb{E}_t\left[\widehat{\ell}_t(i)\right] - \sum_{t=1}^{T}\mathbb{E}_t\left[\widehat{\ell}_t(k)\right]\right] \le \frac{\ln d}{\eta} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{d} p_t(i)\mathbb{E}_t\left[\widehat{\ell}_t(i)^2\right]\right]$$

Loss estimates are unbiased:

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{d} p_t(i)\ell_t(i) - \sum_{t=1}^{T}\ell_t(k)\right] \le \frac{\ln d}{\eta} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{d} p_t(i)\mathbb{E}_t\left[\widehat{\ell}_t(i)^2\right]\right]$$

# Regret analysis (cont.)

Recall where we were:

$$\sum_{t=1}^{T}\sum_{i=1}^{d} p_t(i)\widehat{\ell}_t(i) - \sum_{t=1}^{T}\widehat{\ell}_t(k) \leq \frac{\ln d}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{i=1}^{d} p_t(i)\widehat{\ell}_t(i)^2$$

Take expectation w.r.t. $I_1, \ldots, I_T$

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{d} p_t(i)\mathbb{E}_t\left[\widehat{\ell}_t(i)\right] - \sum_{t=1}^{T}\mathbb{E}_t\left[\widehat{\ell}_t(k)\right]\right] \leq \frac{\ln d}{\eta} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{d} p_t(i)\mathbb{E}_t\left[\widehat{\ell}_t(i)^2\right]\right]$$

This is just the regret

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{d} p_t(i)\ell_t(i) - \sum_{t=1}^{T}\ell_t(k)\right] \leq \frac{\ln d}{\eta} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{d} p_t(i)\mathbb{E}_t\left[\widehat{\ell}_t(i)^2\right]\right]$$

# Regret analysis (cont.)

$$R_T \le \frac{\ln d}{\eta} + \frac{\eta}{2}\,\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{d} p_t(i)\mathbb{E}_t\left[\widehat{\ell}_t(i)^2\right]\right]$$

# Regret analysis (cont.)

$$
\begin{aligned}
R_T &\le \frac{\ln d}{\eta} + \frac{\eta}{2}\,\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{d} p_t(i)\mathbb{E}_t\left[\widehat{\ell}_t(i)^2\right]\right] \\
&\le \frac{\ln d}{\eta} + \frac{\eta}{2}\,\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{d} \frac{p_t(i)}{\mathbb{P}_t\big(\ell_t(i) \text{ is observed}\big)}\right] \qquad \text{(variance bound)}
\end{aligned}
$$

# Regret analysis (cont.)

$$
\begin{aligned}
R_T &\le \frac{\ln d}{\eta} + \frac{\eta}{2}\,\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{d} p_t(i)\,\mathbb{E}_t\left[\widehat{\ell}_t(i)^2\right]\right] \\
&\le \frac{\ln d}{\eta} + \frac{\eta}{2}\,\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{d} \frac{p_t(i)}{\mathbb{P}_t\big(\ell_t(i)\text{ is observed}\big)}\right] && \text{(variance bound)} \\
&= \frac{\ln d}{\eta} + \frac{\eta}{2}\,\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{d} \frac{p_t(i)}{p_t(i) + \sum_{j\in\mathcal{N}_G(i)} p_t(j)}\right] && \text{(observability condition)}
\end{aligned}
$$

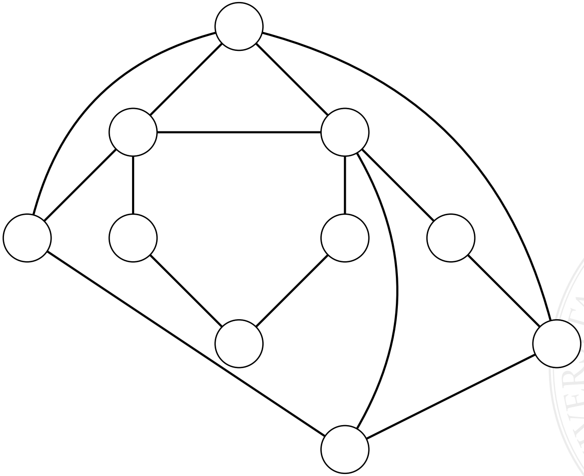
# Regret analysis (cont.)

$$
\begin{aligned}
R_T &\le \frac{\ln d}{\eta} + \frac{\eta}{2}\,\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{d} p_t(i)\,\mathbb{E}_t\left[\widehat{\ell}_t(i)^2\right]\right] \\
&\le \frac{\ln d}{\eta} + \frac{\eta}{2}\,\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{d} \frac{p_t(i)}{\mathbb{P}_t\big(\ell_t(i) \text{ is observed}\big)}\right] && \text{(variance bound)} \\
&= \frac{\ln d}{\eta} + \frac{\eta}{2}\,\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{d} \frac{p_t(i)}{p_t(i) + \sum_{j\in\mathcal{N}_G(i)} p_t(j)}\right] && \text{(observability condition)} \\
&\le \frac{\ln d}{\eta} + \frac{\eta}{2} T\,\alpha(G) && \text{(cool graph-theoretic fact)}
\end{aligned}
$$

$\alpha(G)$ is the independence number of $G$

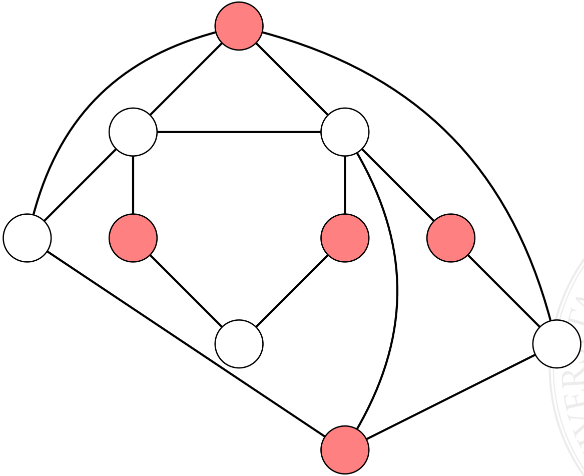
## Independence number $\alpha(G)$

The size of the largest independent set in $G$



# Independence number $\alpha(G)$

The size of the largest independent set in $G$



# Regret bound

$$R_T \le \frac{\ln d}{\eta} + \frac{\eta}{2} T \alpha(G)$$

# Regret bound

$$R_T \leq \frac{\ln d}{\eta} + \frac{\eta}{2} T \alpha(G) = \sqrt{T \alpha(G) \ln d}$$

# Regret bound

$$R_T \leq \frac{\ln d}{\eta} + \frac{\eta}{2} T\alpha(G) = \sqrt{T\alpha(G) \ln d}$$

Note: This bound is tight for all $G$ (up to logarithmic factors)

# Regret bound

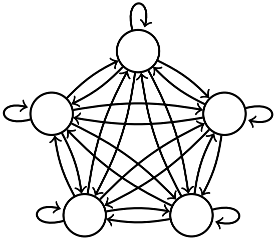
$$R_T \leq \frac{\ln d}{\eta} + \frac{\eta}{2} T\alpha(G) = \sqrt{T\alpha(G)\ln d}$$

Note: This bound is tight for all $G$ (up to logarithmic factors)

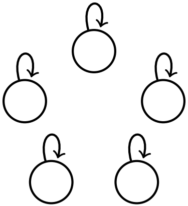
## Special cases

| | | | |
|---|---|---|---|
| Experts (clique): | $\alpha(G) = 1$ | $R_T \leq \sqrt{T \ln d}$ | Hedge algorithm |
| Bandits (edgeless graph): | $\alpha(G) = d$ | $R_T \leq \sqrt{Td \ln d}$ | Exp3 algorithm |

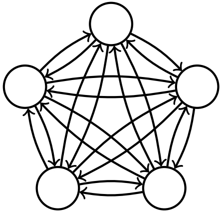
# More general feedback models



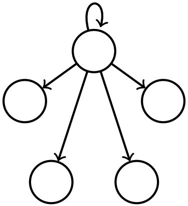
Experts



Bandits



Cops & Robbers



Revealing Action

# Partial monitoring: not observing your own loss

## Dynamic pricing: Perform as the best fixed price

1. Post a T-shirt price
2. Observe if next customer buys or not
3. Adjust price

Feedback does not reveal the player's loss

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | 1 | 2 | 3 | 4 |
| 2 | $c$ | 0 | 1 | 2 | 3 |
| 3 | $c$ | $c$ | 0 | 1 | 2 |
| 4 | $c$ | $c$ | $c$ | 0 | 1 |
| 5 | $c$ | $c$ | $c$ | $c$ | 0 |

Loss

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 0 | 1 | 1 | 1 | 1 |
| 3 | 0 | 0 | 1 | 1 | 1 |
| 4 | 0 | 0 | 0 | 1 | 1 |
| 5 | 0 | 0 | 0 | 0 | 1 |

Feedback

## A general gap theorem

- A constructive characterization of the minimax regret for any partial monitoring game

# A general gap theorem

- A constructive characterization of the minimax regret for any partial monitoring game
- Only three possible rates for nontrivial games:

# A general gap theorem

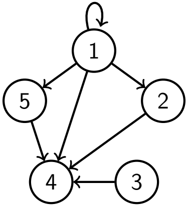
- A constructive characterization of the minimax regret for any partial monitoring game
- Only three possible rates for nontrivial games:
  1. Easy games (e.g., experts, bandits, cops & robbers): $\Theta(\sqrt{T})$

# A general gap theorem

- A constructive characterization of the minimax regret for any partial monitoring game
- Only three possible rates for nontrivial games:
    1. Easy games (e.g., experts, bandits, cops & robbers): $\Theta(\sqrt{T})$
    2. Hard games (e.g., revealing action, dynamic pricing): $\Theta(T^{2/3})$

# A general gap theorem

- A constructive characterization of the minimax regret for any partial monitoring game
- Only three possible rates for nontrivial games:
  1. Easy games (e.g., experts, bandits, cops & robbers): $\Theta(\sqrt{T})$
  2. Hard games (e.g., revealing action, dynamic pricing): $\Theta(T^{2/3})$
  3. Impossible games: $\Theta(T)$



# Contextual bandits

- A policy $\pi$ maps side information (e.g., feature vectors $\boldsymbol{x}_t$) to probabilistic decisions $\pi(\boldsymbol{x}_t) \in \Delta_d$

# Contextual bandits

- A policy $\pi$ maps side information (e.g., feature vectors $\boldsymbol{x}_t$) to probabilistic decisions $\pi(\boldsymbol{x}_t) \in \Delta_d$
- Consider a finite set $\Pi$ of such policies

# Contextual bandits

- A policy $\pi$ maps side information (e.g., feature vectors $\boldsymbol{x}_t$) to probabilistic decisions $\pi(\boldsymbol{x}_t) \in \Delta_d$
- Consider a finite set $\Pi$ of such policies
- Regret against best policy: $R_T^{\text{pol}} = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \min_{\pi \in \Pi} \sum_{t=1}^{T} \ell_t(\pi(\boldsymbol{x}_t))$

# Contextual bandits

- A policy $\pi$ maps side information (e.g., feature vectors $\boldsymbol{x}_t$) to probabilistic decisions $\pi(\boldsymbol{x}_t) \in \Delta_d$
- Consider a finite set $\Pi$ of such policies
- Regret against best policy: $R_T^{\mathrm{pol}} = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \min_{\pi \in \Pi} \sum_{t=1}^{T} \ell_t(\pi(\boldsymbol{x}_t))$
- Exp4 (a variant of Exp3) selects actions $I_t$ based on $\{\pi(\boldsymbol{x}_t) \,:\, \pi \in \Pi\}$

# Contextual bandits

- A policy $\pi$ maps side information (e.g., feature vectors $\boldsymbol{x}_t$) to probabilistic decisions $\pi(\boldsymbol{x}_t) \in \Delta_d$
- Consider a finite set $\Pi$ of such policies
- Regret against best policy: $R_T^{\mathrm{pol}} = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \min_{\pi \in \Pi} \sum_{t=1}^{T} \ell_t(\pi(\boldsymbol{x}_t))$
- Exp4 (a variant of Exp3) selects actions $I_t$ based on $\{\pi(\boldsymbol{x}_t) \,:\, \pi \in \Pi\}$
- Regret bound: $R_T^{\mathrm{pol}} \le \sqrt{Td \ln |\Pi|}$ (with bandit feedback)

# Contextual bandits

- A policy $\pi$ maps side information (e.g., feature vectors $\boldsymbol{x}_t$) to probabilistic decisions $\pi(\boldsymbol{x}_t) \in \Delta_d$
- Consider a finite set $\Pi$ of such policies
- Regret against best policy: $R_T^{\mathrm{pol}} = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \min_{\pi \in \Pi} \sum_{t=1}^{T} \ell_t(\pi(\boldsymbol{x}_t))$
- Exp4 (a variant of Exp3) selects actions $I_t$ based on $\{\pi(\boldsymbol{x}_t) : \pi \in \Pi\}$
- Regret bound: $R_T^{\mathrm{pol}} \le \sqrt{Td \ln |\Pi|}$ (with bandit feedback)
- This holds for all loss sequences, sets of policies, and side information sequences

# Contextual bandits

- A policy $\pi$ maps side information (e.g., feature vectors $\boldsymbol{x}_t$) to probabilistic decisions $\pi(\boldsymbol{x}_t) \in \Delta_d$
- Consider a finite set $\Pi$ of such policies
- Regret against best policy: $R_T^{\mathrm{pol}} = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \min_{\pi\in\Pi} \sum_{t=1}^{T} \ell_t(\pi(\boldsymbol{x}_t))$
- Exp4 (a variant of Exp3) selects actions $I_t$ based on $\{\pi(\boldsymbol{x}_t) \,:\, \pi \in \Pi\}$
- Regret bound: $R_T^{\mathrm{pol}} \le \sqrt{Td \ln |\Pi|}$ (with bandit feedback)
- This holds for all loss sequences, sets of policies, and side information sequences
- Need time linear in $|\Pi|$ at each step

# Online convex optimization

Model space $\mathbb{V} \subseteq \mathbb{R}^d$ convex, closed, and nonempty

For $t = 1, 2, \ldots$

1. The current $h_t \in \mathcal{H}$ is tested on the next data point $(\boldsymbol{x}_t, y_t)$ in the stream
2. $A$ is charged with loss $\ell(y_t, h_t(\boldsymbol{x}_t))$
3. $h_{t+1}$ is computed based on $h_t$ and $(\boldsymbol{x}_t, y_t)$

# Online convex optimization

Model space $\mathbb{V} \subseteq \mathbb{R}^d$ convex, closed, and nonempty

For $t = 1, 2, \ldots$

1. The current $\boldsymbol{w} \in \mathbb{V}$ is tested on the next convex loss function $\ell_t$ in the stream
2. $A$ is charged with loss $\ell(y_t, h_t(\boldsymbol{x}_t))$
3. $h_{t+1}$ is computed based on $h_t$ and $(\boldsymbol{x}_t, y_t)$

# Online convex optimization

Model space $\mathbb{V} \subseteq \mathbb{R}^d$ convex, closed, and nonempty

For $t = 1, 2, \dots$

1. The current $\boldsymbol{w} \in \mathbb{V}$ is tested on the next convex loss function $\ell_t$ in the stream
2. $A$ is charged loss $\ell_t(\boldsymbol{w}_t)$
3. $h_{t+1}$ is computed based on $h_t$ and $(\boldsymbol{x}_t, y_t)$

# Online convex optimization

Model space $\mathbb{V} \subseteq \mathbb{R}^d$ convex, closed, and nonempty

For $t = 1, 2, \ldots$

1. The current $\boldsymbol{w} \in \mathbb{V}$ is tested on the next convex loss function $\ell_t$ in the stream
2. $A$ is charged loss $\ell_t(\boldsymbol{w}_t)$
3. $\boldsymbol{w}_{t+1}$ is computed based on $\boldsymbol{w}_t$ and $\nabla \ell_t(\boldsymbol{w}_t)$ (first-order oracle)

# Online convex optimization

Model space $\mathbb{V} \subseteq \mathbb{R}^d$ convex, closed, and nonempty

For $t = 1, 2, \ldots$

1. The current $\boldsymbol{w} \in \mathbb{V}$ is tested on the next convex loss function $\ell_t$ in the stream
2. $A$ is charged loss $\ell_t(\boldsymbol{w}_t)$
3. $\boldsymbol{w}_{t+1}$ is computed based on $\boldsymbol{w}_t$ and $\boxed{\nabla \ell_t(\boldsymbol{w}_t)}$ (first-order oracle)

Regret

$$R_T(\boldsymbol{u}) = \sum_{t=1}^{T} \ell_t(\boldsymbol{w}_t) - \sum_{t=1}^{T} \ell_t(\boldsymbol{u}) \qquad \boldsymbol{u} \in \mathbb{V}$$

# Online convex optimization

Model space $\mathbb{V} \subseteq \mathbb{R}^d$ convex, closed, and nonempty

For $t = 1, 2, \dots$

1. The current $\boldsymbol{w} \in \mathbb{V}$ is tested on the next convex loss function $\ell_t$ in the stream
2. $A$ is charged loss $\ell_t(\boldsymbol{w}_t)$
3. $\boldsymbol{w}_{t+1}$ is computed based on $\boldsymbol{w}_t$ and $\nabla \ell_t(\boldsymbol{w}_t)$ (first-order oracle)

Regret

$$R_T = \sum_{t=1}^{T} \ell_t(\boldsymbol{w}_t) - \inf_{\boldsymbol{u} \in \mathbb{V}} \sum_{t=1}^{T} \ell_t(\boldsymbol{u})$$

# Stochastic gradient descent

## Minimization of training error

$$\min_{\boldsymbol{w}\in\mathbb{V}} \sum_{i=1}^{m} \ell(\boldsymbol{w}, (\boldsymbol{x}_i, y_i))$$

$\ell(\boldsymbol{w}, (\boldsymbol{x}_i, y_i))$ measures the (convex) loss of $\boldsymbol{w}$ on the training example $(\boldsymbol{x}_i, y_i)$

# Stochastic gradient descent

## Minimization of training error

$$\min_{\boldsymbol{w}\in\mathbb{V}} \sum_{i=1}^{m} \ell(\boldsymbol{w},(\boldsymbol{x}_i,y_i))$$

$\ell(\boldsymbol{w},(\boldsymbol{x}_i,y_i))$ measures the (convex) loss of $\boldsymbol{w}$ on the training example $(\boldsymbol{x}_i,y_i)$

- When $m$ is large we cannot afford to spend more than constant time on each data point

# Stochastic gradient descent

## Minimization of training error

$$\min_{\boldsymbol{w}\in\mathbb{V}} \sum_{i=1}^{m} \ell(\boldsymbol{w},(\boldsymbol{x}_i,y_i))$$

$\ell(\boldsymbol{w},(\boldsymbol{x}_i,y_i))$ measures the (convex) loss of $\boldsymbol{w}$ on the training example $(\boldsymbol{x}_i,y_i)$

- When $m$ is large we cannot afford to spend more than constant time on each data point
- Online convex optimization can be used for stochastic optimization

# Stochastic gradient descent

## Minimization of training error

$$\min_{\boldsymbol{w} \in \mathbb{V}} \sum_{i=1}^{m} \ell(\boldsymbol{w}, (\boldsymbol{x}_i, y_i))$$

$\ell(\boldsymbol{w}, (\boldsymbol{x}_i, y_i))$ measures the (convex) loss of $\boldsymbol{w}$ on the training example $(\boldsymbol{x}_i, y_i)$

- When $m$ is large we cannot afford to spend more than constant time on each data point
- Online convex optimization can be used for stochastic optimization
- Draw $(\boldsymbol{X}_1, Y_1), (\boldsymbol{X}_2, Y_2) \dots$ uniformly i.i.d. from the training set

# Stochastic gradient descent

## Minimization of training error

$$\min_{\boldsymbol{w}\in\mathbb{V}} \sum_{i=1}^{m} \ell(\boldsymbol{w},(\boldsymbol{x}_i,y_i))$$

$\ell(\boldsymbol{w},(\boldsymbol{x}_i,y_i))$ measures the (convex) loss of $\boldsymbol{w}$ on the training example $(\boldsymbol{x}_i,y_i)$

- When $m$ is large we cannot afford to spend more than constant time on each data point
- Online convex optimization can be used for stochastic optimization
- Draw $(\boldsymbol{X}_1,Y_1),(\boldsymbol{X}_2,Y_2)\dots$ uniformly i.i.d. from the training set
- Run online algorithm on the sequence of loss functions $\ell_t = \ell_t(\cdot,(\boldsymbol{X}_t,Y_t))$

# Lower bounds

- $\mathbb{V}$ is a bounded set of diameter $D$ and all $\ell_t$ are Lipschitz with constant $L$

# Lower bounds

- $\mathbb{V}$ is a bounded set of diameter $D$ and all $\ell_t$ are Lipschitz with constant $L$
- Take $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{V}$ such that $\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2 = D$ and set $\boldsymbol{z}_0 = (\boldsymbol{v}_1 - \boldsymbol{v}_2) / \|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2$

# Lower bounds

- $\mathbb{V}$ is a bounded set of diameter $D$ and all $\ell_t$ are Lipschitz with constant $L$
- Take $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{V}$ such that $\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2 = D$ and set $\boldsymbol{z}_0 = (\boldsymbol{v}_1 - \boldsymbol{v}_2) / \|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2$
- Stochastic linear losses $L_t(\boldsymbol{w}) = \varepsilon_t L\, \boldsymbol{w}^\top \boldsymbol{z}_0$ where $\varepsilon_t \in \{-1, 1\}$ are uniform

# Lower bounds

- $\mathbb{V}$ is a bounded set of diameter $D$ and all $\ell_t$ are Lipschitz with constant $L$
- Take $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{V}$ such that $\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2 = D$ and set $\boldsymbol{z}_0 = (\boldsymbol{v}_1 - \boldsymbol{v}_2) / \|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2$
- Stochastic linear losses $L_t(\boldsymbol{w}) = \varepsilon_t L\, \boldsymbol{w}^\top \boldsymbol{z}_0$ where $\varepsilon_t \in \{-1, 1\}$ are uniform

$$\mathbb{E}\left[\max_{\boldsymbol{u} \in \{\boldsymbol{v}_1, \boldsymbol{v}_2\}} R_T(\boldsymbol{u})\right]$$

# Lower bounds

- $\mathbb{V}$ is a bounded set of diameter $D$ and all $\ell_t$ are Lipschitz with constant $L$
- Take $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{V}$ such that $\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2 = D$ and set $\boldsymbol{z}_0 = (\boldsymbol{v}_1 - \boldsymbol{v}_2)/\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2$
- Stochastic linear losses $L_t(\boldsymbol{w}) = \varepsilon_t L\, \boldsymbol{w}^\top \boldsymbol{z}_0$ where $\varepsilon_t \in \{-1, 1\}$ are uniform

$$\mathbb{E}\left[\max_{\boldsymbol{u}\in\{\boldsymbol{v}_1,\boldsymbol{v}_2\}} R_T(\boldsymbol{u})\right] = \mathbb{E}\left[\max_{\boldsymbol{u}\in\{\boldsymbol{v}_1,\boldsymbol{v}_2\}} \sum_{t=1}^{T} L_t(\boldsymbol{u})\right] \qquad \text{(since } \mathbb{E}[L_t(\boldsymbol{w})] = 0\text{)}$$

# Lower bounds

- $\mathbb{V}$ is a bounded set of diameter $D$ and all $\ell_t$ are Lipschitz with constant $L$
- Take $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{V}$ such that $\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2 = D$ and set $\boldsymbol{z}_0 = (\boldsymbol{v}_1 - \boldsymbol{v}_2) / \|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2$
- Stochastic linear losses $L_t(\boldsymbol{w}) = \varepsilon_t L\, \boldsymbol{w}^\top \boldsymbol{z}_0$ where $\varepsilon_t \in \{-1, 1\}$ are uniform

$$
\begin{aligned}
\mathbb{E}\left[\max_{\boldsymbol{u}\in\{\boldsymbol{v}_1,\boldsymbol{v}_2\}} R_T(\boldsymbol{u})\right] &= \mathbb{E}\left[\max_{\boldsymbol{u}\in\{\boldsymbol{v}_1,\boldsymbol{v}_2\}} \sum_{t=1}^{T} L_t(\boldsymbol{u})\right] && \text{(since } \mathbb{E}[L_t(\boldsymbol{w})] = 0\text{)} \\
&= \frac{L}{2}\mathbb{E}\left[\left|\sum_{t=1}^{T} \varepsilon_t \boldsymbol{z}_0^\top (\boldsymbol{v}_1 - \boldsymbol{v}_2)\right|\right] && \text{(using } \max\{a,b\} = \tfrac{1}{2}(a+b+|a-b|)\text{)}
\end{aligned}
$$

# Lower bounds

- $\mathbb{V}$ is a bounded set of diameter $D$ and all $\ell_t$ are Lipschitz with constant $L$
- Take $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{V}$ such that $\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2 = D$ and set $\boldsymbol{z}_0 = (\boldsymbol{v}_1 - \boldsymbol{v}_2)/\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2$
- Stochastic linear losses $L_t(\boldsymbol{w}) = \varepsilon_t L\, \boldsymbol{w}^\top \boldsymbol{z}_0$ where $\varepsilon_t \in \{-1, 1\}$ are uniform

$$
\begin{aligned}
\mathbb{E}\left[\max_{\boldsymbol{u}\in\{\boldsymbol{v}_1,\boldsymbol{v}_2\}} R_T(\boldsymbol{u})\right] &= \mathbb{E}\left[\max_{\boldsymbol{u}\in\{\boldsymbol{v}_1,\boldsymbol{v}_2\}} \sum_{t=1}^T L_t(\boldsymbol{u})\right] && \text{(since } \mathbb{E}[L_t(\boldsymbol{w})] = 0\text{)} \\
&= \frac{L}{2}\mathbb{E}\left[\left|\sum_{t=1}^T \varepsilon_t \boldsymbol{z}_0^\top(\boldsymbol{v}_1 - \boldsymbol{v}_2)\right|\right] && \text{(using } \max\{a,b\} = \tfrac{1}{2}(a+b+|a-b|)\text{)} \\
&= \frac{LD}{2}\mathbb{E}\left[\left|\sum_{t=1}^T \varepsilon_t\right|\right] && \text{(because } \boldsymbol{z}_0^\top(\boldsymbol{v}_1 - \boldsymbol{v}_2) = D\text{)}
\end{aligned}
$$

# Lower bounds

- $\mathbb{V}$ is a bounded set of diameter $D$ and all $\ell_t$ are Lipschitz with constant $L$
- Take $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{V}$ such that $\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2 = D$ and set $\boldsymbol{z}_0 = (\boldsymbol{v}_1 - \boldsymbol{v}_2) / \|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2$
- Stochastic linear losses $L_t(\boldsymbol{w}) = \varepsilon_t L\, \boldsymbol{w}^\top \boldsymbol{z}_0$ where $\varepsilon_t \in \{-1, 1\}$ are uniform

$$
\begin{aligned}
\mathbb{E}\left[\max_{\boldsymbol{u}\in\{\boldsymbol{v}_1,\boldsymbol{v}_2\}} R_T(\boldsymbol{u})\right] &= \mathbb{E}\left[\max_{\boldsymbol{u}\in\{\boldsymbol{v}_1,\boldsymbol{v}_2\}} \sum_{t=1}^{T} L_t(\boldsymbol{u})\right] && \text{(since } \mathbb{E}[L_t(\boldsymbol{w})] = 0\text{)} \\
&= \frac{L}{2}\mathbb{E}\left[\left|\sum_{t=1}^{T} \varepsilon_t \boldsymbol{z}_0^\top (\boldsymbol{v}_1 - \boldsymbol{v}_2)\right|\right] && \text{(using } \max\{a,b\} = \tfrac{1}{2}(a+b+|a-b|)\text{)} \\
&= \frac{LD}{2}\mathbb{E}\left[\left|\sum_{t=1}^{T} \varepsilon_t\right|\right] && \text{(because } \boldsymbol{z}_0^\top(\boldsymbol{v}_1 - \boldsymbol{v}_2) = D\text{)} \\
&\geq LD\sqrt{\frac{T}{8}} && \text{(Khintchine inequality)}
\end{aligned}
$$

# Some remarks

- Let $\mathbb{V}$ be the unit Euclidean ball and assume $\ell_t$ is such that $\|\nabla \ell_t\|_\infty = \Omega(1)$

## Some remarks

- Let $\mathbb{V}$ be the unit Euclidean ball and assume $\ell_t$ is such that $\|\nabla \ell_t\|_\infty = \Omega(1)$
- The previous lower bound suggests $R_T(\boldsymbol{u}) = \Omega(\sqrt{dT})$ for $\|\boldsymbol{u}\| \le 1$

# Some remarks

- Let $\mathbb{V}$ be the unit Euclidean ball and assume $\ell_t$ is such that $\|\nabla \ell_t\|_\infty = \Omega(1)$
- The previous lower bound suggests $R_T(\boldsymbol{u}) = \Omega(\sqrt{dT})$ for $\|\boldsymbol{u}\| \leq 1$

- $\mathbb{V}$ is the simplex $\Delta_d$ and $\ell_t$ is linear with coefficients $\|\boldsymbol{\ell}\|_\infty = \Theta(1)$

# Some remarks

- Let $\mathbb{V}$ be the unit Euclidean ball and assume $\ell_t$ is such that $\|\nabla \ell_t\|_\infty = \Omega(1)$
- The previous lower bound suggests $R_T(\boldsymbol{u}) = \Omega(\sqrt{dT})$ for $\|\boldsymbol{u}\| \le 1$

- $\mathbb{V}$ is the simplex $\Delta_d$ and $\ell_t$ is linear with coefficients $\|\boldsymbol{\ell}\|_\infty = \Theta(1)$
- Hedge (exponential weights) achieves $R_T(\boldsymbol{p}) = \mathcal{O}(\sqrt{T \ln d})$ for $\boldsymbol{p} \in \Delta_d$

## Some remarks

- Let $\mathbb{V}$ be the unit Euclidean ball and assume $\ell_t$ is such that $\|\nabla \ell_t\|_\infty = \Omega(1)$
- The previous lower bound suggests $R_T(\boldsymbol{u}) = \Omega(\sqrt{dT})$ for $\|\boldsymbol{u}\| \le 1$

- $\mathbb{V}$ is the simplex $\Delta_d$ and $\ell_t$ is linear with coefficients $\|\boldsymbol{\ell}\|_\infty = \Theta(1)$
- Hedge (exponential weights) achieves $R_T(\boldsymbol{p}) = \mathcal{O}(\sqrt{T \ln d})$ for $\boldsymbol{p} \in \Delta_d$

The geometry of $\mathbb{V}$ matters

# Gradient descent: from online to offline

- Projected gradient descent: $\boldsymbol{w}_{t+1} = \Pi_{\mathbb{V}}\Big(\boldsymbol{w}_t - \eta_t \nabla F(\boldsymbol{w}_t)\Big)$

# Gradient descent: from online to offline

- Projected gradient descent: $\boldsymbol{w}_{t+1} = \Pi_{\mathbb{V}}\Big(\boldsymbol{w}_t - \eta_t \nabla F(\boldsymbol{w}_t)\Big)$
- Projected GD, optimization form: $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \dfrac{1}{2\eta_t} \|\boldsymbol{w} - \boldsymbol{w}_t\|_2^2 + \boldsymbol{w}^\top \nabla F(\boldsymbol{w}_t)$

## Gradient descent: from online to offline

- Projected gradient descent: $\boldsymbol{w}_{t+1} = \Pi_{\mathbb{V}}\Big(\boldsymbol{w}_t - \eta_t \nabla F(\boldsymbol{w}_t)\Big)$
- Projected GD, optimization form: $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} \frac{1}{2\eta_t} \|\boldsymbol{w} - \boldsymbol{w}_t\|_2^2 + \boldsymbol{w}^\top \nabla F(\boldsymbol{w}_t)$
- Projecte online GD (OGD): $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} \frac{1}{2\eta_t} \|\boldsymbol{w} - \boldsymbol{w}_t\|_2^2 + \boldsymbol{w}^\top \nabla \ell_t(\boldsymbol{w}_t)$

# Gradient descent: from online to offline

- Projected gradient descent: $\boldsymbol{w}_{t+1} = \Pi_{\mathbb{V}}\Big(\boldsymbol{w}_t - \eta_t \nabla F(\boldsymbol{w}_t)\Big)$
- Projected GD, optimization form: $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}_t\|_2^2 + \boldsymbol{w}^\top \nabla F(\boldsymbol{w}_t)$
- Projecte online GD (OGD): $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}_t\|_2^2 + \boldsymbol{w}^\top \nabla \ell_t(\boldsymbol{w}_t)$
- Online Mirror Descent (OMD): $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} \frac{1}{2\eta_t} B_\psi(\boldsymbol{w}, \boldsymbol{w}_t) + \boldsymbol{w}^\top \nabla \ell_t(\boldsymbol{w}_t)$

# Gradient descent: from online to offline

- Projected gradient descent: $\boldsymbol{w}_{t+1} = \Pi_{\mathbb{V}}\Big(\boldsymbol{w}_t - \eta_t \nabla F(\boldsymbol{w}_t)\Big)$
- Projected GD, optimization form: $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \frac{1}{2\eta_t} \|\boldsymbol{w} - \boldsymbol{w}_t\|_2^2 + \boldsymbol{w}^\top \nabla F(\boldsymbol{w}_t)$
- Projecte online GD (OGD): $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \frac{1}{2\eta_t} \|\boldsymbol{w} - \boldsymbol{w}_t\|_2^2 + \boldsymbol{w}^\top \nabla \ell_t(\boldsymbol{w}_t)$
- Online Mirror Descent (OMD): $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \frac{1}{2\eta_t} B_\psi(\boldsymbol{w}, \boldsymbol{w}_t) + \boldsymbol{w}^\top \nabla \ell_t(\boldsymbol{w}_t)$

The Bregman divergence $B_\psi$ measures a generalized squared distance between $\boldsymbol{w}, \boldsymbol{w}_t \in \mathbb{V}$

## Bregman divergences

- Parameterized by strictly convex and differentiable mirror map functions $\psi : \mathbb{R}^d \to \mathbb{R}$

## Bregman divergences

- Parameterized by strictly convex and differentiable mirror map functions $\psi : \mathbb{R}^d \to \mathbb{R}$
- $B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \psi(\boldsymbol{u}) - \psi(\boldsymbol{w}) - \nabla\psi(\boldsymbol{w})^\top(\boldsymbol{u} - \boldsymbol{w})$

## Bregman divergences

- Parameterized by strictly convex and differentiable mirror map functions $\psi : \mathbb{R}^d \to \mathbb{R}$
- $B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \psi(\boldsymbol{u}) - \psi(\boldsymbol{w}) - \nabla\psi(\boldsymbol{w})^\top(\boldsymbol{u} - \boldsymbol{w})$
- Error in first-order Taylor expansion of $\psi$ around $\boldsymbol{w}$

# Bregman divergences

- Parameterized by strictly convex and differentiable mirror map functions $\psi : \mathbb{R}^d \to \mathbb{R}$
- $B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \psi(\boldsymbol{u}) - \psi(\boldsymbol{w}) - \nabla\psi(\boldsymbol{w})^\top(\boldsymbol{u} - \boldsymbol{w})$
- Error in first-order Taylor expansion of $\psi$ around $\boldsymbol{w}$
- If $\psi = \frac{1}{2}\|\cdot\|_2^2$, then $\boxed{B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{w}\|_2^2}$

# Bregman divergences

- Parameterized by strictly convex and differentiable mirror map functions $\psi : \mathbb{R}^d \to \mathbb{R}$
- $B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \psi(\boldsymbol{u}) - \psi(\boldsymbol{w}) - \nabla\psi(\boldsymbol{w})^\top(\boldsymbol{u} - \boldsymbol{w})$
- Error in first-order Taylor expansion of $\psi$ around $\boldsymbol{w}$

- If $\psi = \frac{1}{2}\|\cdot\|_2^2$, then $\boxed{B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{w}\|_2^2}$
- OMD becomes online gradient descent (OGD) with Euclidean projection

# Bregman divergences

- Parameterized by strictly convex and differentiable mirror map functions $\psi : \mathbb{R}^d \to \mathbb{R}$
- $B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \psi(\boldsymbol{u}) - \psi(\boldsymbol{w}) - \nabla\psi(\boldsymbol{w})^\top(\boldsymbol{u} - \boldsymbol{w})$
- Error in first-order Taylor expansion of $\psi$ around $\boldsymbol{w}$

- If $\psi = \frac{1}{2}\|\cdot\|_2^2$, then $B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{w}\|_2^2$
- OMD becomes online gradient descent (OGD) with Euclidean projection

- If $\mathbb{V} = \Delta_d$ and $\psi(\boldsymbol{w}) = \sum_i w_i \ln w_i$, then $B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \sum_i u_i \ln \frac{u_i}{w_i}$ (Kullback-Leibler divergence)

# Bregman divergences

- Parameterized by strictly convex and differentiable mirror map functions $\psi : \mathbb{R}^d \to \mathbb{R}$
- $B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \psi(\boldsymbol{u}) - \psi(\boldsymbol{w}) - \nabla\psi(\boldsymbol{w})^\top(\boldsymbol{u} - \boldsymbol{w})$
- Error in first-order Taylor expansion of $\psi$ around $\boldsymbol{w}$

- If $\psi = \frac{1}{2}\|\cdot\|_2^2$, then $B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{w}\|_2^2$
- OMD becomes online gradient descent (OGD) with Euclidean projection

- If $\mathbb{V} = \Delta_d$ and $\psi(\boldsymbol{w}) = \sum_i w_i \ln w_i$, then $B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \sum_i u_i \ln \frac{u_i}{w_i}$ (Kullback-Leibler divergence)
- OMD becomes the Exponentiated Gradient (EG) algorithm (Hedge for general convex losses)

$$w_{t+1,i} \propto \exp\left(-\eta \sum_{s=1}^{t} \nabla \ell_s(\boldsymbol{w}_s)_i\right) \qquad i = 1, \dots, d$$

## Bregman divergences

- Parameterized by strictly convex and differentiable mirror map functions $\psi : \mathbb{R}^d \to \mathbb{R}$
- $B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \psi(\boldsymbol{u}) - \psi(\boldsymbol{w}) - \nabla\psi(\boldsymbol{w})^\top(\boldsymbol{u} - \boldsymbol{w})$
- Error in first-order Taylor expansion of $\psi$ around $\boldsymbol{w}$

- If $\psi = \frac{1}{2}\|\cdot\|_2^2$, then $\boxed{B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{w}\|_2^2}$
- OMD becomes online gradient descent (OGD) with Euclidean projection

- If $\mathbb{V} = \Delta_d$ and $\psi(\boldsymbol{w}) = \sum_i w_i \ln w_i$, then $\boxed{B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \sum_i u_i \ln \frac{u_i}{w_i}}$ (Kullback-Leibler divergence)
- OMD becomes the Exponentiated Gradient (EG) algorithm (Hedge for general convex losses)

$$p_{t+1}(i) \propto \exp\left(-\eta \sum_{s=1}^{t} \ell_s(i)\right) \qquad i = 1, \dots, d$$

# Strongly convex mirror maps

A differentiable $\psi : \mathbb{R}^d \to \mathbb{R}$ is $\mu$-strongly convex on $\mathbb{V}$ with respect to $\|\cdot\|$ if

$$\psi(\boldsymbol{u}) \geq \psi(\boldsymbol{v}) + \nabla\psi(\boldsymbol{v})^\top(\boldsymbol{u} - \boldsymbol{v}) + \frac{\mu}{2}\|\boldsymbol{u} - \boldsymbol{v}\|^2 \qquad \boldsymbol{u}, \boldsymbol{v} \in \mathbb{V}$$

# Strongly convex mirror maps

A differentiable $\psi : \mathbb{R}^d \to \mathbb{R}$ is $\mu$-strongly convex on $\mathbb{V}$ with respect to $\|\cdot\|$ if

$$\psi(\boldsymbol{u}) \geq \psi(\boldsymbol{v}) + \nabla\psi(\boldsymbol{v})^\top(\boldsymbol{u} - \boldsymbol{v}) + \frac{\mu}{2}\|\boldsymbol{u} - \boldsymbol{v}\|^2 \qquad \boldsymbol{u}, \boldsymbol{v} \in \mathbb{V}$$

Properties of strongly convex mirror maps (helpful picture on next slide)

# Strongly convex mirror maps

A differentiable $\psi : \mathbb{R}^d \to \mathbb{R}$ is $\mu$-strongly convex on $\mathbb{V}$ with respect to $\|\cdot\|$ if

$$\psi(\boldsymbol{u}) \geq \psi(\boldsymbol{v}) + \nabla\psi(\boldsymbol{v})^\top(\boldsymbol{u} - \boldsymbol{v}) + \frac{\mu}{2}\|\boldsymbol{u} - \boldsymbol{v}\|^2 \qquad \boldsymbol{u}, \boldsymbol{v} \in \mathbb{V}$$

Properties of strongly convex mirror maps (helpful picture on next slide)

- $B_\psi(\boldsymbol{u}, \boldsymbol{w}) \geq \frac{\mu}{2}\|\boldsymbol{u} - \boldsymbol{w}\|^2$

# Strongly convex mirror maps

A differentiable $\psi : \mathbb{R}^d \to \mathbb{R}$ is $\mu$-strongly convex on $\mathbb{V}$ with respect to $\|\cdot\|$ if

$$\psi(\boldsymbol{u}) \geq \psi(\boldsymbol{v}) + \nabla\psi(\boldsymbol{v})^\top(\boldsymbol{u} - \boldsymbol{v}) + \frac{\mu}{2}\|\boldsymbol{u} - \boldsymbol{v}\|^2 \qquad \boldsymbol{u}, \boldsymbol{v} \in \mathbb{V}$$

Properties of strongly convex mirror maps (helpful picture on next slide)

- $B_\psi(\boldsymbol{u}, \boldsymbol{w}) \geq \frac{\mu}{2}\|\boldsymbol{u} - \boldsymbol{w}\|^2$
- OMD becomes $\boldsymbol{w}_{t+1} = \nabla\psi^\star_{\mathbb{V}}\Big(\nabla\psi_{\mathbb{V}}(\boldsymbol{w}_t) - \eta_t\nabla\ell_t(\boldsymbol{w}_t)\Big)$ ($\psi_{\mathbb{V}}$ is the restriction of $\psi$ to $\mathbb{V}$)

# Strongly convex mirror maps

A differentiable $\psi : \mathbb{R}^d \to \mathbb{R}$ is $\mu$-strongly convex on $\mathbb{V}$ with respect to $\|\cdot\|$ if

$$\psi(\boldsymbol{u}) \geq \psi(\boldsymbol{v}) + \nabla\psi(\boldsymbol{v})^\top(\boldsymbol{u} - \boldsymbol{v}) + \frac{\mu}{2}\|\boldsymbol{u} - \boldsymbol{v}\|^2 \qquad \boldsymbol{u}, \boldsymbol{v} \in \mathbb{V}$$

## Properties of strongly convex mirror maps (helpful picture on next slide)

- $B_\psi(\boldsymbol{u}, \boldsymbol{w}) \geq \frac{\mu}{2}\|\boldsymbol{u} - \boldsymbol{w}\|^2$
- OMD becomes $\boldsymbol{w}_{t+1} = \nabla\psi_{\mathbb{V}}^\star\Big(\nabla\psi_{\mathbb{V}}(\boldsymbol{w}_t) - \eta_t \nabla\ell_t(\boldsymbol{w}_t)\Big)$ ($\psi_{\mathbb{V}}$ is the restriction of $\psi$ to $\mathbb{V}$)
- The function $\psi_{\mathbb{V}}^\star : \mathbb{R}^d \to \mathbb{R}$ is the Fenchel conjugate of $\psi_{\mathbb{V}}$

$$\psi_{\mathbb{V}}^\star(\boldsymbol{\theta}) = \max_{\boldsymbol{w} \in \mathbb{R}^d} \Big(\boldsymbol{w}^\top\boldsymbol{\theta} - \psi_{\mathbb{V}}(\boldsymbol{w})\Big)$$

## Strongly convex mirror maps

A differentiable $\psi : \mathbb{R}^d \to \mathbb{R}$ is $\mu$-strongly convex on $\mathbb{V}$ with respect to $\|\cdot\|$ if

$$\psi(\boldsymbol{u}) \geq \psi(\boldsymbol{v}) + \nabla\psi(\boldsymbol{v})^\top(\boldsymbol{u}-\boldsymbol{v}) + \frac{\mu}{2}\|\boldsymbol{u}-\boldsymbol{v}\|^2 \qquad \boldsymbol{u},\boldsymbol{v} \in \mathbb{V}$$

Properties of strongly convex mirror maps (helpful picture on next slide)

- $\boxed{B_\psi(\boldsymbol{u},\boldsymbol{w}) \geq \frac{\mu}{2}\|\boldsymbol{u}-\boldsymbol{w}\|^2}$
- OMD becomes $\boxed{\boldsymbol{w}_{t+1} = \nabla\psi_{\mathbb{V}}^\star\Big(\nabla\psi_{\mathbb{V}}(\boldsymbol{w}_t) - \eta_t\nabla\ell_t(\boldsymbol{w}_t)\Big)}$ ($\psi_{\mathbb{V}}$ is the restriction of $\psi$ to $\mathbb{V}$)
- The function $\psi_{\mathbb{V}}^\star : \mathbb{R}^d \to \mathbb{R}$ is the Fenchel conjugate of $\psi_{\mathbb{V}}$

$$\psi_{\mathbb{V}}^\star(\boldsymbol{\theta}) = \max_{\boldsymbol{w}\in\mathbb{R}^d}\Big(\boldsymbol{w}^\top\boldsymbol{\theta} - \psi_{\mathbb{V}}(\boldsymbol{w})\Big)$$

- $\psi_{\mathbb{V}}^\star$ is differentiable

## Strongly convex mirror maps

A differentiable $\psi : \mathbb{R}^d \to \mathbb{R}$ is $\mu$-strongly convex on $\mathbb{V}$ with respect to $\|\cdot\|$ if

$$\psi(\boldsymbol{u}) \geq \psi(\boldsymbol{v}) + \nabla\psi(\boldsymbol{v})^\top(\boldsymbol{u}-\boldsymbol{v}) + \frac{\mu}{2}\|\boldsymbol{u}-\boldsymbol{v}\|^2 \qquad \boldsymbol{u},\boldsymbol{v} \in \mathbb{V}$$

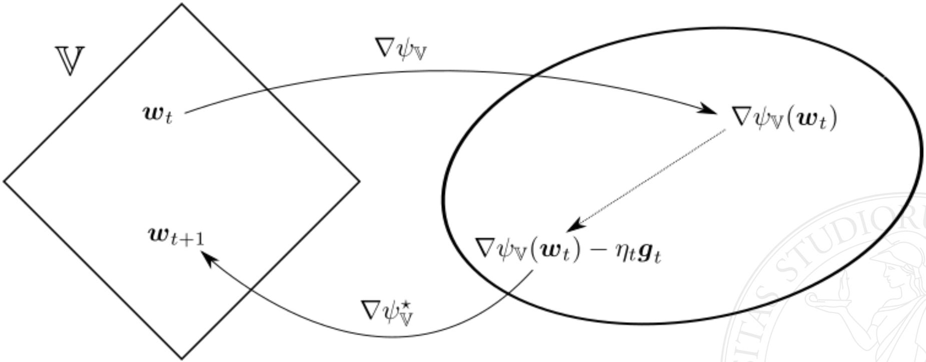
### Properties of strongly convex mirror maps (helpful picture on next slide)

- $B_\psi(\boldsymbol{u},\boldsymbol{w}) \geq \frac{\mu}{2}\|\boldsymbol{u}-\boldsymbol{w}\|^2$
- OMD becomes $\boldsymbol{w}_{t+1} = \nabla\psi^\star_{\mathbb{V}}\Big(\nabla\psi_{\mathbb{V}}(\boldsymbol{w}_t) - \eta_t\nabla\ell_t(\boldsymbol{w}_t)\Big)$ ($\psi_{\mathbb{V}}$ is the restriction of $\psi$ to $\mathbb{V}$)
- The function $\psi^\star_{\mathbb{V}} : \mathbb{R}^d \to \mathbb{R}$ is the Fenchel conjugate of $\psi_{\mathbb{V}}$

$$\psi^\star_{\mathbb{V}}(\boldsymbol{\theta}) = \max_{\boldsymbol{w}\in\mathbb{R}^d}\Big(\boldsymbol{w}^\top\boldsymbol{\theta} - \psi_{\mathbb{V}}(\boldsymbol{w})\Big)$$

- $\psi^\star_{\mathbb{V}}$ is differentiable
- $\nabla\psi^\star_{\mathbb{V}}$ is the functional inverse of $\nabla\psi_{\mathbb{V}}$

# The mirror step




$$\boldsymbol{w}_{t+1} = \nabla\psi^\star_{\mathbb{V}}\Big(\nabla\psi_{\mathbb{V}}(\boldsymbol{w}_t) - \eta_t \underbrace{\nabla\ell_t(\boldsymbol{w}_t)}_{\boldsymbol{g}_t}\Big)$$

# Regret analysis

## Two basic inequalities

$\boxed{\boldsymbol{g}_t = \nabla \ell_t(\boldsymbol{w}_t)}$

- Linearized regret: $\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{u}) \le \boldsymbol{g}_t^\top (\boldsymbol{w}_t - \boldsymbol{u})$

# Regret analysis

## Two basic inequalities

$\boldsymbol{g}_t = \nabla \ell_t(\boldsymbol{w}_t)$

- Linearized regret: $\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{u}) \leq \boldsymbol{g}_t^\top(\boldsymbol{w}_t - \boldsymbol{u})$
- Bregman's progress: $\eta \boldsymbol{g}_t^\top(\boldsymbol{w}_t - \boldsymbol{u}) \leq B_\psi(\boldsymbol{u}, \boldsymbol{w}_t) - B_\psi(\boldsymbol{u}, \boldsymbol{w}_{t+1}) + \frac{\eta^2}{2\mu}\|\boldsymbol{g}_t\|_\star^2$

# Regret analysis

## Two basic inequalities

$\boldsymbol{g}_t = \nabla \ell_t(\boldsymbol{w}_t)$

- Linearized regret: $\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{u}) \leq \boldsymbol{g}_t^\top(\boldsymbol{w}_t - \boldsymbol{u})$
- Bregman's progress: $\eta \boldsymbol{g}_t^\top(\boldsymbol{w}_t - \boldsymbol{u}) \leq B_\psi(\boldsymbol{u}, \boldsymbol{w}_t) - B_\psi(\boldsymbol{u}, \boldsymbol{w}_{t+1}) + \frac{\eta^2}{2\mu}\|\boldsymbol{g}_t\|_\star^2$

$$R_T(\boldsymbol{u}) = \sum_{t=1}^T \left(\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{u})\right)$$

# Regret analysis

## Two basic inequalities

$\boldsymbol{g}_t = \nabla \ell_t(\boldsymbol{w}_t)$

- Linearized regret: $\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{u}) \leq \boldsymbol{g}_t^\top (\boldsymbol{w}_t - \boldsymbol{u})$
- Bregman's progress: $\eta \boldsymbol{g}_t^\top (\boldsymbol{w}_t - \boldsymbol{u}) \leq B_\psi(\boldsymbol{u}, \boldsymbol{w}_t) - B_\psi(\boldsymbol{u}, \boldsymbol{w}_{t+1}) + \frac{\eta^2}{2\mu} \|\boldsymbol{g}_t\|_\star^2$

$$\begin{aligned} R_T(\boldsymbol{u}) &= \sum_{t=1}^{T} \left(\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{u})\right) \\ &\leq \sum_{t=1}^{T} \boldsymbol{g}_t^\top (\boldsymbol{w}_t - \boldsymbol{u}) \qquad \text{(linearized regret)} \end{aligned}$$

# Regret analysis

## Two basic inequalities

$\boldsymbol{g}_t = \nabla \ell_t(\boldsymbol{w}_t)$

- Linearized regret: $\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{u}) \le \boldsymbol{g}_t^\top(\boldsymbol{w}_t - \boldsymbol{u})$
- Bregman's progress: $\eta \boldsymbol{g}_t^\top(\boldsymbol{w}_t - \boldsymbol{u}) \le B_\psi(\boldsymbol{u}, \boldsymbol{w}_t) - B_\psi(\boldsymbol{u}, \boldsymbol{w}_{t+1}) + \frac{\eta^2}{2\mu}\|\boldsymbol{g}_t\|_\star^2$

$$
\begin{aligned}
R_T(\boldsymbol{u}) &= \sum_{t=1}^T \left(\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{u})\right) \\
&\le \sum_{t=1}^T \boldsymbol{g}_t^\top(\boldsymbol{w}_t - \boldsymbol{u}) && \text{(linearized regret)} \\
&\le \sum_{t=1}^T \left(\frac{B_\psi(\boldsymbol{u}, \boldsymbol{w}_t)}{\eta_t} - \frac{B_\psi(\boldsymbol{u}, \boldsymbol{w}_{t+1})}{\eta_t}\right) + \frac{1}{2\mu}\sum_{t=1}^T \eta_t \|\boldsymbol{g}_t\|_\star^2 && \text{(Bregman's progress)}
\end{aligned}
$$

# Regret analysis (cont.)

$$\sum_{t=1}^{T} \left( \frac{B_\psi(\boldsymbol{u}, \boldsymbol{w}_t)}{\eta_t} - \frac{B_\psi(\boldsymbol{u}, \boldsymbol{w}_{t+1})}{\eta_t} \right) + \frac{1}{2\mu} \sum_{t=1}^{T} \eta_t \left\| \boldsymbol{g}_t \right\|_\star^2$$

# Regret analysis (cont.)

$$\sum_{t=1}^{T}\left(\frac{B_\psi(\boldsymbol{u},\boldsymbol{w}_t)}{\eta_t}-\frac{B_\psi(\boldsymbol{u},\boldsymbol{w}_{t+1})}{\eta_t}\right)+\square$$

# Regret analysis (cont.)

$$\sum_{t=1}^{T}\left(\frac{B_\psi(\boldsymbol{u},\boldsymbol{w}_t)}{\eta_t}-\frac{B_\psi(\boldsymbol{u},\boldsymbol{w}_{t+1})}{\eta_t}\right)+\square$$

$$=\frac{B_\psi(\boldsymbol{u},\boldsymbol{w}_1)}{\eta_1}-\frac{B_\psi(\boldsymbol{u},\boldsymbol{w}_{T+1})}{\eta_{T+1}}+\sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}}-\frac{1}{\eta_t}\right)B_\psi(\boldsymbol{u},\boldsymbol{w}_{t+1})+\square \quad \text{(fix telescoping)}$$

# Regret analysis (cont.)

$$
\begin{aligned}
&\sum_{t=1}^{T}\left(\frac{B_\psi(\boldsymbol{u},\boldsymbol{w}_t)}{\eta_t}-\frac{B_\psi(\boldsymbol{u},\boldsymbol{w}_{t+1})}{\eta_t}\right)+\square \\
&\quad=\frac{B_\psi(\boldsymbol{u},\boldsymbol{w}_1)}{\eta_1}-\frac{B_\psi(\boldsymbol{u},\boldsymbol{w}_{T+1})}{\eta_{T+1}}+\sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}}-\frac{1}{\eta_t}\right)B_\psi(\boldsymbol{u},\boldsymbol{w}_{t+1})+\square \qquad \text{(fix telescoping)} \\
&\quad\le\frac{D^2}{\eta_1}+\left(\frac{1}{\eta_T}-\frac{1}{\eta_1}\right)D^2+\square \qquad \text{(where } D^2=\max_{\boldsymbol{u},\boldsymbol{w}\in\mathbb{V}} B_\psi(\boldsymbol{u},\boldsymbol{w})\text{)}
\end{aligned}
$$

# Regret analysis (cont.)

$$
\begin{aligned}
&\sum_{t=1}^{T}\left(\frac{B_\psi(\boldsymbol{u},\boldsymbol{w}_t)}{\eta_t}-\frac{B_\psi(\boldsymbol{u},\boldsymbol{w}_{t+1})}{\eta_t}\right)+\square \\
&\quad=\frac{B_\psi(\boldsymbol{u},\boldsymbol{w}_1)}{\eta_1}-\frac{B_\psi(\boldsymbol{u},\boldsymbol{w}_{T+1})}{\eta_{T+1}}+\sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}}-\frac{1}{\eta_t}\right)B_\psi(\boldsymbol{u},\boldsymbol{w}_{t+1})+\square \qquad \text{(fix telescoping)} \\
&\quad\le\frac{D^2}{\eta_1}+\left(\frac{1}{\eta_T}-\frac{1}{\eta_1}\right)D^2+\square \qquad \text{(where } D^2=\max_{\boldsymbol{u},\boldsymbol{w}\in\mathbb{V}} B_\psi(\boldsymbol{u},\boldsymbol{w})\text{)} \\
&\quad=\frac{D^2}{\eta_T}+\square
\end{aligned}
$$

# The final bound

- We proved $R_T(\boldsymbol{u}) \leq \frac{D^2}{\eta_T} + \frac{1}{2\mu}\sum_{t=1}^{T} \eta_t \left\|\boldsymbol{g}_t\right\|_\star^2$

# The final bound

- We proved $$R_T(\boldsymbol{u}) \le \frac{D^2}{\eta_T} + \frac{1}{2\mu}\sum_{t=1}^{T} \eta_t \|\boldsymbol{g}_t\|_\star^2$$
- Setting $$\eta_t = D\sqrt{\frac{\mu}{\sum_{s=1}^{t} \|\boldsymbol{g}_s\|_\star^2}}$$

# The final bound

- We proved $$R_T(\boldsymbol{u}) \le \frac{D^2}{\eta_T} + \frac{1}{2\mu} \sum_{t=1}^{T} \eta_t \|\boldsymbol{g}_t\|_\star^2$$
- Setting $$\eta_t = D\sqrt{\frac{\mu}{\sum_{s=1}^t \|\boldsymbol{g}_s\|_\star^2}}$$
- We get $$R_T(\boldsymbol{u}) \le 2D\sqrt{\frac{1}{\mu}\sum_{t=1}^{T} \|\boldsymbol{g}_t\|_\star^2}$$

# Matching the mirror map to the geometry of the model space

# Matching the mirror map to the geometry of the model space

OGD

# Matching the mirror map to the geometry of the model space

## OGD

- $\mathbb{V}$ is the closed Euclidean ball of radius $\frac{D}{2}$

# Matching the mirror map to the geometry of the model space

## OGD

- $\mathbb{V}$ is the closed Euclidean ball of radius $\frac{D}{2}$
- $\psi = \dfrac{1}{2} \|\cdot\|_2^2$ is $1$-strongly convex with respect to $\|\cdot\|_2$

# Matching the mirror map to the geometry of the model space

## OGD

- $\mathbb{V}$ is the closed Euclidean ball of radius $\frac{D}{2}$
- $\psi = \dfrac{1}{2} \|\cdot\|_2^2$ is $1$-strongly convex with respect to $\|\cdot\|_2$
- Bregman divergence: $B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \dfrac{1}{2} \|\boldsymbol{u} - \boldsymbol{w}\|_2^2$

# Matching the mirror map to the geometry of the model space

## OGD

- $\mathbb{V}$ is the closed Euclidean ball of radius $\frac{D}{2}$
- $\psi = \dfrac{1}{2} \|\cdot\|_2^2$ is $1$-strongly convex with respect to $\|\cdot\|_2$
- Bregman divergence: $B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \dfrac{1}{2} \|\boldsymbol{u} - \boldsymbol{w}\|_2^2$
- Assume $\|\boldsymbol{g}_t\|_\star^2 = \|\boldsymbol{g}_t\|_2^2 = \mathcal{O}(d)$

# Matching the mirror map to the geometry of the model space

## OGD

- $\mathbb{V}$ is the closed Euclidean ball of radius $\frac{D}{2}$
- $\psi = \frac{1}{2} \|\cdot\|_2^2$ is $1$-strongly convex with respect to $\|\cdot\|_2$
- Bregman divergence: $B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \frac{1}{2} \|\boldsymbol{u} - \boldsymbol{w}\|_2^2$
- Assume $\|\boldsymbol{g}_t\|_\star^2 = \|\boldsymbol{g}_t\|_2^2 = \mathcal{O}(d)$
- $R_T = \mathcal{O}\big(D\sqrt{dT}\big)$

# Matching the mirror map to the geometry of the model space

EG (with constant stepsize $\eta = \sqrt{(\ln d)/T}$)

# Matching the mirror map to the geometry of the model space

## EG (with constant stepsize $\eta = \sqrt{(\ln d)/T}$)

- $\mathbb{V}$ is the probability simplex

# Matching the mirror map to the geometry of the model space

EG (with constant stepsize $\eta = \sqrt{(\ln d)/T}$)

- $\mathbb{V}$ is the probability simplex
- $\psi(\boldsymbol{p}) = \sum_i p_i \ln p_i$ is $1$-strongly convex with respect to $\|\cdot\|_1$

# Matching the mirror map to the geometry of the model space

EG (with constant stepsize $\eta = \sqrt{(\ln d)/T}$)

- $\mathbb{V}$ is the probability simplex
- $\psi(\boldsymbol{p}) = \sum_i p_i \ln p_i$ is $1$-strongly convex with respect to $\|\cdot\|_1$
- Bregman divergence: $B_\psi(\boldsymbol{q}, \boldsymbol{p}) = \sum_{i=1}^d q_i \ln \frac{q_i}{p_i}$

# Matching the mirror map to the geometry of the model space

EG (with constant stepsize $\eta = \sqrt{(\ln d)/T}$)

- $\mathbb{V}$ is the probability simplex
- $\psi(\boldsymbol{p}) = \sum_i p_i \ln p_i$ is $1$-strongly convex with respect to $\|\cdot\|_1$
- Bregman divergence: $B_\psi(\boldsymbol{q}, \boldsymbol{p}) = \sum_{i=1}^d q_i \ln \frac{q_i}{p_i}$
- Problem: $D^2 = \max_{\boldsymbol{p},\boldsymbol{q} \in \Delta_d} B_\psi(\boldsymbol{q}, \boldsymbol{p}) = \infty$

# Matching the mirror map to the geometry of the model space

EG (with constant stepsize $\eta = \sqrt{(\ln d)/T}$)

- $\mathbb{V}$ is the probability simplex
- $\psi(\boldsymbol{p}) = \sum_i p_i \ln p_i$ is $1$-strongly convex with respect to $\|\cdot\|_1$
- Bregman divergence: $B_\psi(\boldsymbol{q}, \boldsymbol{p}) = \sum_{i=1}^d q_i \ln \frac{q_i}{p_i}$
- Problem: $D^2 = \max_{\boldsymbol{p},\boldsymbol{q} \in \Delta_d} B_\psi(\boldsymbol{q}, \boldsymbol{p}) = \infty$
- OMD analysis for constant learning rate: $R_T(\boldsymbol{q}) \leq \dfrac{B_\psi(\boldsymbol{q}, \boldsymbol{p}_1)}{\eta} + \dfrac{\eta}{2} \sum_{t=1}^T \|\boldsymbol{g}_t\|_\star^2$

# Matching the mirror map to the geometry of the model space

## EG (with constant stepsize $\eta = \sqrt{(\ln d)/T}$)

- $\mathbb{V}$ is the probability simplex
- $\psi(\boldsymbol{p}) = \sum_i p_i \ln p_i$ is $1$-strongly convex with respect to $\|\cdot\|_1$
- Bregman divergence: $B_\psi(\boldsymbol{q}, \boldsymbol{p}) = \sum_{i=1}^d q_i \ln \frac{q_i}{p_i}$
- Problem: $D^2 = \max_{\boldsymbol{p},\boldsymbol{q} \in \Delta_d} B_\psi(\boldsymbol{q}, \boldsymbol{p}) = \infty$
- OMD analysis for constant learning rate: $R_T(\boldsymbol{q}) \le \frac{B_\psi(\boldsymbol{q}, \boldsymbol{p}_1)}{\eta} + \frac{\eta}{2} \sum_{t=1}^T \|\boldsymbol{g}_t\|_\star^2$
- Choosing $\boldsymbol{p}_1 = \left(\frac{1}{d}, \dots, \frac{1}{d}\right)$ we get $B_\psi(\boldsymbol{q}, \boldsymbol{p}_1) \le \ln d$

# Matching the mirror map to the geometry of the model space

## EG (with constant stepsize $\eta = \sqrt{(\ln d)/T}$)

- $\mathbb{V}$ is the probability simplex
- $\psi(\boldsymbol{p}) = \sum_i p_i \ln p_i$ is $1$-strongly convex with respect to $\|\cdot\|_1$
- Bregman divergence: $B_\psi(\boldsymbol{q}, \boldsymbol{p}) = \sum_{i=1}^d q_i \ln \frac{q_i}{p_i}$
- Problem: $D^2 = \max_{\boldsymbol{p},\boldsymbol{q} \in \Delta_d} B_\psi(\boldsymbol{q}, \boldsymbol{p}) = \infty$
- OMD analysis for constant learning rate: $R_T(\boldsymbol{q}) \le \dfrac{B_\psi(\boldsymbol{q}, \boldsymbol{p}_1)}{\eta} + \dfrac{\eta}{2} \sum_{t=1}^T \|\boldsymbol{g}_t\|_\star^2$
- Choosing $\boldsymbol{p}_1 = \left(\frac{1}{d}, \dots, \frac{1}{d}\right)$ we get $B_\psi(\boldsymbol{q}, \boldsymbol{p}_1) \le \ln d$
- Assume $\|\boldsymbol{g}_t\|_\star^2 = \|\boldsymbol{g}_t\|_\infty^2 = \mathcal{O}(1)$

# Matching the mirror map to the geometry of the model space

## EG (with constant stepsize $\eta = \sqrt{(\ln d)/T}$)

- $\mathbb{V}$ is the probability simplex
- $\psi(\boldsymbol{p}) = \sum_i p_i \ln p_i$ is $1$-strongly convex with respect to $\|\cdot\|_1$
- Bregman divergence: $B_\psi(\boldsymbol{q}, \boldsymbol{p}) = \sum_{i=1}^d q_i \ln \frac{q_i}{p_i}$
- Problem: $D^2 = \max_{\boldsymbol{p},\boldsymbol{q} \in \Delta_d} B_\psi(\boldsymbol{q}, \boldsymbol{p}) = \infty$
- OMD analysis for constant learning rate: $R_T(\boldsymbol{q}) \le \dfrac{B_\psi(\boldsymbol{q}, \boldsymbol{p}_1)}{\eta} + \dfrac{\eta}{2} \displaystyle\sum_{t=1}^T \|\boldsymbol{g}_t\|_\star^2$
- Choosing $\boldsymbol{p}_1 = \left(\frac{1}{d}, \dots, \frac{1}{d}\right)$ we get $B_\psi(\boldsymbol{q}, \boldsymbol{p}_1) \le \ln d$
- Assume $\|\boldsymbol{g}_t\|_\star^2 = \|\boldsymbol{g}_t\|_\infty^2 = \mathcal{O}(1)$
- $R_T = \mathcal{O}(\sqrt{T \ln d})$

# Some remarks

- We can interpolate between OGD and EG using a $p$-norm as a mirror map:
  $\psi(\boldsymbol{w}) = \frac{1}{2}\left(\sum_{i=1}^{d} |w_i|^p\right)^{2/p}$ for $1 < p \leq 2$

# Some remarks

- We can interpolate between OGD and EG using a $p$-norm as a mirror map:
  $\psi(\boldsymbol{w}) = \frac{1}{2}\left(\sum_{i=1}^{d} |w_i|^p\right)^{2/p}$ for $1 < p \le 2$
- Choosing $p = \frac{2\ln d}{2\ln d - 1}$ gives bound similar to EG without the tuning problem

# AdaGrad (diagonal version)

# AdaGrad (diagonal version)

- Independence w.r.t. rescaling of the coordinates

# AdaGrad (diagonal version)

- Independence w.r.t. rescaling of the coordinates
- Useful in neural network training where range of gradient components varies across layers

# AdaGrad (diagonal version)

- Independence w.r.t. rescaling of the coordinates
- Useful in neural network training where range of gradient components varies across layers

- $\mathbb{V}$ is the hyperrectangle $[a_1, b_1] \times \cdots \times [a_d, b_d] \in \mathbb{R}^d$

# AdaGrad (diagonal version)

- Independence w.r.t. rescaling of the coordinates
- Useful in neural network training where range of gradient components varies across layers

- $\mathbb{V}$ is the hyperrectangle $[a_1, b_1] \times \cdots \times [a_d, b_d] \in \mathbb{R}^d$
- Run OMD with Euclidean mirror map independently on each coordinate:

$$w_{t+1,i} = \max\left\{ \min\{w_{t,i} - \eta_{t,i} g_{t,i}\}, a_i \right\} \qquad i = 1, \dots, d$$

# AdaGrad (diagonal version)

- Independence w.r.t. rescaling of the coordinates
- Useful in neural network training where range of gradient components varies across layers

- $\mathbb{V}$ is the hyperrectangle $[a_1, b_1] \times \cdots \times [a_d, b_d] \in \mathbb{R}^d$
- Run OMD with Euclidean mirror map independently on each coordinate:

$$w_{t+1,i} = \max\left\{ \min\{w_{t,i} - \eta_{t,i} g_{t,i}\}, a_i \right\} \qquad i = 1, \dots, d$$

- With learning rate

$$\eta_{t,i} = \frac{b_i - a_i}{\sqrt{2 \sum_{s=1}^{t} g_{s,i}^2}} \qquad i = 1, \dots, d$$

# AdaGrad analysis

## AdaGrad analysis

By applying OMD analysis on each coordinate $R_T \le \sum_{i=1}^{d} (b_i - a_i) \sqrt{2 \sum_{t=1}^{T} g_{t,i}^2}$

# AdaGrad analysis

By applying OMD analysis on each coordinate $$R_T \leq \sum_{i=1}^{d} (b_i - a_i) \sqrt{2 \sum_{t=1}^{T} g_{t,i}^2}$$

## Comparing with OGD bound

# AdaGrad analysis

By applying OMD analysis on each coordinate $\quad R_T \leq \sum_{i=1}^{d} (b_i - a_i) \sqrt{2 \sum_{t=1}^{T} g_{t,i}^2}$

## Comparing with OGD bound

- For simplicity, take $b_i - a_i = 1$ for $i = 1, \ldots, d$

# AdaGrad analysis

By applying OMD analysis on each coordinate $\quad R_T \leq \sum_{i=1}^{d} (b_i - a_i) \sqrt{2 \sum_{t=1}^{T} g_{t,i}^2}$

## Comparing with OGD bound

- For simplicity, take $b_i - a_i = 1$ for $i = 1, \ldots, d$
- The diameter of $\mathbb{V}$ is then $D = \sqrt{d}$

# AdaGrad analysis

By applying OMD analysis on each coordinate $$R_T \leq \sum_{i=1}^{d} (b_i - a_i) \sqrt{2 \sum_{t=1}^{T} g_{t,i}^2}$$

## Comparing with OGD bound

- For simplicity, take $b_i - a_i = 1$ for $i = 1, \ldots, d$
- The diameter of $\mathbb{V}$ is then $D = \sqrt{d}$
- OGD update: $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \boldsymbol{g}_t$ followed by projection onto $\mathbb{V}$

# AdaGrad analysis

By applying OMD analysis on each coordinate $$R_T \leq \sum_{i=1}^{d}(b_i - a_i)\sqrt{2\sum_{t=1}^{T} g_{t,i}^2}$$

## Comparing with OGD bound

- For simplicity, take $b_i - a_i = 1$ for $i = 1, \ldots, d$
- The diameter of $\mathbb{V}$ is then $D = \sqrt{d}$
- OGD update: $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \boldsymbol{g}_t$ followed by projection onto $\mathbb{V}$
- OGD learning rate: $$\eta_t = \sqrt{\frac{d}{\sum_{s=1}^{t} \|\boldsymbol{g}_s\|^2}}$$

# AdaGrad analysis

By applying OMD analysis on each coordinate $$R_T \le \sum_{i=1}^{d}(b_i - a_i)\sqrt{2\sum_{t=1}^{T} g_{t,i}^2}$$

## Comparing with OGD bound

- For simplicity, take $b_i - a_i = 1$ for $i = 1, \ldots, d$
- The diameter of $\mathbb{V}$ is then $D = \sqrt{d}$
- OGD update: $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \boldsymbol{g}_t$ followed by projection onto $\mathbb{V}$
- OGD learning rate: $$\eta_t = \sqrt{\frac{d}{\sum_{s=1}^{t} \|\boldsymbol{g}_s\|^2}}$$
- By Jensen's inequality $$\underbrace{\sum_{i=1}^{d}\sqrt{\sum_{t=1}^{T} g_{t,i}^2}}_{\text{AdaGrad}} \le \underbrace{\sqrt{d}\sqrt{\sum_{t=1}^{T}\|\boldsymbol{g}_t\|_2^2}}_{\text{OGD}}$$

# Exploiting curvature of the losses

# Exploiting curvature of the losses

- Convex losses: OGD with $\eta_t \approx \frac{1}{\sqrt{t}}$ achieves $R_T = \mathcal{O}(\sqrt{dT})$

# Exploiting curvature of the losses

- Convex losses: OGD with $\eta_t \approx \frac{1}{\sqrt{t}}$ achieves $R_T = \mathcal{O}(\sqrt{dT})$
- Strongly convex losses: OGD with $\eta_t \approx \frac{1}{t}$ achieves $R_T = \mathcal{O}(d \ln T)$ (unconstrained!)

# Exploiting curvature of the losses

- Convex losses: OGD with $\eta_t \approx \frac{1}{\sqrt{t}}$ achieves $R_T = \mathcal{O}(\sqrt{dT})$
- Strongly convex losses: OGD with $\eta_t \approx \frac{1}{t}$ achieves $R_T = \mathcal{O}(d \ln T)$ (unconstrained!)

## Strong convexity in the direction of the gradient

$$\ell_t(\boldsymbol{u}) \geq \ell_t(\boldsymbol{w}) + \boldsymbol{g}^\top(\boldsymbol{u} - \boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{u} - \boldsymbol{w}\|^2_{\boldsymbol{g}\boldsymbol{g}^\top} \qquad \boldsymbol{u}, \boldsymbol{w} \in \mathbb{V}$$

where $\boldsymbol{g} = \nabla \ell_t(\boldsymbol{w})$ and $\|\boldsymbol{w}\|^2_M = \boldsymbol{w}^\top M \boldsymbol{w}$

# Exploiting curvature of the losses

- Convex losses: OGD with $\eta_t \approx \frac{1}{\sqrt{t}}$ achieves $R_T = \mathcal{O}(\sqrt{dT})$
- Strongly convex losses: OGD with $\eta_t \approx \frac{1}{t}$ achieves $R_T = \mathcal{O}(d \ln T)$ (unconstrained!)

## Strong convexity in the direction of the gradient

$$\ell_t(\boldsymbol{u}) \geq \ell_t(\boldsymbol{w}) + \boldsymbol{g}^\top(\boldsymbol{u} - \boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{u} - \boldsymbol{w}\|^2_{\boldsymbol{g}\boldsymbol{g}^\top} \qquad \boldsymbol{u}, \boldsymbol{w} \in \mathbb{V}$$

where $\boldsymbol{g} = \nabla \ell_t(\boldsymbol{w})$ and $\|\boldsymbol{w}\|^2_M = \boldsymbol{w}^\top M \boldsymbol{w}$

## Some losses satisfying the condition

- Square loss $\ell(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{w}^\top \boldsymbol{x} - y)^2$ for bounded $|\boldsymbol{w}^\top \boldsymbol{x}|, |y|$

# Exploiting curvature of the losses

- Convex losses: OGD with $\eta_t \approx \frac{1}{\sqrt{t}}$ achieves $R_T = \mathcal{O}(\sqrt{dT})$
- Strongly convex losses: OGD with $\eta_t \approx \frac{1}{t}$ achieves $R_T = \mathcal{O}(d \ln T)$ (unconstrained!)

## Strong convexity in the direction of the gradient

$$\ell_t(\boldsymbol{u}) \geq \ell_t(\boldsymbol{w}) + \boldsymbol{g}^\top(\boldsymbol{u} - \boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{u} - \boldsymbol{w}\|^2_{\boldsymbol{g}\boldsymbol{g}^\top} \qquad \boldsymbol{u}, \boldsymbol{w} \in \mathbb{V}$$

where $\boldsymbol{g} = \nabla \ell_t(\boldsymbol{w})$ and $\|\boldsymbol{w}\|^2_M = \boldsymbol{w}^\top M \boldsymbol{w}$

## Some losses satisfying the condition

- Square loss $\ell(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{w}^\top \boldsymbol{x} - y)^2$ for bounded $|\boldsymbol{w}^\top \boldsymbol{x}|, |y|$
- Logistic loss $\ell_t(\boldsymbol{w}) = \ln\left(1 + \exp(-\boldsymbol{w}^\top \boldsymbol{x}_t)\right)$ for bounded $\|\boldsymbol{w}\|$

# Follow the Regularized Leader

- FTRL is not formulated as gradient descent, but as regularized error minimization

# Follow the Regularized Leader

- FTRL is not formulated as gradient descent, but as regularized error minimization
- OMD learning rates $\eta_t$ are replaced by time-dependent regularizers (mirror maps) $\psi_t$

# Follow the Regularized Leader

- FTRL is not formulated as gradient descent, but as regularized error minimization
- OMD learning rates $\eta_t$ are replaced by time-dependent regularizers (mirror maps) $\psi_t$
- $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \psi_{t+1}(\boldsymbol{w}) + \sum_{s=1}^{t} \ell_s(\boldsymbol{w})$

# Follow the Regularized Leader

- FTRL is not formulated as gradient descent, but as regularized error minimization
- OMD learning rates $\eta_t$ are replaced by time-dependent regularizers (mirror maps) $\psi_t$
- $\boldsymbol{w}_{t+1} = \underset{\boldsymbol{w} \in \mathbb{V}}{\operatorname{argmin}} \, \psi_{t+1}(\boldsymbol{w}) + \sum_{s=1}^{t} \nabla \ell_s(\boldsymbol{w}_s)^\top \boldsymbol{w}$

# Follow the Regularized Leader

- FTRL is not formulated as gradient descent, but as regularized error minimization
- OMD learning rates $\eta_t$ are replaced by time-dependent regularizers (mirror maps) $\psi_t$
- $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \psi_{t+1}(\boldsymbol{w}) + \sum_{s=1}^{t} \nabla \ell_s(\boldsymbol{w}_s)^\top \boldsymbol{w}$
- If $\psi_{\mathbb{V},t}$ are all strongly convex, then $\boldsymbol{w}_{t+1} = \nabla \psi^{\star}_{\mathbb{V},t+1}\left(-\sum_{s=1}^{t} \nabla \ell_s(\boldsymbol{w}_s)\right)$

# Follow the Regularized Leader

- FTRL is not formulated as gradient descent, but as regularized error minimization
- OMD learning rates $\eta_t$ are replaced by time-dependent regularizers (mirror maps) $\psi_t$
- $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} \psi_{t+1}(\boldsymbol{w}) + \sum_{s=1}^{t} \nabla\ell_s(\boldsymbol{w}_s)^\top \boldsymbol{w}$
- If $\psi_{\mathbb{V},t}$ are all strongly convex, then $\boldsymbol{w}_{t+1} = \nabla\psi^\star_{\mathbb{V},t+1}\left(-\sum_{s=1}^{t}\nabla\ell_s(\boldsymbol{w}_s)\right)$
- Recall OMD: $\boldsymbol{w}_{t+1} = \nabla\psi^\star_{\mathbb{V}}(\boldsymbol{\theta}'_{t+1})$ where $\boldsymbol{\theta}'_{t+1} = \nabla\psi_{\mathbb{V}}(\boldsymbol{w}_t) - \eta_t\nabla\ell_t(\boldsymbol{w}_t)$

# Follow the Regularized Leader

- FTRL is not formulated as gradient descent, but as regularized error minimization
- OMD learning rates $\eta_t$ are replaced by time-dependent regularizers (mirror maps) $\psi_t$
- $\boldsymbol{w}_{t+1} = \underset{\boldsymbol{w} \in \mathbb{V}}{\operatorname{argmin}} \, \psi_{t+1}(\boldsymbol{w}) + \sum_{s=1}^{t} \nabla \ell_s(\boldsymbol{w}_s)^\top \boldsymbol{w}$
- If $\psi_{\mathbb{V},t}$ are all strongly convex, then $\boldsymbol{w}_{t+1} = \nabla \psi^\star_{\mathbb{V},t+1}\left(-\sum_{s=1}^{t} \nabla \ell_s(\boldsymbol{w}_s)\right)$
- Recall OMD: $\boldsymbol{w}_{t+1} = \nabla \psi^\star_{\mathbb{V}}(\boldsymbol{\theta}'_{t+1})$ where $\boldsymbol{\theta}'_{t+1} = \nabla \psi_{\mathbb{V}}(\boldsymbol{w}_t) - \eta_t \nabla \ell_t(\boldsymbol{w}_t)$
- OMD throws away information by projecting after each update

# Some differences between OMD and FTRL

- FTRL keeps a state variable $\boldsymbol{\theta}_t$ in the dual space of gradients

# Some differences between OMD and FTRL

- FTRL keeps a state variable $\boldsymbol{\theta}_t$ in the dual space of gradients
- This is then mapped to the primal space of iterates everytime a prediction is needed

# Some differences between OMD and FTRL

- FTRL keeps a state variable $\boldsymbol{\theta}_t$ in the dual space of gradients
- This is then mapped to the primal space of iterates everytime a prediction is needed
- OMD keeps its state $\boldsymbol{w}_t$ in the primal space of iterates

# Some differences between OMD and FTRL

- FTRL keeps a state variable $\boldsymbol{\theta}_t$ in the dual space of gradients
- This is then mapped to the primal space of iterates everytime a prediction is needed
- OMD keeps its state $\boldsymbol{w}_t$ in the primal space of iterates
- This is then mapped to the dual space of gradients everytime an update must be computed

# Some differences between OMD and FTRL

- FTRL keeps a state variable $\boldsymbol{\theta}_t$ in the dual space of gradients
- This is then mapped to the primal space of iterates everytime a prediction is needed
- OMD keeps its state $\boldsymbol{w}_t$ in the primal space of iterates
- This is then mapped to the dual space of gradients everytime an update must be computed
- OMD and FTRL have similar regret bounds and become identical in certain cases

# Some differences between OMD and FTRL

- FTRL keeps a state variable $\boldsymbol{\theta}_t$ in the dual space of gradients
- This is then mapped to the primal space of iterates everytime a prediction is needed
- OMD keeps its state $\boldsymbol{w}_t$ in the primal space of iterates
- This is then mapped to the dual space of gradients everytime an update must be computed
- OMD and FTRL have similar regret bounds and become identical in certain cases
- Time-dependent regularizers are generally more flexible than time-dependent learning rates

## Online Newton Step

Choose the model minimizing a second-order approximation of the true loss:

$$\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} \sum_{s=1}^{t} \widehat{\ell}_s(\boldsymbol{w}) \qquad \text{(Follow-the-Leader approach)}$$

$$\widehat{\ell}_t(\boldsymbol{w}) = \ell_t(\boldsymbol{w}_t) + \boldsymbol{g}_t^\top(\boldsymbol{w}-\boldsymbol{w}_t) + \frac{\lambda}{2}\left\|\boldsymbol{w}-\boldsymbol{w}_t\right\|^2_{\boldsymbol{g}_t\boldsymbol{g}_t^\top} \qquad \boldsymbol{g}_t = \nabla\ell_t(\boldsymbol{w}_t)$$

# Online Newton Step

Choose the model minimizing a second-order approximation of the true loss:

$$\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \sum_{s=1}^{t} \widehat{\ell}_s(\boldsymbol{w}) \qquad \text{(Follow-the-Leader approach)}$$

$$\widehat{\ell}_t(\boldsymbol{w}) = \ell_t(\boldsymbol{w}_t) + \boldsymbol{g}_t^\top (\boldsymbol{w} - \boldsymbol{w}_t) + \frac{\lambda}{2} \|\boldsymbol{w} - \boldsymbol{w}_t\|^2_{\boldsymbol{g}_t \boldsymbol{g}_t^\top} \qquad \boldsymbol{g}_t = \nabla \ell_t(\boldsymbol{w}_t)$$

Regret analysis:

# Online Newton Step

Choose the model minimizing a second-order approximation of the true loss:

$$\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} \sum_{s=1}^{t} \widehat{\ell}_s(\boldsymbol{w}) \qquad \text{(Follow-the-Leader approach)}$$

$$\widehat{\ell}_t(\boldsymbol{w}) = \ell_t(\boldsymbol{w}_t) + \boldsymbol{g}_t^\top(\boldsymbol{w} - \boldsymbol{w}_t) + \frac{\lambda}{2}\|\boldsymbol{w} - \boldsymbol{w}_t\|^2_{\boldsymbol{g}_t\boldsymbol{g}_t^\top} \qquad \boldsymbol{g}_t = \nabla\ell_t(\boldsymbol{w}_t)$$

Regret analysis:

- $\widehat{\ell}_t(\boldsymbol{w}_t) = \ell_t(\boldsymbol{w}_t)$

# Online Newton Step

Choose the model minimizing a second-order approximation of the true loss:

$$\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} \sum_{s=1}^{t} \widehat{\ell}_s(\boldsymbol{w}) \qquad \text{(Follow-the-Leader approach)}$$

$$\widehat{\ell}_t(\boldsymbol{w}) = \ell_t(\boldsymbol{w}_t) + \boldsymbol{g}_t^\top(\boldsymbol{w}-\boldsymbol{w}_t) + \frac{\lambda}{2}\left\|\boldsymbol{w}-\boldsymbol{w}_t\right\|^2_{\boldsymbol{g}_t\boldsymbol{g}_t^\top} \qquad \boldsymbol{g}_t = \nabla\ell_t(\boldsymbol{w}_t)$$

## Regret analysis:

- $\widehat{\ell}_t(\boldsymbol{w}_t) = \ell_t(\boldsymbol{w}_t)$
- $\widehat{\ell}_t(\boldsymbol{u}) \le \ell_t(\boldsymbol{u})$ for all $\boldsymbol{u} \in \mathbb{V}$

# Online Newton Step

Choose the model minimizing a second-order approximation of the true loss:

$$\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \sum_{s=1}^{t} \widehat{\ell}_s(\boldsymbol{w}) \qquad \text{(Follow-the-Leader approach)}$$

$$\widehat{\ell}_t(\boldsymbol{w}) = \ell_t(\boldsymbol{w}_t) + \boldsymbol{g}_t^\top(\boldsymbol{w} - \boldsymbol{w}_t) + \frac{\lambda}{2}\left\|\boldsymbol{w} - \boldsymbol{w}_t\right\|^2_{\boldsymbol{g}_t\boldsymbol{g}_t^\top} \qquad \boldsymbol{g}_t = \nabla \ell_t(\boldsymbol{w}_t)$$

Regret analysis:

- $\widehat{\ell}_t(\boldsymbol{w}_t) = \ell_t(\boldsymbol{w}_t)$
- $\widehat{\ell}_t(\boldsymbol{u}) \le \ell_t(\boldsymbol{u})$ for all $\boldsymbol{u} \in \mathbb{V}$
- Regret bound: $\quad R_T(\boldsymbol{u}) \le \sum_{t=1}^{T} \widehat{\ell}_t(\boldsymbol{w}_t) - \sum_{t=1}^{T} \widehat{\ell}_t(\boldsymbol{u}) = \mathcal{O}(d \ln T)$

# Online Newton Step

Choose the model minimizing a second-order approximation of the true loss:

$$\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} \sum_{s=1}^{t} \widehat{\ell}_s(\boldsymbol{w}) \qquad \text{(Follow-the-Leader approach)}$$

$$\widehat{\ell}_t(\boldsymbol{w}) = \ell_t(\boldsymbol{w}_t) + \boldsymbol{g}_t^\top(\boldsymbol{w}-\boldsymbol{w}_t) + \frac{\lambda}{2}\|\boldsymbol{w}-\boldsymbol{w}_t\|^2_{\boldsymbol{g}_t\boldsymbol{g}_t^\top} \qquad \boldsymbol{g}_t = \nabla\ell_t(\boldsymbol{w}_t)$$

Regret analysis:

- $\widehat{\ell}_t(\boldsymbol{w}_t) = \ell_t(\boldsymbol{w}_t)$
- $\widehat{\ell}_t(\boldsymbol{u}) \le \ell_t(\boldsymbol{u})$ for all $\boldsymbol{u} \in \mathbb{V}$
- Regret bound: $R_T(\boldsymbol{u}) \le \sum_{t=1}^{T} \widehat{\ell}_t(\boldsymbol{w}_t) - \sum_{t=1}^{T} \widehat{\ell}_t(\boldsymbol{u}) = \mathcal{O}(d\ln T)$
- $\mathcal{O}(d\ln T)$ matches the bound for strongly convex losses

# Unconstrained online convex optimization

- Model space is unconstrained: $\mathbb{V} = \mathbb{R}^d$

# Unconstrained online convex optimization

- Model space is unconstrained: $\mathbb{V} = \mathbb{R}^d$
- Run OGD with fixed learning rate $\eta = \alpha/\sqrt{T}$ for $\alpha > 0$

# Unconstrained online convex optimization

- Model space is unconstrained: $\mathbb{V} = \mathbb{R}^d$
- Run OGD with fixed learning rate $\eta = \alpha/\sqrt{T}$ for $\alpha > 0$
- $R_T(\boldsymbol{u}) \leq \frac{1}{2}\left(\frac{\|\boldsymbol{u}\|_2^2}{\alpha} + \alpha\right)\sqrt{T} \qquad \forall \boldsymbol{u} \in \mathbb{R}^d$

# Unconstrained online convex optimization

- Model space is unconstrained: $\mathbb{V} = \mathbb{R}^d$
- Run OGD with fixed learning rate $\eta = \alpha/\sqrt{T}$ for $\alpha > 0$
- $R_T(\boldsymbol{u}) \le \frac{1}{2}\left(\frac{\|\boldsymbol{u}\|_2^2}{\alpha} + \alpha\right)\sqrt{T} \qquad \forall \boldsymbol{u} \in \mathbb{R}^d$
- $R_T(\boldsymbol{u}) \le \|\boldsymbol{u}\|_2 \sqrt{T}$ for $\alpha = \|\boldsymbol{u}\|_2$

# Unconstrained online convex optimization

- Model space is unconstrained: $\mathbb{V} = \mathbb{R}^d$
- Run OGD with fixed learning rate $\eta = \alpha/\sqrt{T}$ for $\alpha > 0$
- $R_T(\boldsymbol{u}) \leq \frac{1}{2}\left(\frac{\|\boldsymbol{u}\|_2^2}{\alpha} + \alpha\right)\sqrt{T} \qquad \forall \boldsymbol{u} \in \mathbb{R}^d$
- $R_T(\boldsymbol{u}) \leq \|\boldsymbol{u}\|_2 \sqrt{T}$ for $\alpha = \|\boldsymbol{u}\|_2$
- Equivalent to running OGD with projection in the Euclidean ball of radius $\|\boldsymbol{u}\|_2$

# Unconstrained online convex optimization

- Model space is unconstrained: $\mathbb{V} = \mathbb{R}^d$
- Run OGD with fixed learning rate $\eta = \alpha/\sqrt{T}$ for $\alpha > 0$
- $R_T(\boldsymbol{u}) \le \frac{1}{2}\left(\frac{\|\boldsymbol{u}\|_2^2}{\alpha} + \alpha\right)\sqrt{T} \qquad \forall \boldsymbol{u} \in \mathbb{R}^d$
- $R_T(\boldsymbol{u}) \le \|\boldsymbol{u}\|_2 \sqrt{T}$ for $\alpha = \|\boldsymbol{u}\|_2$
- Equivalent to running OGD with projection in the Euclidean ball of radius $\|\boldsymbol{u}\|_2$
- This bound cannot be simultaneously achieved for all $\boldsymbol{u}$!

## Main idea

- Control $R_T(\boldsymbol{u})$ by learning length $w = \|\boldsymbol{u}\|$ and direction $\boldsymbol{v} = \boldsymbol{u}/\|\boldsymbol{u}\|$ separately

# Main idea

- Control $R_T(\boldsymbol{u})$ by learning length $w = \|\boldsymbol{u}\|$ and direction $\boldsymbol{v} = \boldsymbol{u}/\|\boldsymbol{u}\|$ separately
- The direction can be learned via OMD run in the unit ball

# Main idea

- Control $R_T(\boldsymbol{u})$ by learning length $w = \|\boldsymbol{u}\|$ and direction $\boldsymbol{v} = \boldsymbol{u}/\|\boldsymbol{u}\|$ separately
- The direction can be learned via OMD run in the unit ball
- The length is learned using a parameterless $1$-dimensional online learning algorithm

# Main idea

- Control $R_T(\boldsymbol{u})$ by learning length $w = \|\boldsymbol{u}\|$ and direction $\boldsymbol{v} = \boldsymbol{u}/\|\boldsymbol{u}\|$ separately
- The direction can be learned via OMD run in the unit ball
- The length is learned using a parameterless $1$-dimensional online learning algorithm
- One such algorithm has regret $R_T(w) = \mathcal{O}\big(|w|\sqrt{T\ln(T)}\big)$ for all $w \in R$

# Analysis

# Analysis

$$R_T(\boldsymbol{u}) = \sum_{t=1}^{T} \ell_t(w_t \boldsymbol{v}_t) - \sum_{t=1}^{T} \ell_t(\boldsymbol{u})$$

# Analysis

$$
\begin{aligned}
R_T(\boldsymbol{u}) &= \sum_{t=1}^{T} \ell_t(w_t \boldsymbol{v}_t) - \sum_{t=1}^{T} \ell_t(\boldsymbol{u}) \\
&\le \sum_{t=1}^{T} \boldsymbol{g}_t^\top (w_t \boldsymbol{v}_t - \boldsymbol{u}) && \text{(linearized regret)}
\end{aligned}
$$

# Analysis

$$
\begin{aligned}
R_T(\boldsymbol{u}) &= \sum_{t=1}^{T} \ell_t(w_t \boldsymbol{v}_t) - \sum_{t=1}^{T} \ell_t(\boldsymbol{u}) \\
&\le \sum_{t=1}^{T} \boldsymbol{g}_t^\top (w_t \boldsymbol{v}_t - \boldsymbol{u}) && \text{(linearized regret)} \\
&= \sum_{t=1}^{T} \Big( w_t \, \boldsymbol{g}_t^\top \boldsymbol{v}_t - \|\boldsymbol{u}\| \, \boldsymbol{g}_t^\top \boldsymbol{v}_t \Big) + \|\boldsymbol{u}\| \sum_{t=1}^{T} \left( \boldsymbol{g}_t^\top \boldsymbol{v}_t - \boldsymbol{g}_t^\top \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|} \right)
\end{aligned}
$$

# Analysis

$$
\begin{aligned}
R_T(\boldsymbol{u}) &= \sum_{t=1}^{T} \ell_t(w_t \boldsymbol{v}_t) - \sum_{t=1}^{T} \ell_t(\boldsymbol{u}) \\
&\leq \sum_{t=1}^{T} \boldsymbol{g}_t^\top (w_t \boldsymbol{v}_t - \boldsymbol{u}) && \text{(linearized regret)} \\
&= \sum_{t=1}^{T} \underbrace{\left(w_t\, \ell_t'(w_t) - \|\boldsymbol{u}\|\, \ell_t'(w_t)\right)}_{\text{parameterless}} + \|\boldsymbol{u}\| \sum_{t=1}^{T} \underbrace{\left(\boldsymbol{g}_t^\top \boldsymbol{v}_t - \boldsymbol{g}_t^\top \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|}\right)}_{\text{OMD}}
\end{aligned}
$$

# Analysis

$$
\begin{aligned}
R_T(\boldsymbol{u}) &= \sum_{t=1}^{T} \ell_t(w_t \boldsymbol{v}_t) - \sum_{t=1}^{T} \ell_t(\boldsymbol{u}) \\
&\leq \sum_{t=1}^{T} \boldsymbol{g}_t^\top (w_t \boldsymbol{v}_t - \boldsymbol{u}) && \text{(linearized regret)} \\
&= \sum_{t=1}^{T} \underbrace{\left(w_t\, \ell_t'(w_t) - \|\boldsymbol{u}\|\, \ell_t'(w_t)\right)}_{\text{parameterless}} + \|\boldsymbol{u}\| \sum_{t=1}^{T} \underbrace{\left(\boldsymbol{g}_t^\top \boldsymbol{v}_t - \boldsymbol{g}_t^\top \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|}\right)}_{\text{OMD}}
\end{aligned}
$$

$$
R_T(\boldsymbol{u}) = \mathcal{O}\left(\left(\sqrt{\ln\left(\|\boldsymbol{u}\|^2 T + 1\right)} + 1\right) \|\boldsymbol{u}\| \sqrt{T} + 1\right) \qquad \forall \boldsymbol{u} \in \mathbb{R}^d
$$

# Analysis

$$
\begin{aligned}
R_T(\boldsymbol{u}) &= \sum_{t=1}^{T} \ell_t(w_t \boldsymbol{v}_t) - \sum_{t=1}^{T} \ell_t(\boldsymbol{u}) \\
&\le \sum_{t=1}^{T} \boldsymbol{g}_t^\top (w_t \boldsymbol{v}_t - \boldsymbol{u}) && \text{(linearized regret)} \\
&= \sum_{t=1}^{T} \underbrace{\left(w_t\, \ell_t'(w_t) - \|\boldsymbol{u}\|\, \ell_t'(w_t)\right)}_{\text{parameterless}} + \|\boldsymbol{u}\| \underbrace{\sum_{t=1}^{T} \left(\boldsymbol{g}_t^\top \boldsymbol{v}_t - \boldsymbol{g}_t^\top \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|}\right)}_{\text{OMD}}
\end{aligned}
$$

$$
R_T(\boldsymbol{u}) = \mathcal{O}\left(\left(\underbrace{\sqrt{\ln\left(\|\boldsymbol{u}\|^2 T + 1\right)}}_{\text{unavoidable}} + 1\right) \|\boldsymbol{u}\| \sqrt{T} + 1\right) \qquad \forall \boldsymbol{u} \in \mathbb{R}^d
$$

# Learning and investing

$1$-dimensional parameterless online algorithms extracted from investment strategies

# Learning and investing

1-dimensional parameterless online algorithms extracted from investment strategies

The betting game

# Learning and investing

$1$-dimensional parameterless online algorithms extracted from investment strategies

## The betting game

- The bettor starts out with an initial wealth of $C_0 = 1$

# Learning and investing

$1$-dimensional parameterless online algorithms extracted from investment strategies

## The betting game

- The bettor starts out with an initial wealth of $C_0 = 1$
- In each round $t = 1, 2, \ldots$ of the game

# Learning and investing

1-dimensional parameterless online algorithms extracted from investment strategies

## The betting game

- The bettor starts out with an initial wealth of $C_0 = 1$
- In each round $t = 1, 2, \ldots$ of the game
   1. The bettor bets $\alpha_t \in [-1, 1]$

# Learning and investing

$1$-dimensional parameterless online algorithms extracted from investment strategies

## The betting game

- The bettor starts out with an initial wealth of $C_0 = 1$
- In each round $t = 1, 2, \ldots$ of the game
  1. The bettor bets $\alpha_t \in [-1, 1]$
  2. The market reveals $x_t \in [-1, 1]$

# Learning and investing

$1$-dimensional parameterless online algorithms extracted from investment strategies

## The betting game

- The bettor starts out with an initial wealth of $C_0 = 1$
- In each round $t = 1, 2, \dots$ of the game
  1. The bettor bets $\alpha_t \in [-1, 1]$
  2. The market reveals $x_t \in [-1, 1]$
  3. The bettor's wealth is $C_{t+1} = (1 + \alpha_t x_t) C_t$

# Learning and investing

$1$-dimensional parameterless online algorithms extracted from investment strategies

## The betting game

- The bettor starts out with an initial wealth of $C_0 = 1$
- In each round $t = 1, 2, \dots$ of the game
  1. The bettor bets $\alpha_t \in [-1, 1]$
  2. The market reveals $x_t \in [-1, 1]$
  3. The bettor's wealth is $C_{t+1} = (1 + \alpha_t x_t) C_t$

## Reduction to learning

# Learning and investing

$1$-dimensional parameterless online algorithms extracted from investment strategies

## The betting game

- The bettor starts out with an initial wealth of $C_0 = 1$
- In each round $t = 1, 2, \ldots$ of the game
  1. The bettor bets $\alpha_t \in [-1, 1]$
  2. The market reveals $x_t \in [-1, 1]$
  3. The bettor's wealth is $C_{t+1} = (1 + \alpha_t x_t) C_t$

## Reduction to learning

- $w_t = \alpha_t \, C_t$

# Learning and investing

$1$-dimensional parameterless online algorithms extracted from investment strategies

## The betting game

- The bettor starts out with an initial wealth of $C_0 = 1$
- In each round $t = 1, 2, \ldots$ of the game
  1. The bettor bets $\alpha_t \in [-1, 1]$
  2. The market reveals $x_t \in [-1, 1]$
  3. The bettor's wealth is $C_{t+1} = (1 + \alpha_t x_t) C_t$

## Reduction to learning

- $w_t = \alpha_t \, C_t$
- $x_t = -\ell_t'(w_t)$

# Learning and investing

$1$-dimensional parameterless online algorithms extracted from investment strategies

## The betting game

- The bettor starts out with an initial wealth of $C_0 = 1$
- In each round $t = 1, 2, \ldots$ of the game
  1. The bettor bets $\alpha_t \in [-1, 1]$
  2. The market reveals $x_t \in [-1, 1]$
  3. The bettor's wealth is $C_{t+1} = (1 + \alpha_t x_t) C_t$

## Reduction to learning

- $w_t = \alpha_t \, C_t$
- $x_t = -\ell_t'(w_t)$
- $C_T = \prod_{t=1}^T (1 + \alpha_t x_t) = 1 + \sum_{t=1}^T w_t x_t = 1 - \sum_{t=1}^T w_t \, \ell_t'(w_t)$

# Learning and investing

$1$-dimensional parameterless online algorithms extracted from investment strategies

## The betting game

- The bettor starts out with an initial wealth of $C_0 = 1$
- In each round $t = 1, 2, \ldots$ of the game
  1. The bettor bets $\alpha_t \in [-1, 1]$
  2. The market reveals $x_t \in [-1, 1]$
  3. The bettor's wealth is $C_{t+1} = (1 + \alpha_t x_t) C_t$

## Reduction to learning

- $w_t = \alpha_t \, C_t$
- $x_t = -\ell_t'(w_t)$
- $C_T = \prod_{t=1}^T (1 + \alpha_t x_t) = 1 + \sum_{t=1}^T w_t x_t = 1 - \sum_{t=1}^T w_t \, \ell_t'(w_t)$
- A lower bound on $C_T$ implies an upper bound on $R_T(w)$ for all $w \in \mathbb{R}$

# Other notions of regret

# Other notions of regret

- If the loss sequence $\ell_1, \ell_2, \dots$ is such that no $\boldsymbol{u} \in \mathbb{V}$ achieves a small cumulative loss $\ell_1(\boldsymbol{u}) + \ell_2(\boldsymbol{u}) + \cdots$, then regret bounds are meaningless

## Other notions of regret

- If the loss sequence $\ell_1, \ell_2, \ldots$ is such that no $\boldsymbol{u} \in \mathbb{V}$ achieves a small cumulative loss $\ell_1(\boldsymbol{u}) + \ell_2(\boldsymbol{u}) + \cdots$, then regret bounds are meaningless
- Lack of a single good minimizer in $\mathbb{V}$ caused by a highly nonstationary data sequence

# Other notions of regret

- If the loss sequence $\ell_1, \ell_2, \dots$ is such that no $\boldsymbol{u} \in \mathbb{V}$ achieves a small cumulative loss $\ell_1(\boldsymbol{u}) + \ell_2(\boldsymbol{u}) + \cdots$, then regret bounds are meaningless
- Lack of a single good minimizer in $\mathbb{V}$ caused by a highly nonstationary data sequence
- In this case, the regret should be replaced by more robust measures

# Other notions of regret

- If the loss sequence $\ell_1, \ell_2, \dots$ is such that no $\boldsymbol{u} \in \mathbb{V}$ achieves a small cumulative loss $\ell_1(\boldsymbol{u}) + \ell_2(\boldsymbol{u}) + \cdots$, then regret bounds are meaningless
- Lack of a single good minimizer in $\mathbb{V}$ caused by a highly nonstationary data sequence
- In this case, the regret should be replaced by more robust measures

- Dynamic regret $\quad R_T^{\mathrm{dyn}}(\boldsymbol{u}_1, \dots, \boldsymbol{u}_T) = \sum_{t=1}^{T} \ell_t(\boldsymbol{w}_t) - \sum_{t=1}^{T} \ell_t(\boldsymbol{u}_t) \qquad$ where $\boldsymbol{u}_1, \dots, \boldsymbol{u}_T \in \mathbb{V}$

# Other notions of regret

- If the loss sequence $\ell_1, \ell_2, \ldots$ is such that no $\boldsymbol{u} \in \mathbb{V}$ achieves a small cumulative loss $\ell_1(\boldsymbol{u}) + \ell_2(\boldsymbol{u}) + \cdots$, then regret bounds are meaningless
- Lack of a single good minimizer in $\mathbb{V}$ caused by a highly nonstationary data sequence
- In this case, the regret should be replaced by more robust measures

- Dynamic regret $\quad R_T^{\mathrm{dyn}}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T) = \sum_{t=1}^{T} \ell_t(\boldsymbol{w}_t) - \sum_{t=1}^{T} \ell_t(\boldsymbol{u}_t) \quad$ where $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T \in \mathbb{V}$

- Complexity parameter: $\Pi_T = \sum_{t=1}^{T-1} \|\boldsymbol{u}_{t+1} - \boldsymbol{u}_t\|$

## Other notions of regret

- If the loss sequence $\ell_1, \ell_2, \dots$ is such that no $\boldsymbol{u} \in \mathbb{V}$ achieves a small cumulative loss $\ell_1(\boldsymbol{u}) + \ell_2(\boldsymbol{u}) + \cdots$, then regret bounds are meaningless
- Lack of a single good minimizer in $\mathbb{V}$ caused by a highly nonstationary data sequence
- In this case, the regret should be replaced by more robust measures

- Dynamic regret $\quad R_T^{\mathrm{dyn}}(\boldsymbol{u}_1, \dots, \boldsymbol{u}_T) = \sum_{t=1}^{T} \ell_t(\boldsymbol{w}_t) - \sum_{t=1}^{T} \ell_t(\boldsymbol{u}_t) \quad$ where $\boldsymbol{u}_1, \dots, \boldsymbol{u}_T \in \mathbb{V}$
- Complexity parameter: $\Pi_T = \sum_{t=1}^{T-1} \|\boldsymbol{u}_{t+1} - \boldsymbol{u}_t\|$
- Lower bound: $\Omega\big(L\sqrt{(D + \Pi_T) DT}\big)$

# Other notions of regret

- If the loss sequence $\ell_1, \ell_2, \dots$ is such that no $\boldsymbol{u} \in \mathbb{V}$ achieves a small cumulative loss $\ell_1(\boldsymbol{u}) + \ell_2(\boldsymbol{u}) + \cdots$, then regret bounds are meaningless
- Lack of a single good minimizer in $\mathbb{V}$ caused by a highly nonstationary data sequence
- In this case, the regret should be replaced by more robust measures

- Dynamic regret $\quad R_T^{\mathrm{dyn}}(\boldsymbol{u}_1, \dots, \boldsymbol{u}_T) = \sum_{t=1}^{T} \ell_t(\boldsymbol{w}_t) - \sum_{t=1}^{T} \ell_t(\boldsymbol{u}_t) \qquad$ where $\boldsymbol{u}_1, \dots, \boldsymbol{u}_T \in \mathbb{V}$
- Complexity parameter: $\Pi_T = \sum_{t=1}^{T-1} \|\boldsymbol{u}_{t+1} - \boldsymbol{u}_t\|$
- Lower bound: $\Omega\big(L\sqrt{(D + \Pi_T)DT}\big)$
- When $\Pi_T = 0$ this reduces to the standard lower bound $\Omega(LD\sqrt{T})$

# Other notions of regret

- If the loss sequence $\ell_1, \ell_2, \dots$ is such that no $\boldsymbol{u} \in \mathbb{V}$ achieves a small cumulative loss $\ell_1(\boldsymbol{u}) + \ell_2(\boldsymbol{u}) + \cdots$, then regret bounds are meaningless
- Lack of a single good minimizer in $\mathbb{V}$ caused by a highly nonstationary data sequence
- In this case, the regret should be replaced by more robust measures

- Dynamic regret $\quad R_T^{\mathrm{dyn}}(\boldsymbol{u}_1, \dots, \boldsymbol{u}_T) = \sum_{t=1}^{T} \ell_t(\boldsymbol{w}_t) - \sum_{t=1}^{T} \ell_t(\boldsymbol{u}_t) \quad$ where $\boldsymbol{u}_1, \dots, \boldsymbol{u}_T \in \mathbb{V}$

- Complexity parameter: $\Pi_T = \sum_{t=1}^{T-1} \|\boldsymbol{u}_{t+1} - \boldsymbol{u}_t\|$
- Lower bound: $\Omega\big(L\sqrt{(D + \Pi_T)DT}\big)$
- When $\Pi_T = 0$ this reduces to the standard lower bound $\Omega(LD\sqrt{T})$
- Matching upper bound obtained by using Hedge to aggregate $\mathcal{O}(\ln T)$ instances of OGD each tuned to a different $\Pi_T$

# Adaptive regret

# Adaptive regret

- Evaluate the performance of the online algorithm against that of the best fixed comparator in any interval of time

# Adaptive regret

- Evaluate the performance of the online algorithm against that of the best fixed comparator in any interval of time
- $R^{\mathrm{ada}}_{\tau,T} = \max_{s=1,\ldots,T-\tau+1} \left( \sum_{t=s}^{s+\tau-1} \ell_t(\boldsymbol{w}_t) - \min_{\boldsymbol{u}\in\mathbb{V}} \sum_{t=s}^{s+\tau-1} \ell_t(\boldsymbol{u}) \right)$ where $\tau \in \{1,\ldots,T\}$

# Adaptive regret

- Evaluate the performance of the online algorithm against that of the best fixed comparator in any interval of time
- $R^{\text{ada}}_{\tau,T} = \max_{s=1,\dots,T-\tau+1} \left( \sum_{t=s}^{s+\tau-1} \ell_t(\boldsymbol{w}_t) - \min_{\boldsymbol{u}\in\mathbb{V}} \sum_{t=s}^{s+\tau-1} \ell_t(\boldsymbol{u}) \right)$ where $\tau \in \{1,\dots,T\}$
- Best known upper bound: $R^{\text{ada}}_{\tau,T}(\boldsymbol{u}) = \mathcal{O}\big(DL\sqrt{\tau} + \sqrt{(\ln T)\tau}\big)$

# Adaptive regret

- Evaluate the performance of the online algorithm against that of the best fixed comparator in any interval of time
- $R^{\text{ada}}_{\tau,T} = \max_{s=1,\ldots,T-\tau+1} \left( \sum_{t=s}^{s+\tau-1} \ell_t(\boldsymbol{w}_t) - \min_{\boldsymbol{u}\in\mathbb{V}} \sum_{t=s}^{s+\tau-1} \ell_t(\boldsymbol{u}) \right)$ where $\tau \in \{1,\ldots,T\}$
- Best known upper bound: $R^{\text{ada}}_{\tau,T}(\boldsymbol{u}) = \mathcal{O}\big(DL\sqrt{\tau} + \sqrt{(\ln T)\tau}\big)$
- Obtained by combining several instances of a standard online algorithm each run in a specific interval of time

# Adaptive regret

- Evaluate the performance of the online algorithm against that of the best fixed comparator in any interval of time
- $R^{\mathrm{ada}}_{\tau,T} = \max_{s=1,\dots,T-\tau+1} \left( \sum_{t=s}^{s+\tau-1} \ell_t(\boldsymbol{w}_t) - \min_{\boldsymbol{u}\in\mathbb{V}} \sum_{t=s}^{s+\tau-1} \ell_t(\boldsymbol{u}) \right)$ where $\tau \in \{1,\dots,T\}$
- Best known upper bound: $R^{\mathrm{ada}}_{\tau,T}(\boldsymbol{u}) = \mathcal{O}\big(DL\sqrt{\tau} + \sqrt{(\ln T)\tau}\big)$
- Obtained by combining several instances of a standard online algorithm each run in a specific interval of time
- The set of intervals is carefully designed so that the overall number of instances to be run is $\mathcal{O}(\ln T)$

# From sequential to statistical learning

- Assume $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots$ is the realization of i.i.d. draws from $\mathcal{D}$

# From sequential to statistical learning

- Assume $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots$ is the realization of i.i.d. draws from $\mathcal{D}$
- Let $\overline{\boldsymbol{w}} = \dfrac{1}{T}\displaystyle\sum_{t=1}^{T} \boldsymbol{w}_t$

# From sequential to statistical learning

- Assume $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots$ is the realization of i.i.d. draws from $\mathcal{D}$
- Let $\overline{\boldsymbol{w}} = \dfrac{1}{T}\displaystyle\sum_{t=1}^{T} \boldsymbol{w}_t$
- Linear prediction with convex loss $\ell(\boldsymbol{w}^\top \boldsymbol{x}, y_t)$

$$\ell_{\mathcal{D}}(\overline{\boldsymbol{w}}) = \mathbb{E}\Big[\ell(\overline{\boldsymbol{w}}^\top \boldsymbol{X}), Y)\Big] \le \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \ell_t(\boldsymbol{w}_t^\top \boldsymbol{X}, Y)\right] = \frac{1}{T}\sum_{t=1}^{T} \ell_{\mathcal{D}}(\boldsymbol{w}_t)$$

# From sequential to statistical learning

- Assume $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots$ is the realization of i.i.d. draws from $\mathcal{D}$
- Let $\overline{\boldsymbol{w}} = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{w}_t$
- Linear prediction with convex loss $\ell(\boldsymbol{w}^\top \boldsymbol{x}, y_t)$

$$\ell_{\mathcal{D}}(\overline{\boldsymbol{w}}) = \mathbb{E}\Big[\ell(\overline{\boldsymbol{w}}^\top \boldsymbol{X}), Y)\Big] \le \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \ell_t(\boldsymbol{w}_t^\top \boldsymbol{X}, Y)\right] = \frac{1}{T}\sum_{t=1}^{T} \ell_{\mathcal{D}}(\boldsymbol{w}_t)$$

- Note also that $\mathbb{E}\Big[\ell_{\mathcal{D}}(\boldsymbol{w}_t) - \ell(\boldsymbol{w}_t^\top \boldsymbol{X}_t, Y_t) \,\Big|\, (\boldsymbol{X}_1, Y_1), \dots, (\boldsymbol{X}_{t-1}, Y_{t-1})\Big] = 0$

# From sequential to statistical learning

- Assume $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots$ is the realization of i.i.d. draws from $\mathcal{D}$
- Let $\overline{\boldsymbol{w}} = \dfrac{1}{T}\displaystyle\sum_{t=1}^{T} \boldsymbol{w}_t$
- Linear prediction with convex loss $\ell(\boldsymbol{w}^\top \boldsymbol{x}, y_t)$

$$\ell_{\mathcal{D}}(\overline{\boldsymbol{w}}) = \mathbb{E}\Big[\ell(\overline{\boldsymbol{w}}^\top \boldsymbol{X}), Y)\Big] \le \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \ell_t(\boldsymbol{w}_t^\top \boldsymbol{X}, Y)\right] = \frac{1}{T}\sum_{t=1}^{T} \ell_{\mathcal{D}}(\boldsymbol{w}_t)$$

- Note also that $\mathbb{E}\Big[\ell_{\mathcal{D}}(\boldsymbol{w}_t) - \ell(\boldsymbol{w}_t^\top \boldsymbol{X}_t, Y_t) \,\Big|\, (\boldsymbol{X}_1, Y_1), \dots, (\boldsymbol{X}_{t-1}, Y_{t-1})\Big] = 0$
- Then, by the bounded martingale concentration law,

$$\frac{1}{T}\sum_{t=1}^{T} \ell_{\mathcal{D}}(\boldsymbol{w}_t) \le \frac{1}{T}\sum_{t=1}^{T} \ell(\boldsymbol{w}_t^\top \boldsymbol{X}_t, Y_t) + \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) \qquad \text{w.h.p.}$$

# Contextual bandits (reloaded)

In practice, actions in bandit problems have features (ads, items on sale, etc.)

# Contextual bandits (reloaded)

In practice, actions in bandit problems have features (ads, items on sale, etc.)

For $t = 1, 2, \dots$

1. Observe finite set $C_t \subset \mathbb{R}^d$ of contexts (feature vectors)

# Contextual bandits (reloaded)

In practice, actions in bandit problems have features (ads, items on sale, etc.)

For $t = 1, 2, \ldots$

1. Observe finite set $C_t \subset \mathbb{R}^d$ of contexts (feature vectors)
2. Choose $\boldsymbol{x}_t \in C_t$

# Contextual bandits (reloaded)

In practice, actions in bandit problems have features (ads, items on sale, etc.)

For $t = 1, 2, \dots$

1. Observe finite set $C_t \subset \mathbb{R}^d$ of contexts (feature vectors)
2. Choose $\boldsymbol{x}_t \in C_t$
3. Get reward $Y_t$

# Contextual bandits (reloaded)

In practice, actions in bandit problems have features (ads, items on sale, etc.)

For $t = 1, 2, \ldots$

1. Observe finite set $C_t \subset \mathbb{R}^d$ of contexts (feature vectors)
2. Choose $\boldsymbol{x}_t \in C_t$
3. Get reward $Y_t$

We assume a linear model: $Y_t = \boldsymbol{w}^\top \boldsymbol{x}_t + Z_t$

# Contextual bandits (reloaded)

In practice, actions in bandit problems have features (ads, items on sale, etc.)

For $t = 1, 2, \ldots$

1. Observe finite set $C_t \subset \mathbb{R}^d$ of contexts (feature vectors)
2. Choose $\boldsymbol{x}_t \in C_t$
3. Get reward $Y_t$

We assume a linear model: $Y_t = \boldsymbol{w}^\top \boldsymbol{x}_t + Z_t$

- $\boldsymbol{w} \in \mathbb{R}^d$ is fixed and unknown, but $\|\boldsymbol{w}\| \leq D$ with $D$ known

# Contextual bandits (reloaded)

In practice, actions in bandit problems have features (ads, items on sale, etc.)

For $t = 1, 2, \dots$

1. Observe finite set $C_t \subset \mathbb{R}^d$ of contexts (feature vectors)
2. Choose $\boldsymbol{x}_t \in C_t$
3. Get reward $Y_t$

We assume a linear model: $Y_t = \boldsymbol{w}^\top \boldsymbol{x}_t + Z_t$

- $\boldsymbol{w} \in \mathbb{R}^d$ is fixed and unknown, but $\|\boldsymbol{w}\| \le D$ with $D$ known
- $Z_t$ are zero-mean with a known bound $R$ on the variance

# Contextual bandits (reloaded)

In practice, actions in bandit problems have features (ads, items on sale, etc.)

For $t = 1, 2, \ldots$

1. Observe finite set $C_t \subset \mathbb{R}^d$ of contexts (feature vectors)
2. Choose $\boldsymbol{x}_t \in C_t$
3. Get reward $Y_t$

We assume a linear model: $Y_t = \boldsymbol{w}^\top \boldsymbol{x}_t + Z_t$

- $\boldsymbol{w} \in \mathbb{R}^d$ is fixed and unknown, but $\|\boldsymbol{w}\| \leq D$ with $D$ known
- $Z_t$ are zero-mean with a known bound $R$ on the variance

Regret: $$R_T^{\mathrm{cont}} = \sum_{t=1}^{T} \max_{\boldsymbol{x} \in C_t} \boldsymbol{w}^\top \boldsymbol{x} - \sum_{t=1}^{T} \boldsymbol{w}^\top \boldsymbol{x}_t$$

# The confidence ellipsoid

Fix a sequence of contexts $C_1, \dots, C_t$ and choices $\boldsymbol{x}_s \in C_s$, $s = 1, \dots, t$
RLS estimate

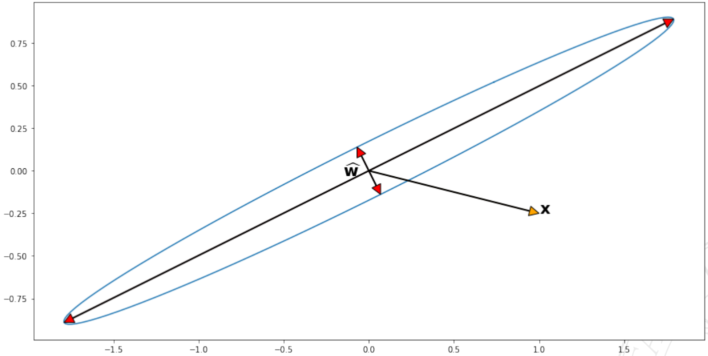
$$\widehat{\boldsymbol{w}}_t = V_t^{-1} \sum_{s=1}^{t} Y_s \boldsymbol{x}_s \qquad V_t = \lambda I_d + \underbrace{[\boldsymbol{x}_1, \dots, \boldsymbol{x}_t]}_{d \times t} [\boldsymbol{x}_1, \dots, \boldsymbol{x}_t]^\top$$

With high probability, $\boldsymbol{w} \in \mathcal{E}_t \equiv \left\{ \boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u} - \widehat{\boldsymbol{w}}\|_{V_t} \le \beta_t \right\}$

$\beta_t$ of order $D + R\sqrt{1 + d \ln\left(1 + \frac{t}{d}\right)}$

Think of $\mathcal{E}_t$ as a $d$-dimensional confidence interval

# The LinUCB/OFUL algorithm




Optimism in the face of uncertainty

$$\boldsymbol{x}_{t+1} = \operatorname*{argmax}_{\boldsymbol{x}\in C_{t+1}} \max_{\boldsymbol{u}\in\mathcal{E}_t} \boldsymbol{u}^\top \boldsymbol{x} = \operatorname*{argmax}_{\boldsymbol{x}\in C_t} \left( \widehat{\boldsymbol{w}}_t^\top \boldsymbol{x} + \beta_t \|\boldsymbol{x}\|_{V_t^{-1}} \right)$$

# Regret

- $R_T^{\text{cont}} = \mathcal{O}\left((d \ln T)\sqrt{T}\right)$

# Regret

- $R_T^{\text{cont}} = \mathcal{O}\left((d \ln T)\sqrt{T}\right)$
- Update time: $\Theta(d^2)$

# Regret

- $R_T^{\text{cont}} = \mathcal{O}\left((d \ln T)\sqrt{T}\right)$
- Update time: $\Theta(d^2)$
- This can be reduced to $\Theta(md)$ by sketching $\left[\boldsymbol{x}_1, \dots, \boldsymbol{x}_t\right]$ with a $d \times m$ matrix

# Regret

- $R_T^{\text{cont}} = \mathcal{O}\left((d \ln T)\sqrt{T}\right)$
- Update time: $\Theta(d^2)$
- This can be reduced to $\Theta(md)$ by sketching $[\boldsymbol{x}_1, \dots, \boldsymbol{x}_t]$ with a $d \times m$ matrix
- The spectral error $\varepsilon_m$ is bounded by the sum of the last $d - m + 1$ eigenvalues of $V_T$

# Regret

- $R_T^{\text{cont}} = \mathcal{O}\left((d \ln T)\sqrt{T}\right)$
- Update time: $\Theta(d^2)$
- This can be reduced to $\Theta(md)$ by sketching $\left[\boldsymbol{x}_1, \dots, \boldsymbol{x}_t\right]$ with a $d \times m$ matrix
- The spectral error $\varepsilon_m$ is bounded by the sum of the last $d - m + 1$ eigenvalues of $V_T$
- The regret becomes $R_T^{\text{cont}} = \widetilde{\mathcal{O}}\left((1+\varepsilon_m)^{3/2}(m + d\ln(1+\varepsilon_m))\sqrt{T}\right)$

# Regret

- $R_T^{\text{cont}} = \mathcal{O}\left((d \ln T)\sqrt{T}\right)$
- Update time: $\Theta(d^2)$
- This can be reduced to $\Theta(md)$ by sketching $[\boldsymbol{x}_1, \dots, \boldsymbol{x}_t]$ with a $d \times m$ matrix
- The spectral error $\varepsilon_m$ is bounded by the sum of the last $d - m + 1$ eigenvalues of $V_T$
- The regret becomes $R_T^{\text{cont}} = \widetilde{\mathcal{O}}\left((1+\varepsilon_m)^{3/2}(m + d\ln(1+\varepsilon_m))\sqrt{T}\right)$
- If the span of $\boldsymbol{x}_1, \dots, \boldsymbol{x}_T$ has dimension $m$, then $\varepsilon_m = 0$

# Regret

- $R_T^{\mathrm{cont}} = \mathcal{O}\left((d \ln T)\sqrt{T}\right)$
- Update time: $\Theta(d^2)$
- This can be reduced to $\Theta(md)$ by sketching $\left[\boldsymbol{x}_1, \dots, \boldsymbol{x}_t\right]$ with a $d \times m$ matrix
- The spectral error $\varepsilon_m$ is bounded by the sum of the last $d - m + 1$ eigenvalues of $V_T$
- The regret becomes $R_T^{\mathrm{cont}} = \widetilde{\mathcal{O}}\left((1+\varepsilon_m)^{3/2}(m + d \ln(1+\varepsilon_m))\sqrt{T}\right)$
- If the span of $\boldsymbol{x}_1, \dots, \boldsymbol{x}_T$ has dimension $m$, then $\varepsilon_m = 0$
- In this case, $R_T^{\mathrm{cont}} = \mathcal{O}\left((m \ln T)\sqrt{T}\right)$ for both algorithms

# Some references

List of wonderful references goes here