

# Tutorial on Bandits Games

**Sébastien Bubeck**



# Online Learning with Full Information

Adversary



Player

# Online Learning with Full Information

Adversary



Player

$A \in \{1, \dots, d\}$

# Online Learning with Full Information

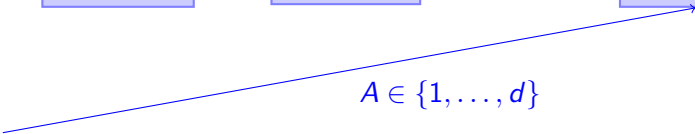
Adversary



---

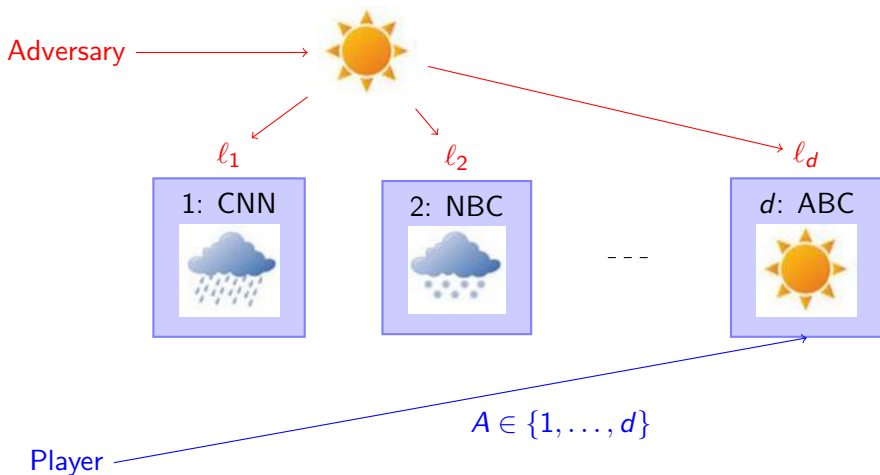


Player

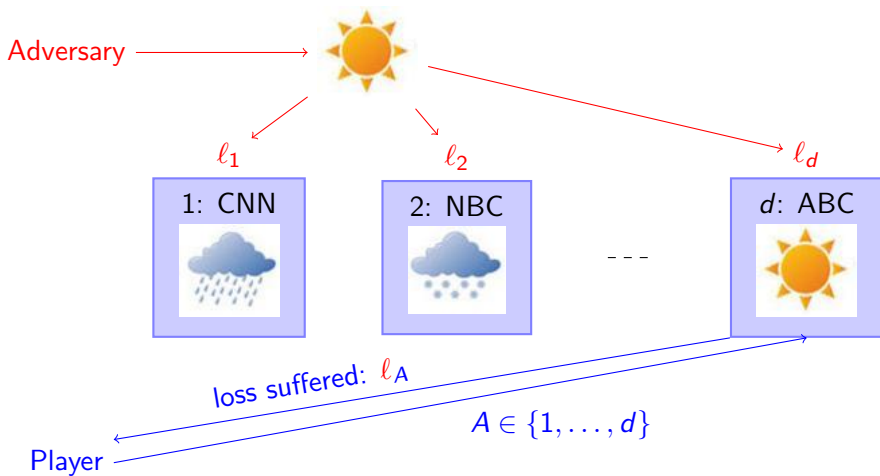


$A \in \{1, \dots, d\}$

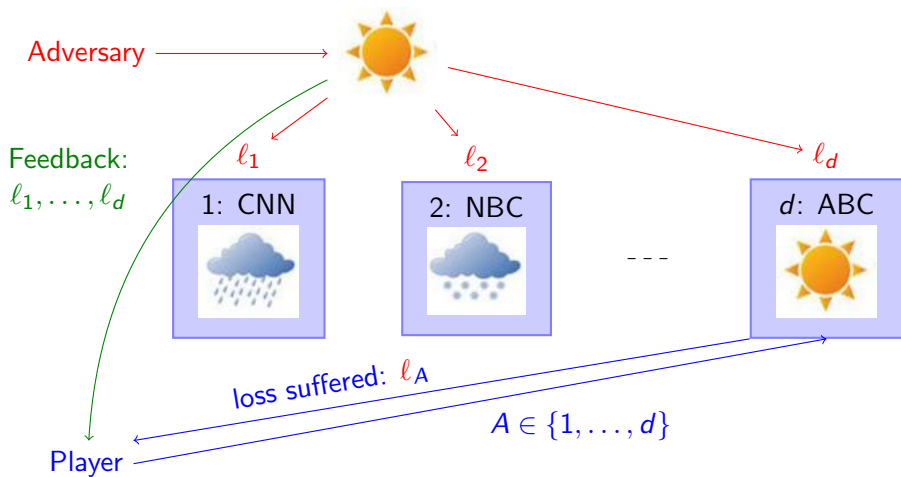
# Online Learning with Full Information



# Online Learning with Full Information



# Online Learning with Full Information



# Online Learning with Bandit Feedback

Adversary



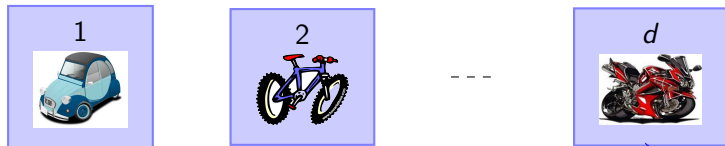
---



Player

# Online Learning with Bandit Feedback

Adversary



Player

$$A \in \{1, \dots, d\}$$

# Online Learning with Bandit Feedback

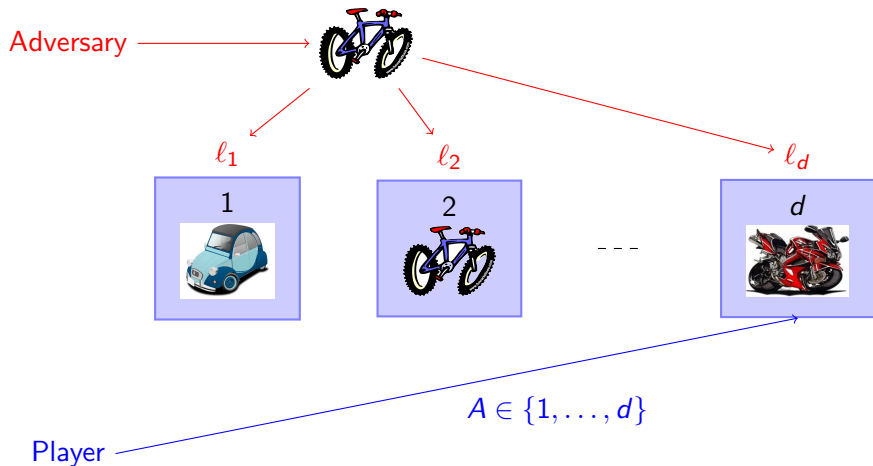


---

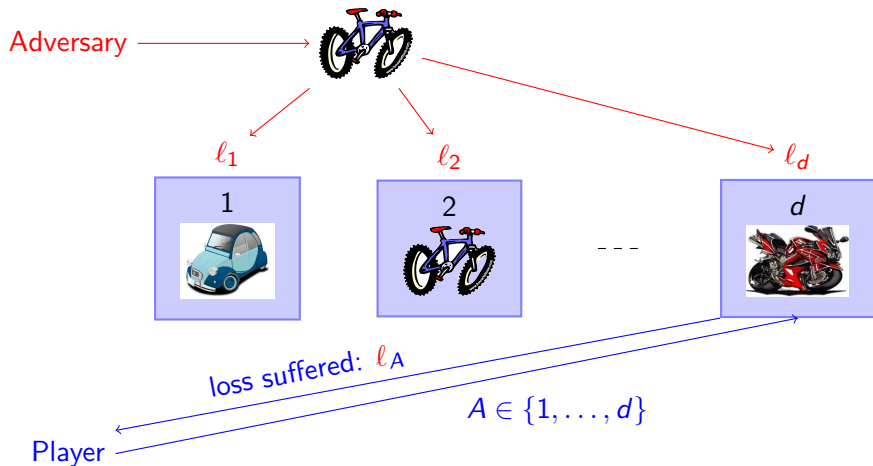


Player  $\longrightarrow$   $A \in \{1, \dots, d\}$

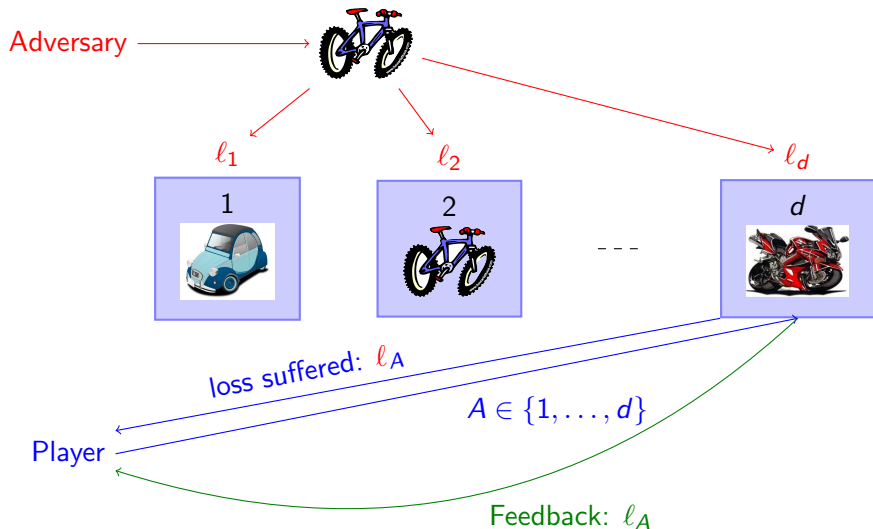
# Online Learning with Bandit Feedback



# Online Learning with Bandit Feedback



# Online Learning with Bandit Feedback



# Some Applications

Computer Go



Brain computer interface



Medical trials



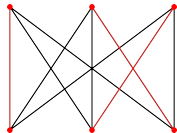
Packets routing



Ads placement

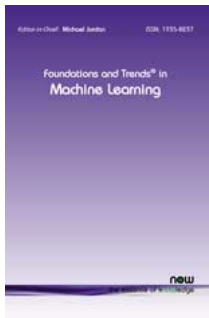


Dynamic allocation



# A little bit of advertising

Survey on multi-armed bandits to appear in



S. Bubeck and N. Cesa-Bianchi.

Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems.

To appear in *Foundations and Trends in Machine Learning*, 2012. (Draft available on my webpage.)

# Notation

For each round  $t = 1, 2, \dots, n$ ;

- 1 The player chooses an arm  $I_t \in \{1, \dots, d\}$ , possibly with the help of an external randomization.
- 2 Simultaneously the adversary chooses a loss vector  $\ell_t = (\ell_{1,t}, \dots, \ell_{d,t}) \in [0, 1]^d$ .
- 3 The player incurs the loss  $\ell_{I_t,t}$ , and observes:
  - The loss vector  $\ell_t$  in the full information setting.
  - Only the loss incurred  $\ell_{I_t,t}$  in the bandit setting.

**Goal:** Minimize the cumulative loss incurred. We consider the regret:

$$R_n = \mathbb{E} \sum_{t=1}^n \ell_{I_t,t} - \min_{i=1,\dots,d} \mathbb{E} \sum_{t=1}^n \ell_{i,t}.$$

# Notation

For each round  $t = 1, 2, \dots, n$ ;

- 1 The player chooses an arm  $I_t \in \{1, \dots, d\}$ , possibly with the help of an external randomization.
- 2 Simultaneously the adversary chooses a loss vector  $\ell_t = (\ell_{1,t}, \dots, \ell_{d,t}) \in [0, 1]^d$ .
- 3 The player incurs the loss  $\ell_{I_t,t}$ , and observes:
  - The loss vector  $\ell_t$  in the full information setting.
  - Only the loss incurred  $\ell_{I_t,t}$  in the bandit setting.

**Goal:** Minimize the cumulative loss incurred. We consider the regret:

$$R_n = \mathbb{E} \sum_{t=1}^n \ell_{I_t,t} - \min_{i=1,\dots,d} \mathbb{E} \sum_{t=1}^n \ell_{i,t}.$$

# Notation

For each round  $t = 1, 2, \dots, n$ ;

- 1 The player chooses an arm  $I_t \in \{1, \dots, d\}$ , possibly with the help of an external randomization.
- 2 Simultaneously the adversary chooses a loss vector  $\ell_t = (\ell_{1,t}, \dots, \ell_{d,t}) \in [0, 1]^d$ .
- 3 The player incurs the loss  $\ell_{I_t,t}$ , and observes:
  - The loss vector  $\ell_t$  in the full information setting.
  - Only the loss incurred  $\ell_{I_t,t}$  in the bandit setting.

**Goal:** Minimize the cumulative loss incurred. We consider the regret:

$$R_n = \mathbb{E} \sum_{t=1}^n \ell_{I_t,t} - \min_{i=1,\dots,d} \mathbb{E} \sum_{t=1}^n \ell_{i,t}.$$

# Notation

For each round  $t = 1, 2, \dots, n$ ;

- 1 The player chooses an arm  $I_t \in \{1, \dots, d\}$ , possibly with the help of an external randomization.
- 2 Simultaneously the adversary chooses a loss vector  $\ell_t = (\ell_{1,t}, \dots, \ell_{d,t}) \in [0, 1]^d$ .
- 3 The player incurs the loss  $\ell_{I_t,t}$ , and observes:
  - The loss vector  $\ell_t$  in the full information setting.
  - Only the loss incurred  $\ell_{I_t,t}$  in the bandit setting.

**Goal:** Minimize the cumulative loss incurred. We consider the regret:

$$R_n = \mathbb{E} \sum_{t=1}^n \ell_{I_t,t} - \min_{i=1,\dots,d} \mathbb{E} \sum_{t=1}^n \ell_{i,t}.$$

# Notation

For each round  $t = 1, 2, \dots, n$ ;

- 1 The player chooses an arm  $I_t \in \{1, \dots, d\}$ , possibly with the help of an external randomization.
- 2 Simultaneously the adversary chooses a loss vector  $\ell_t = (\ell_{1,t}, \dots, \ell_{d,t}) \in [0, 1]^d$ .
- 3 The player incurs the loss  $\ell_{I_t,t}$ , and observes:
  - The loss vector  $\ell_t$  in the full information setting.
  - Only the loss incurred  $\ell_{I_t,t}$  in the bandit setting.

**Goal:** Minimize the cumulative loss incurred. We consider the regret:

$$R_n = \mathbb{E} \sum_{t=1}^n \ell_{I_t,t} - \min_{i=1,\dots,d} \mathbb{E} \sum_{t=1}^n \ell_{i,t}.$$

# Notation

For each round  $t = 1, 2, \dots, n$ ;

- 1 The player chooses an arm  $I_t \in \{1, \dots, d\}$ , possibly with the help of an external randomization.
- 2 Simultaneously the adversary chooses a loss vector  $\ell_t = (\ell_{1,t}, \dots, \ell_{d,t}) \in [0, 1]^d$ .
- 3 The player incurs the loss  $\ell_{I_t,t}$ , and observes:
  - The loss vector  $\ell_t$  in the full information setting.
  - Only the loss incurred  $\ell_{I_t,t}$  in the bandit setting.

**Goal:** Minimize the cumulative loss incurred. We consider the regret:

$$R_n = \mathbb{E} \sum_{t=1}^n \ell_{I_t,t} - \min_{i=1,\dots,d} \mathbb{E} \sum_{t=1}^n \ell_{i,t}.$$

# Notation

For each round  $t = 1, 2, \dots, n$ ;

- 1 The player chooses an arm  $I_t \in \{1, \dots, d\}$ , possibly with the help of an external randomization.
- 2 Simultaneously the adversary chooses a loss vector  $\ell_t = (\ell_{1,t}, \dots, \ell_{d,t}) \in [0, 1]^d$ .
- 3 The player incurs the loss  $\ell_{I_t,t}$ , and observes:
  - The loss vector  $\ell_t$  in the full information setting.
  - Only the loss incurred  $\ell_{I_t,t}$  in the bandit setting.

**Goal:** Minimize the cumulative loss incurred. We consider the regret:

$$R_n = \mathbb{E} \sum_{t=1}^n \ell_{I_t,t} - \min_{i=1,\dots,d} \mathbb{E} \sum_{t=1}^n \ell_{i,t}.$$

# Exponential Weights (EW, EWA, MW, Hedge, ect)

Draw  $I_t$  at random from  $p_t$  where

$$p_t(i) = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \ell_{i,s}\right)}{\sum_{j=1}^d \exp\left(-\eta \sum_{s=1}^{t-1} \ell_{j,s}\right)}$$

Theorem (Cesa-Bianchi, Freund, Haussler, Helmbold, Schapire and Warmuth [1997])

*Exp satisfies*

$$R_n \leq \sqrt{\frac{n \log d}{2}}.$$

*Moreover for any strategy,*

$$\sup_{\text{adversaries}} R_n \geq \sqrt{\frac{n \log d}{2}} + o(\sqrt{n \log d}).$$

# Exponential Weights (EW, EWA, MW, Hedge, ect)

Draw  $I_t$  at random from  $p_t$  where

$$p_t(i) = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \ell_{i,s}\right)}{\sum_{j=1}^d \exp\left(-\eta \sum_{s=1}^{t-1} \ell_{j,s}\right)}$$

Theorem (Cesa-Bianchi, Freund, Haussler, Helmbold, Schapire and Warmuth [1997])

*Exp* satisfies

$$R_n \leq \sqrt{\frac{n \log d}{2}}.$$

Moreover for *any strategy*,

$$\sup_{\text{adversaries}} R_n \geq \sqrt{\frac{n \log d}{2}} + o(\sqrt{n \log d}).$$

# Exponential Weights (EW, EWA, MW, Hedge, ect)

Draw  $I_t$  at random from  $p_t$  where

$$p_t(i) = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \ell_{i,s}\right)}{\sum_{j=1}^d \exp\left(-\eta \sum_{s=1}^{t-1} \ell_{j,s}\right)}$$

Theorem (Cesa-Bianchi, Freund, Haussler, Helmbold, Schapire and Warmuth [1997])

*Exp* satisfies

$$R_n \leq \sqrt{\frac{n \log d}{2}}.$$

Moreover for *any strategy*,

$$\sup_{\text{adversaries}} R_n \geq \sqrt{\frac{n \log d}{2}} + o(\sqrt{n \log d}).$$

# Magic trick for bandit feedback

$$\tilde{\ell}_{i,t} = \frac{\ell_{i,t}}{p_t(i)} \mathbb{1}_{I_t=i},$$

is an unbiased estimate of  $\ell_{i,t}$ . We call **Exp3** the Exp strategy run on the estimated losses.

Theorem (Auer, Cesa-Bianchi, Freund and Schapire [2003])

*Exp3* satisfies:

$$R_n \leq \sqrt{2nd \log d}.$$

Moreover for any strategy,

$$\sup_{\text{adversaries}} R_n \geq \frac{1}{4} \sqrt{nd} + o(\sqrt{nd}).$$

# Magic trick for bandit feedback

$$\tilde{\ell}_{i,t} = \frac{\ell_{i,t}}{p_t(i)} \mathbb{1}_{I_t=i},$$

is an unbiased estimate of  $\ell_{i,t}$ . We call **Exp3** the Exp strategy run on the estimated losses.

Theorem (Auer, Cesa-Bianchi, Freund and Schapire [2003])

*Exp3* satisfies:

$$R_n \leq \sqrt{2nd \log d}.$$

Moreover for *any strategy*,

$$\sup_{\text{adversaries}} R_n \geq \frac{1}{4} \sqrt{nd} + o(\sqrt{nd}).$$

# Magic trick for bandit feedback

$$\tilde{\ell}_{i,t} = \frac{\ell_{i,t}}{p_t(i)} \mathbb{1}_{I_t=i},$$

is an unbiased estimate of  $\ell_{i,t}$ . We call **Exp3** the Exp strategy run on the estimated losses.

Theorem (Auer, Cesa-Bianchi, Freund and Schapire [2003])

*Exp3* satisfies:

$$R_n \leq \sqrt{2nd \log d}.$$

Moreover for *any strategy*,

$$\sup_{\text{adversaries}} R_n \geq \frac{1}{4} \sqrt{nd} + o(\sqrt{nd}).$$

# High probability bounds

What about bounds directly on the *true* regret

$$\sum_{t=1}^n \ell_{I_t,t} - \min_{i=1,\dots,d} \sum_{t=1}^n \ell_{i,t} ?$$

Auer et al. [2003] proposed [Exp3.P](#):

$$p_t(i) = (1 - \gamma) \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_{i,s}\right)}{\sum_{j=1}^d \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_{j,s}\right)} + \frac{\gamma}{d},$$

where

$$\tilde{\ell}_{i,t} = \frac{\ell_{i,t}}{p_t(i)} \mathbb{1}_{I_t=i} + \frac{\beta}{p_t(i)}.$$

# High probability bounds

What about bounds directly on the *true* regret

$$\sum_{t=1}^n \ell_{I_t, t} - \min_{i=1, \dots, d} \sum_{t=1}^n \ell_{i, t} ?$$

Auer et al. [2003] proposed [Exp3.P](#):

$$p_t(i) = (1 - \gamma) \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_{i,s}\right)}{\sum_{j=1}^d \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_{j,s}\right)} + \frac{\gamma}{d},$$

where

$$\tilde{\ell}_{i,t} = \frac{\ell_{i,t}}{p_t(i)} \mathbb{1}_{I_t=i} + \frac{\beta}{p_t(i)}.$$

# High probability bounds

What about bounds directly on the *true* regret

$$\sum_{t=1}^n \ell_{I_t,t} - \min_{i=1,\dots,d} \sum_{t=1}^n \ell_{i,t} ?$$

Auer et al. [2003] proposed [Exp3.P](#):

$$p_t(i) = (1 - \gamma) \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_{i,s}\right)}{\sum_{j=1}^d \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_{j,s}\right)} + \frac{\gamma}{d},$$

where

$$\tilde{\ell}_{i,t} = \frac{\ell_{i,t}}{p_t(i)} \mathbb{1}_{I_t=i} + \frac{\beta}{p_t(i)}.$$

# High probability bounds

Theorem (Auer et al. [2003], Audibert and Bubeck [2011])

Let  $\delta \in (0, 1)$ , with  $\beta = \sqrt{\frac{\log(d\delta^{-1})}{nd}}$ ,  $\eta = 0.95\sqrt{\frac{\log d}{nd}}$  and  $\gamma = 1.05\sqrt{\frac{d \log d}{n}}$ , Exp3.P satisfies with probability at least  $1 - \delta$ :

$$\sum_{t=1}^n \ell_{I_t, t} - \min_{i=1, \dots, d} \sum_{t=1}^n \ell_{i, t} \leq 5.15 \sqrt{nd \log(d\delta^{-1})}.$$

On the other hand with  $\beta = \sqrt{\frac{\log d}{nd}}$ ,  $\eta = 0.95\sqrt{\frac{\log d}{nd}}$  and  $\gamma = 1.05\sqrt{\frac{d \log d}{n}}$ , Exp3.P satisfies, for any  $\delta \in (0, 1)$ , with probability at least  $1 - \delta$ :

$$\sum_{t=1}^n \ell_{I_t, t} - \min_{i=1, \dots, d} \sum_{t=1}^n \ell_{i, t} \leq \sqrt{\frac{nd}{\log d}} \log(\delta^{-1}) + 5.15 \sqrt{nd \log d}.$$

# High probability bounds

Theorem (Auer et al. [2003], Audibert and Bubeck [2011])

Let  $\delta \in (0, 1)$ , with  $\beta = \sqrt{\frac{\log(d\delta^{-1})}{nd}}$ ,  $\eta = 0.95\sqrt{\frac{\log d}{nd}}$  and  $\gamma = 1.05\sqrt{\frac{d \log d}{n}}$ , *Exp3.P* satisfies with probability at least  $1 - \delta$ :

$$\sum_{t=1}^n \ell_{I_t, t} - \min_{i=1, \dots, d} \sum_{t=1}^n \ell_{i, t} \leq 5.15 \sqrt{nd \log(d\delta^{-1})}.$$

On the other hand with  $\beta = \sqrt{\frac{\log d}{nd}}$ ,  $\eta = 0.95\sqrt{\frac{\log d}{nd}}$  and  $\gamma = 1.05\sqrt{\frac{d \log d}{n}}$ , *Exp3.P* satisfies, for any  $\delta \in (0, 1)$ , with probability at least  $1 - \delta$ :

$$\sum_{t=1}^n \ell_{I_t, t} - \min_{i=1, \dots, d} \sum_{t=1}^n \ell_{i, t} \leq \sqrt{\frac{nd}{\log d}} \log(\delta^{-1}) + 5.15 \sqrt{nd \log d}.$$

## Other types of normalization

- **INF** (Implicitly Normalized Forecaster) is based on a potential function  $\psi : \mathbb{R}_-^* \rightarrow \mathbb{R}_+^*$  increasing, convex, twice continuously differentiable, and such that  $(0, 1] \subset \psi(\mathbb{R}_-^*)$ .
- At each time step INF computes the new probability distribution as follows:

$$p_t(i) = \psi \left( C_t - \sum_{s=1}^{t-1} \tilde{\ell}_{i,s} \right),$$

where  $C_t$  is the unique real number such that  $\sum_{i=1}^d p_t(i) = 1$ .

- $\psi(x) = \exp(\eta x) + \frac{\gamma}{d}$  corresponds exactly to the **Exp3** strategy.
- $\psi(x) = (-\eta x)^{-1/2} + \frac{\gamma}{d}$  is the **quadratic INF** strategy.

## Other types of normalization

- **INF** (Implicitly Normalized Forecaster) is based on a potential function  $\psi : \mathbb{R}_-^* \rightarrow \mathbb{R}_+^*$  increasing, convex, twice continuously differentiable, and such that  $(0, 1] \subset \psi(\mathbb{R}_-^*)$ .
- At each time step INF computes the new probability distribution as follows:

$$p_t(i) = \psi \left( C_t - \sum_{s=1}^{t-1} \tilde{\ell}_{i,s} \right),$$

where  $C_t$  is the unique real number such that  $\sum_{i=1}^d p_t(i) = 1$ .

- $\psi(x) = \exp(\eta x) + \frac{\gamma}{d}$  corresponds exactly to the **Exp3** strategy.
- $\psi(x) = (-\eta x)^{-1/2} + \frac{\gamma}{d}$  is the **quadratic INF** strategy.

## Other types of normalization

- **INF** (Implicitly Normalized Forecaster) is based on a potential function  $\psi : \mathbb{R}_-^* \rightarrow \mathbb{R}_+^*$  increasing, convex, twice continuously differentiable, and such that  $(0, 1] \subset \psi(\mathbb{R}_-^*)$ .
- At each time step INF computes the new probability distribution as follows:

$$p_t(i) = \psi \left( C_t - \sum_{s=1}^{t-1} \tilde{\ell}_{i,s} \right),$$

where  $C_t$  is the unique real number such that  $\sum_{i=1}^d p_t(i) = 1$ .

- $\psi(x) = \exp(\eta x) + \frac{\gamma}{d}$  corresponds exactly to the **Exp3** strategy.
- $\psi(x) = (-\eta x)^{-1/2} + \frac{\gamma}{d}$  is the **quadratic INF** strategy.

## Other types of normalization

- **INF** (Implicitly Normalized Forecaster) is based on a potential function  $\psi : \mathbb{R}_-^* \rightarrow \mathbb{R}_+^*$  increasing, convex, twice continuously differentiable, and such that  $(0, 1] \subset \psi(\mathbb{R}_-^*)$ .
- At each time step INF computes the new probability distribution as follows:

$$p_t(i) = \psi \left( C_t - \sum_{s=1}^{t-1} \tilde{\ell}_{i,s} \right),$$

where  $C_t$  is the unique real number such that  $\sum_{i=1}^d p_t(i) = 1$ .

- $\psi(x) = \exp(\eta x) + \frac{\gamma}{d}$  corresponds exactly to the **Exp3** strategy.
- $\psi(x) = (-\eta x)^{-1/2} + \frac{\gamma}{d}$  is the **quadratic INF** strategy.

Theorem (Audibert and Bubeck [2009], Audibert and Bubeck [2010], Audibert, Bubeck and Lugosi [2011])

*Quadratic INF* satisfies:

$$R_n \leq 2\sqrt{2nd}.$$

# Stochastic Assumption

## Assumption (Robbins [1952])

The sequence of losses  $(\ell_t)_{1 \leq t \leq n}$  is a sequence of i.i.d random variables.

For historical reasons in this setting we consider gains rather than losses and we introduce different notation:

- Let  $\nu_i$  be the unknown reward distribution underlying arm  $i$ ,  $\mu_i$  the mean of  $\nu_i$ ,  $\mu^* = \max_{1 \leq i \leq d} \mu_i$  and  $\Delta_i = \mu^* - \mu_i$ .
- Let  $X_{i,s} \sim \nu_i$  be the reward obtained when pulling arm  $i$  for the  $s^{\text{th}}$  time, and  $T_i(t) = \sum_{s=1}^t \mathbb{1}_{I_s=i}$  the number of times arm  $i$  was pulled up to time  $t$ .
- Thus here

$$R_n = n\mu^* - \mathbb{E} \sum_{t=1}^n \mu_{I_t} = \sum_{i=1}^d \Delta_i \mathbb{E} T_i(n).$$

# Stochastic Assumption

## Assumption (Robbins [1952])

The sequence of losses  $(\ell_t)_{1 \leq t \leq n}$  is a sequence of i.i.d random variables.

For historical reasons in this setting we consider gains rather than losses and we introduce different notation:

- Let  $\nu_i$  be the unknown reward distribution underlying arm  $i$ ,  $\mu_i$  the mean of  $\nu_i$ ,  $\mu^* = \max_{1 \leq i \leq d} \mu_i$  and  $\Delta_i = \mu^* - \mu_i$ .
- Let  $X_{i,s} \sim \nu_i$  be the reward obtained when pulling arm  $i$  for the  $s^{\text{th}}$  time, and  $T_i(t) = \sum_{s=1}^t \mathbb{1}_{I_s=i}$  the number of times arm  $i$  was pulled up to time  $t$ .
- Thus here

$$R_n = n\mu^* - \mathbb{E} \sum_{t=1}^n \mu_{I_t} = \sum_{i=1}^d \Delta_i \mathbb{E} T_i(n).$$

# Stochastic Assumption

## Assumption (Robbins [1952])

The sequence of losses  $(\ell_t)_{1 \leq t \leq n}$  is a sequence of i.i.d random variables.

For historical reasons in this setting we consider gains rather than losses and we introduce different notation:

- Let  $\nu_i$  be the unknown reward distribution underlying arm  $i$ ,  $\mu_i$  the mean of  $\nu_i$ ,  $\mu^* = \max_{1 \leq i \leq d} \mu_i$  and  $\Delta_i = \mu^* - \mu_i$ .
- Let  $X_{i,s} \sim \nu_i$  be the reward obtained when pulling arm  $i$  for the  $s^{\text{th}}$  time, and  $T_i(t) = \sum_{s=1}^t \mathbb{1}_{I_s=i}$  the number of times arm  $i$  was pulled up to time  $t$ .
- Thus here

$$R_n = n\mu^* - \mathbb{E} \sum_{t=1}^n \mu_{I_t} = \sum_{i=1}^d \Delta_i \mathbb{E} T_i(n).$$

## Assumption (Robbins [1952])

The sequence of losses  $(\ell_t)_{1 \leq t \leq n}$  is a sequence of i.i.d random variables.

For historical reasons in this setting we consider gains rather than losses and we introduce different notation:

- Let  $\nu_i$  be the unknown reward distribution underlying arm  $i$ ,  $\mu_i$  the mean of  $\nu_i$ ,  $\mu^* = \max_{1 \leq i \leq d} \mu_i$  and  $\Delta_i = \mu^* - \mu_i$ .
- Let  $X_{i,s} \sim \nu_i$  be the reward obtained when pulling arm  $i$  for the  $s^{\text{th}}$  time, and  $T_i(t) = \sum_{s=1}^t \mathbb{1}_{I_s=i}$  the number of times arm  $i$  was pulled up to time  $t$ .
- Thus here

$$R_n = n\mu^* - \mathbb{E} \sum_{t=1}^n \mu_{I_t} = \sum_{i=1}^d \Delta_i \mathbb{E} T_i(n).$$

## Assumption (Robbins [1952])

The sequence of losses  $(\ell_t)_{1 \leq t \leq n}$  is a sequence of i.i.d random variables.

For historical reasons in this setting we consider gains rather than losses and we introduce different notation:

- Let  $\nu_i$  be the unknown reward distribution underlying arm  $i$ ,  $\mu_i$  the mean of  $\nu_i$ ,  $\mu^* = \max_{1 \leq i \leq d} \mu_i$  and  $\Delta_i = \mu^* - \mu_i$ .
- Let  $X_{i,s} \sim \nu_i$  be the reward obtained when pulling arm  $i$  for the  $s^{\text{th}}$  time, and  $T_i(t) = \sum_{s=1}^t \mathbb{1}_{I_s=i}$  the number of times arm  $i$  was pulled up to time  $t$ .
- Thus here

$$R_n = n\mu^* - \mathbb{E} \sum_{t=1}^n \mu_{I_t} = \sum_{i=1}^d \Delta_i \mathbb{E} T_i(n).$$

# Optimism in face of uncertainty

**General principle:** given some observations from an unknown environment, build (with some probabilistic argument) a set of *possible* environments  $\Omega$ , then act as if the real environment was the most favorable one in  $\Omega$ .

**Application to stochastic bandits:** given the past rewards, build confidence intervals for the means ( $\mu_i$ ) (in particular build upper confidence bounds), then play the arm with the highest upper confidence bound.

# Optimism in face of uncertainty

**General principle:** given some observations from an unknown environment, build (with some probabilistic argument) a set of *possible* environments  $\Omega$ , then act as if the real environment was the most favorable one in  $\Omega$ .

**Application to stochastic bandits:** given the past rewards, build confidence intervals for the means  $(\mu_i)$  (in particular build upper confidence bounds), then play the arm with the highest upper confidence bound.

# UCB (Upper Confidence Bounds)

## Theorem (Hoeffding [1963])

Let  $X, X_1, \dots, X_t$  be i.i.d random variables in  $[0, 1]$ , then with probability at least  $1 - \delta$ ,

$$\mathbb{E}X \leq \frac{1}{t} \sum_{s=1}^t X_s + \sqrt{\frac{\log \delta^{-1}}{2t}}.$$

This directly suggests the famous UCB strategy of Auer, Cesa-Bianchi and Fischer [2002]:

$$I_t \in \operatorname{argmax}_{1 \leq i \leq d} \frac{1}{T_i(t-1)} \sum_{s=1}^{T_i(t-1)} X_{i,s} + \sqrt{\frac{2 \log t}{T_i(t-1)}}.$$

Auer et al. proved the following regret bound:

$$R_n \leq \sum_{i: \Delta_i > 0} \frac{10 \log n}{\Delta_i}.$$

# UCB (Upper Confidence Bounds)

## Theorem (Hoeffding [1963])

Let  $X, X_1, \dots, X_t$  be i.i.d random variables in  $[0, 1]$ , then with probability at least  $1 - \delta$ ,

$$\mathbb{E}X \leq \frac{1}{t} \sum_{s=1}^t X_s + \sqrt{\frac{\log \delta^{-1}}{2t}}.$$

This directly suggests the famous **UCB** strategy of Auer, Cesa-Bianchi and Fischer [2002]:

$$I_t \in \operatorname{argmax}_{1 \leq i \leq d} \frac{1}{T_i(t-1)} \sum_{s=1}^{T_i(t-1)} X_{i,s} + \sqrt{\frac{2 \log t}{T_i(t-1)}}.$$

Auer et al. proved the following regret bound:

$$R_n \leq \sum_{i: \Delta_i > 0} \frac{10 \log n}{\Delta_i}.$$

# UCB (Upper Confidence Bounds)

## Theorem (Hoeffding [1963])

Let  $X, X_1, \dots, X_t$  be i.i.d random variables in  $[0, 1]$ , then with probability at least  $1 - \delta$ ,

$$\mathbb{E}X \leq \frac{1}{t} \sum_{s=1}^t X_s + \sqrt{\frac{\log \delta^{-1}}{2t}}.$$

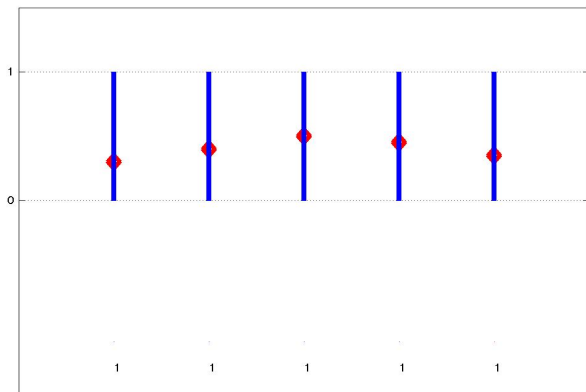
This directly suggests the famous **UCB** strategy of Auer, Cesa-Bianchi and Fischer [2002]:

$$I_t \in \operatorname{argmax}_{1 \leq i \leq d} \frac{1}{T_i(t-1)} \sum_{s=1}^{T_i(t-1)} X_{i,s} + \sqrt{\frac{2 \log t}{T_i(t-1)}}.$$

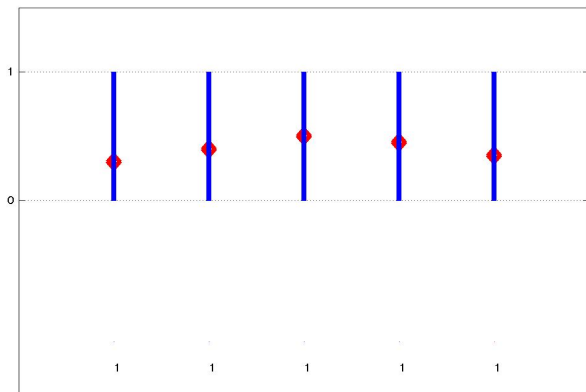
Auer et al. proved the following regret bound:

$$R_n \leq \sum_{i: \Delta_i > 0} \frac{10 \log n}{\Delta_i}.$$

# Illustration of UCB



# Illustration of UCB



## Lower bound

For any  $p, q \in [0, 1]$ , let

$$\text{kl}(p, q) = p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q}.$$

Theorem (Lai and Robbins [1985])

*Consider a consistent strategy, i.e. s.t.  $\forall a > 0$ , we have  $\mathbb{E}T_i(n) = o(n^a)$  if  $\Delta_i > 0$ . Then for any Bernoulli reward distributions,*

$$\liminf_{n \rightarrow +\infty} \frac{R_n}{\log n} \geq \sum_{i: \Delta_i > 0} \frac{\Delta_i}{\text{kl}(\mu_i, \mu^*)}.$$

Note that

$$\frac{1}{2\Delta_i} \geq \frac{\Delta_i}{\text{kl}(\mu_i, \mu^*)} \geq \frac{\mu^*(1 - \mu^*)}{2\Delta_i}.$$

## Lower bound

For any  $p, q \in [0, 1]$ , let

$$\text{kl}(p, q) = p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q}.$$

### Theorem (Lai and Robbins [1985])

Consider a consistent strategy, i.e. s.t.  $\forall a > 0$ , we have  $\mathbb{E}T_i(n) = o(n^a)$  if  $\Delta_i > 0$ . Then for any Bernoulli reward distributions,

$$\liminf_{n \rightarrow +\infty} \frac{R_n}{\log n} \geq \sum_{i: \Delta_i > 0} \frac{\Delta_i}{\text{kl}(\mu_i, \mu^*)}.$$

Note that

$$\frac{1}{2\Delta_i} \geq \frac{\Delta_i}{\text{kl}(\mu_i, \mu^*)} \geq \frac{\mu^*(1 - \mu^*)}{2\Delta_i}.$$

## Lower bound

For any  $p, q \in [0, 1]$ , let

$$\text{kl}(p, q) = p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q}.$$

### Theorem (Lai and Robbins [1985])

Consider a consistent strategy, i.e. s.t.  $\forall a > 0$ , we have  $\mathbb{E} T_i(n) = o(n^a)$  if  $\Delta_i > 0$ . Then for any Bernoulli reward distributions,

$$\liminf_{n \rightarrow +\infty} \frac{R_n}{\log n} \geq \sum_{i: \Delta_i > 0} \frac{\Delta_i}{\text{kl}(\mu_i, \mu^*)}.$$

Note that

$$\frac{1}{2\Delta_i} \geq \frac{\Delta_i}{\text{kl}(\mu_i, \mu^*)} \geq \frac{\mu^*(1 - \mu^*)}{2\Delta_i}.$$

**Theorem (Chernoff's inequality)**

Let  $X, X_1, \dots, X_t$  be i.i.d random variables in  $[0, 1]$ , then

$$\mathbb{P} \left( \frac{1}{t} \sum_{s=1}^t X_s \leq \mathbb{E}X - \epsilon \right) \leq \exp(-t \text{kl}(\mathbb{E}X - \epsilon, \mathbb{E}X)).$$

In particular this implies that with probability at least  $1 - \delta$ :

$$\mathbb{E}X \leq \max \left\{ q \in [0, 1] : \text{kl} \left( \frac{1}{t} \sum_{s=1}^t X_s, q \right) \leq \frac{\log \delta^{-1}}{t} \right\}.$$

## Theorem (Chernoff's inequality)

Let  $X, X_1, \dots, X_t$  be i.i.d random variables in  $[0, 1]$ , then

$$\mathbb{P} \left( \frac{1}{t} \sum_{s=1}^t X_s \leq \mathbb{E}X - \epsilon \right) \leq \exp(-t \text{kl}(\mathbb{E}X - \epsilon, \mathbb{E}X)).$$

In particular this implies that with probability at least  $1 - \delta$ :

$$\mathbb{E}X \leq \max \left\{ q \in [0, 1] : \text{kl} \left( \frac{1}{t} \sum_{s=1}^t X_s, q \right) \leq \frac{\log \delta^{-1}}{t} \right\}.$$

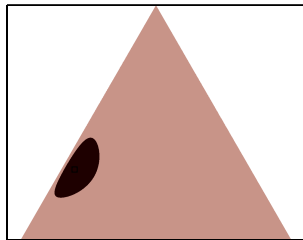
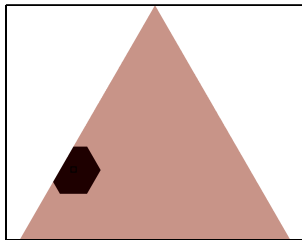
## Theorem (Chernoff's inequality)

Let  $X, X_1, \dots, X_t$  be i.i.d random variables in  $[0, 1]$ , then

$$\mathbb{P} \left( \frac{1}{t} \sum_{s=1}^t X_s \leq \mathbb{E}X - \epsilon \right) \leq \exp(-t \text{kl}(\mathbb{E}X - \epsilon, \mathbb{E}X)).$$

In particular this implies that with probability at least  $1 - \delta$ :

$$\mathbb{E}X \leq \max \left\{ q \in [0, 1] : \text{kl} \left( \frac{1}{t} \sum_{s=1}^t X_s, q \right) \leq \frac{\log \delta^{-1}}{t} \right\}.$$



Thus Chernoff's bound suggests the **KL-UCB** strategy of Garivier and Cappé [2011] (see also Honda and Takemura [2010], Maillard, Munos and Stoltz [2011]) :

$$I_t \in \operatorname{argmax}_{1 \leq i \leq d} \max \left\{ q \in [0, 1] : \operatorname{kl} \left( \frac{1}{T_i(t-1)} \sum_{s=1}^{T_i(t-1)} X_{i,s}, q \right) \leq \frac{(1+\epsilon) \log t}{T_i(t-1)} \right\}.$$

Garivier and Cappé proved the following regret bound for  $n$  large enough:

$$R_n \leq \sum_{i: \Delta_i > 0} (1 + 2\epsilon) \frac{\Delta_i}{\operatorname{kl}(\mu_i, \mu^*)} \log n.$$

Thus Chernoff's bound suggests the **KL-UCB** strategy of Garivier and Cappé [2011] (see also Honda and Takemura [2010], Maillard, Munos and Stoltz [2011]) :

$$I_t \in \operatorname{argmax}_{1 \leq i \leq d} \max \left\{ q \in [0, 1] : \operatorname{kl} \left( \frac{1}{T_i(t-1)} \sum_{s=1}^{T_i(t-1)} X_{i,s}, q \right) \leq \frac{(1+\epsilon) \log t}{T_i(t-1)} \right\}.$$

Garivier and Cappé proved the following regret bound for  $n$  large enough:

$$R_n \leq \sum_{i: \Delta_i > 0} (1 + 2\epsilon) \frac{\Delta_i}{\operatorname{kl}(\mu_i, \mu^*)} \log n.$$

Thus Chernoff's bound suggests the **KL-UCB** strategy of Garivier and Cappé [2011] (see also Honda and Takemura [2010], Maillard, Munos and Stoltz [2011]) :

$$I_t \in \operatorname{argmax}_{1 \leq i \leq d} \max \left\{ q \in [0, 1] : \operatorname{kl} \left( \frac{1}{T_i(t-1)} \sum_{s=1}^{T_i(t-1)} X_{i,s}, q \right) \leq \frac{(1+\epsilon) \log t}{T_i(t-1)} \right\}.$$

Garivier and Cappé proved the following regret bound for  $n$  large enough:

$$R_n \leq \sum_{i: \Delta_i > 0} (1 + 2\epsilon) \frac{\Delta_i}{\operatorname{kl}(\mu_i, \mu^*)} \log n.$$

# Heavy-tailed distributions

The standard UCB works for all  $\sigma^2$  - **subgaussian** distributions (not only bounded distributions), i.e. such that

$$\mathbb{E} \exp(\lambda(X - \mathbb{E}X)) \leq \frac{\sigma^2 \lambda^2}{2}, \forall \lambda \in \mathbb{R}.$$

It is easy to see that this equivalent to

$$\exists \alpha > 0 \text{ s.t. } \mathbb{E} \exp(\alpha X^2) < +\infty.$$

What happens for distributions with **heavier tails**? Can we get logarithmic regret if the distributions only have a **finite variance**?

# Heavy-tailed distributions

The standard UCB works for all  $\sigma^2$  - **subgaussian** distributions (not only bounded distributions), i.e. such that

$$\mathbb{E} \exp(\lambda(X - \mathbb{E}X)) \leq \frac{\sigma^2 \lambda^2}{2}, \forall \lambda \in \mathbb{R}.$$

It is easy to see that this equivalent to

$$\exists \alpha > 0 \text{ s.t. } \mathbb{E} \exp(\alpha X^2) < +\infty.$$

What happens for distributions with **heavier tails**? Can we get logarithmic regret if the distributions only have a **finite variance**?

# Heavy-tailed distributions

The standard UCB works for all  $\sigma^2$  - **subgaussian** distributions (not only bounded distributions), i.e. such that

$$\mathbb{E} \exp(\lambda(X - \mathbb{E}X)) \leq \frac{\sigma^2 \lambda^2}{2}, \forall \lambda \in \mathbb{R}.$$

It is easy to see that this equivalent to

$$\exists \alpha > 0 \text{ s.t. } \mathbb{E} \exp(\alpha X^2) < +\infty.$$

What happens for distributions with **heavier tails**? Can we get logarithmic regret if the distributions only have a **finite variance**?

# Median of means, Alon, Gibbons, Matias and Szegedy [2002]

## Lemma

Let  $X, X_1, \dots, X_t$  be i.i.d random variables such that

$\mathbb{E}(X - \mathbb{E}X)^2 \leq 1$ . Let  $\delta \in (0, 1)$ ,  $k = 8 \log \delta^{-1}$  and  $N = \frac{n}{8 \log \delta^{-1}}$ .

Then with probability at least  $1 - \delta$ ,

$$\mathbb{E}X \leq \text{median} \left( \frac{1}{N} \sum_{s=1}^N X_s, \dots, \frac{1}{N} \sum_{s=(k-1)N+1}^{kN} X_s \right) + 8 \sqrt{\frac{8 \log(\delta^{-1})}{n}}.$$

# Median of means, Alon, Gibbons, Matias and Szegedy [2002]

## Lemma

Let  $X, X_1, \dots, X_t$  be i.i.d random variables such that

$\mathbb{E}(X - \mathbb{E}X)^2 \leq 1$ . Let  $\delta \in (0, 1)$ ,  $k = 8 \log \delta^{-1}$  and  $N = \frac{n}{8 \log \delta^{-1}}$ .

Then with probability at least  $1 - \delta$ ,

$$\mathbb{E}X \leq \text{median} \left( \frac{1}{N} \sum_{s=1}^N X_s, \dots, \frac{1}{N} \sum_{s=(k-1)N+1}^{kN} X_s \right) + 8 \sqrt{\frac{8 \log(\delta^{-1})}{n}}.$$

# Median of means, Alon, Gibbons, Matias and Szegedy [2002]

## Lemma

Let  $X, X_1, \dots, X_t$  be i.i.d random variables such that

$\mathbb{E}(X - \mathbb{E}X)^2 \leq 1$ . Let  $\delta \in (0, 1)$ ,  $k = 8 \log \delta^{-1}$  and  $N = \frac{n}{8 \log \delta^{-1}}$ .

Then with probability at least  $1 - \delta$ ,

$$\mathbb{E}X \leq \text{median} \left( \frac{1}{N} \sum_{s=1}^N X_s, \dots, \frac{1}{N} \sum_{s=(k-1)N+1}^{kN} X_s \right) + 8 \sqrt{\frac{8 \log(\delta^{-1})}{n}}.$$

# Median of means, Alon, Gibbons, Matias and Szegedy [2002]

## Lemma

Let  $X, X_1, \dots, X_t$  be i.i.d random variables such that

$\mathbb{E}(X - \mathbb{E}X)^2 \leq 1$ . Let  $\delta \in (0, 1)$ ,  $k = 8 \log \delta^{-1}$  and  $N = \frac{n}{8 \log \delta^{-1}}$ .

Then with probability at least  $1 - \delta$ ,

$$\mathbb{E}X \leq \text{median} \left( \frac{1}{N} \sum_{s=1}^N X_s, \dots, \frac{1}{N} \sum_{s=(k-1)N+1}^{kN} X_s \right) + 8 \sqrt{\frac{8 \log(\delta^{-1})}{n}}.$$

This suggests LT-UCB, Bubeck, Cesa-Bianchi and Lugosi [2012]:

$$I_t \in \operatorname{argmax}_{1 \leq i \leq d} \operatorname{median} \left( \frac{1}{N_{i,t}} \sum_{s=1}^{N_{i,t}} X_{i,s}, \dots, \frac{1}{N_{i,t}} \sum_{s=(k_t-1)N_{i,t}+1}^{k_t N_{i,t}} X_{i,s} \right) \\ + 32 \sqrt{\frac{\log t}{T_i(t-1)}},$$

with  $k_t = 16 \log t$  and  $N_{i,t} = \frac{T_i(t-1)}{16 \log t}$ . The following regret bound can be proved for any set of distributions with variance bounded by 1:

$$R_n \leq \sum_{i: \Delta_i > 0} \frac{200 \log n}{\Delta_i}.$$

This suggests LT-UCB, Bubeck, Cesa-Bianchi and Lugosi [2012]:

$$I_t \in \operatorname{argmax}_{1 \leq i \leq d} \operatorname{median} \left( \frac{1}{N_{i,t}} \sum_{s=1}^{N_{i,t}} X_{i,s}, \dots, \frac{1}{N_{i,t}} \sum_{s=(k_t-1)N_{i,t}+1}^{k_t N_{i,t}} X_{i,s} \right) + 32 \sqrt{\frac{\log t}{T_i(t-1)}},$$

with  $k_t = 16 \log t$  and  $N_{i,t} = \frac{T_i(t-1)}{16 \log t}$ . The following regret bound can be proved for any set of distributions with variance bounded by 1:

$$R_n \leq \sum_{i: \Delta_i > 0} \frac{200 \log n}{\Delta_i}.$$

## Assumption

The sequence  $(X_{i,t})_{t \geq 1}$  forms an *aperiodic irreducible finite-state Markov chain* with unknown transition matrix  $P_i$ .

Again in this framework it is possible to design a UCB strategy with logarithmic regret (Tekin and Liu, [2011]), using the following result:

## Theorem (Lezaud [1998])

Let  $X_1, \dots, X_t$  be an *aperiodic irreducible finite-state Markov chain* with transition matrix  $P$ . Let  $\lambda_2$  be the *second largest eigenvalue* of the multiplicative symmetrization of  $P$  and  $\epsilon = 1 - \lambda_2$ . Let  $\mu$  be the *expectation of  $X_1$  under the stationary distribution*. There exists  $C > 0$  such that for any  $\gamma \in (0, 1]$ ,

$$\mathbb{P} \left( \frac{1}{t} \sum_{s=1}^t X_s \geq \mu + \gamma \right) \leq C \exp \left( -\frac{t\gamma^2\epsilon}{28} \right).$$

## Assumption

The sequence  $(X_{i,t})_{t \geq 1}$  forms an *aperiodic irreducible finite-state Markov chain* with unknown transition matrix  $P_i$ .

Again in this framework it is possible to design a UCB strategy with logarithmic regret (Tekin and Liu, [2011]), using the following result:

## Theorem (Lezaud [1998])

Let  $X_1, \dots, X_t$  be an *aperiodic irreducible finite-state Markov chain* with transition matrix  $P$ . Let  $\lambda_2$  be the *second largest eigenvalue* of the multiplicative symmetrization of  $P$  and  $\epsilon = 1 - \lambda_2$ . Let  $\mu$  be the *expectation of  $X_1$  under the stationary distribution*. There exists  $C > 0$  such that for any  $\gamma \in (0, 1]$ ,

$$\mathbb{P} \left( \frac{1}{t} \sum_{s=1}^t X_s \geq \mu + \gamma \right) \leq C \exp \left( -\frac{t\gamma^2\epsilon}{28} \right).$$

## Assumption

The sequence  $(X_{i,t})_{t \geq 1}$  forms an *aperiodic irreducible finite-state Markov chain* with unknown transition matrix  $P_i$ .

Again in this framework it is possible to design a UCB strategy with logarithmic regret (Tekin and Liu, [2011]), using the following result:

## Theorem (Lezaud [1998])

Let  $X_1, \dots, X_t$  be an *aperiodic irreducible finite-state Markov chain* with transition matrix  $P$ . Let  $\lambda_2$  be the *second largest eigenvalue* of the multiplicative symmetrization of  $P$  and  $\epsilon = 1 - \lambda_2$ . Let  $\mu$  be the *expectation of  $X_1$  under the stationary distribution*. There exists  $C > 0$  such that for any  $\gamma \in (0, 1]$ ,

$$\mathbb{P} \left( \frac{1}{t} \sum_{s=1}^t X_s \geq \mu + \gamma \right) \leq C \exp \left( -\frac{t\gamma^2\epsilon}{28} \right).$$

# Online Lipschitz and Stochastic Optimization

Stochastic multi-armed bandit where  $\{1, \dots, K\}$  is replaced by  $\mathcal{X}$ .  
At time  $t$ , select  $x_t \in \mathcal{X}$ , then receive a random variable  $r_t \in [0, 1]$  such that  $\mathbb{E}[r_t | x_t] = f(x_t)$ .

## Assumption

$\mathcal{X}$  is equipped with a symmetric function  $\rho : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$  such that  $\rho(x, x) = 0$ .  $f$  is Lipschitz with respect to  $\rho$ , that is

$$|f(x) - f(y)| \leq \rho(x, y), \forall x, y \in \mathcal{X}.$$

$$R_n = nf^* - \mathbb{E} \sum_{t=1}^n f(x_t),$$

where  $f^* = \sup_{x \in \mathcal{X}} f(x)$ .

# Online Lipschitz and Stochastic Optimization

Stochastic multi-armed bandit where  $\{1, \dots, K\}$  is replaced by  $\mathcal{X}$ .  
At time  $t$ , select  $x_t \in \mathcal{X}$ , then receive a random variable  $r_t \in [0, 1]$  such that  $\mathbb{E}[r_t | x_t] = f(x_t)$ .

## Assumption

$\mathcal{X}$  is equipped with a symmetric function  $\rho : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$  such that  $\rho(x, x) = 0$ .  $f$  is Lipschitz with respect to  $\rho$ , that is

$$|f(x) - f(y)| \leq \rho(x, y), \forall x, y \in \mathcal{X}.$$

$$R_n = nf^* - \mathbb{E} \sum_{t=1}^n f(x_t),$$

where  $f^* = \sup_{x \in \mathcal{X}} f(x)$ .

# Online Lipschitz and Stochastic Optimization

Stochastic multi-armed bandit where  $\{1, \dots, K\}$  is replaced by  $\mathcal{X}$ .  
At time  $t$ , select  $x_t \in \mathcal{X}$ , then receive a random variable  $r_t \in [0, 1]$   
such that  $\mathbb{E}[r_t | x_t] = f(x_t)$ .

## Assumption

$\mathcal{X}$  is equipped with a symmetric function  $\rho : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$  such  
that  $\rho(x, x) = 0$ .  $f$  is Lipschitz with respect to  $\rho$ , that is

$$|f(x) - f(y)| \leq \rho(x, y), \forall x, y \in \mathcal{X}.$$

$$R_n = nf^* - \mathbb{E} \sum_{t=1}^n f(x_t),$$

where  $f^* = \sup_{x \in \mathcal{X}} f(x)$ .

Stochastic multi-armed bandit where  $\{1, \dots, K\}$  is replaced by  $\mathcal{X}$ .  
At time  $t$ , select  $x_t \in \mathcal{X}$ , then receive a random variable  $r_t \in [0, 1]$  such that  $\mathbb{E}[r_t | x_t] = f(x_t)$ .

## Assumption

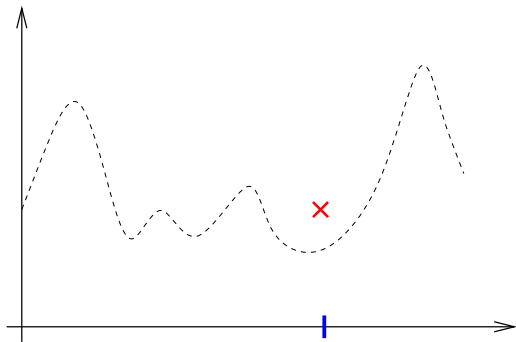
$\mathcal{X}$  is equipped with a symmetric function  $\rho : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$  such that  $\rho(x, x) = 0$ .  $f$  is Lipschitz with respect to  $\rho$ , that is

$$|f(x) - f(y)| \leq \rho(x, y), \forall x, y \in \mathcal{X}.$$

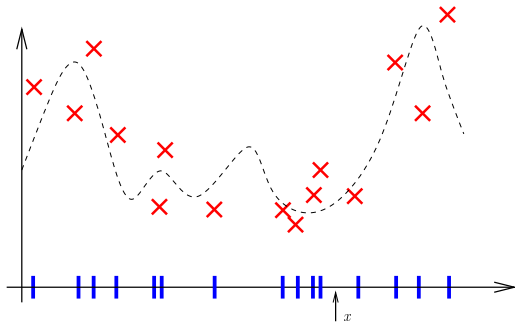
$$R_n = nf^* - \mathbb{E} \sum_{t=1}^n f(x_t),$$

where  $f^* = \sup_{x \in \mathcal{X}} f(x)$ .

# Example in 1d

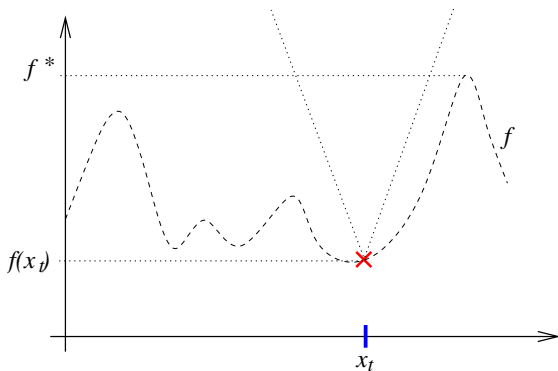


# Where should one sample next?



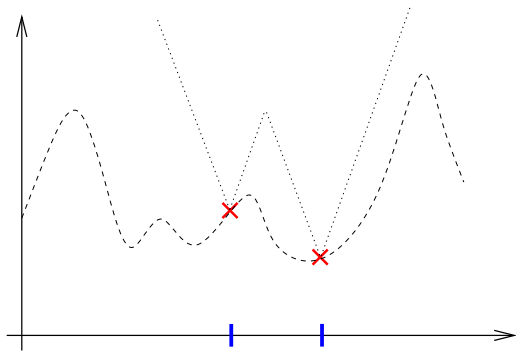
How to define a high probability upper bound at any state  $x$ ?

# Noiseless case, $r_t = f(x_t)$



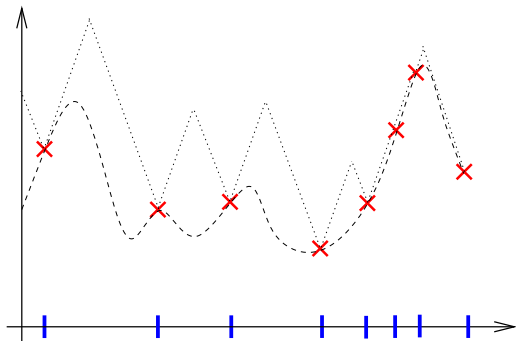
Lipschitz property  $\rightarrow$  the evaluation of  $f$  at  $x_t$  provides a first upper-bound on  $f$ .

Noiseless case,  $r_t = f(x_t)$

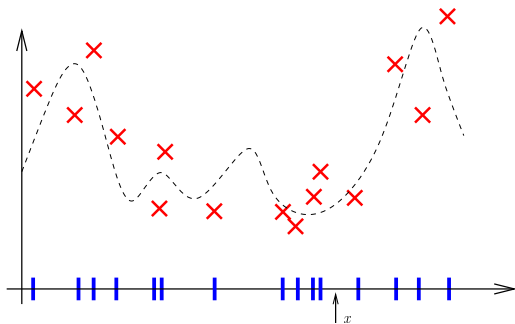


New point  $\rightarrow$  refined upper-bound on  $f$ .

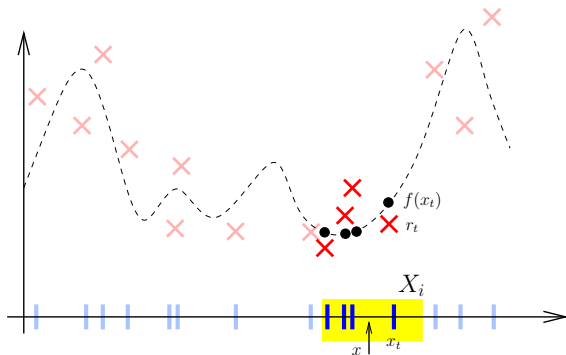
Noiseless case,  $r_t = f(x_t)$



# Back to the noisy case



# UCB in a given domain

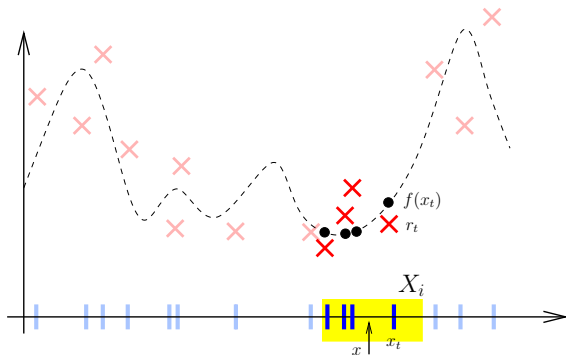


For a **fixed domain**  $X_i \ni x$  containing  $n_i$  points  $\{x_t\} \in X_i$ , we have that  $\sum_{t=1}^{n_i} r_t - f(x_t)$  is a **martingale**. Thus by **Azuma's inequality**,

$$\frac{1}{n_i} \sum_{t=1}^{n_i} r_t + \sqrt{\frac{\log 1/\delta}{2n_i}} \geq \frac{1}{n_i} \sum_{t=1}^{n_i} f(x_t) \geq f(x) - \text{diam}(X_i),$$

since  $f$  is Lipschitz (where  $\text{diam}(X_i) = \sup_{x,y \in X_i} \rho(x,y)$ ).

# UCB in a given domain

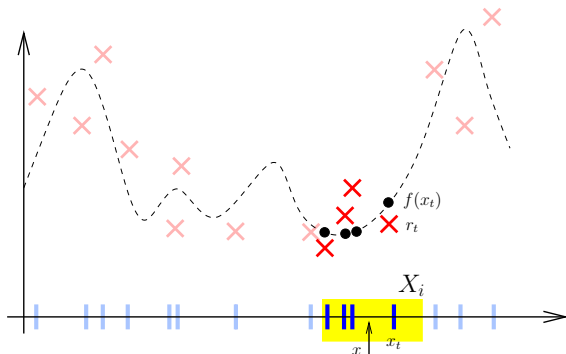


For a **fixed domain**  $X_i \ni x$  containing  $n_i$  points  $\{x_t\} \in X_i$ , we have that  $\sum_{t=1}^{n_i} r_t - f(x_t)$  is a **martingale**. Thus by **Azuma's inequality**,

$$\frac{1}{n_i} \sum_{t=1}^{n_i} r_t + \sqrt{\frac{\log 1/\delta}{2n_i}} \geq \frac{1}{n_i} \sum_{t=1}^{n_i} f(x_t) \geq f(x) - \text{diam}(X_i),$$

since  $f$  is Lipschitz (where  $\text{diam}(X_i) = \sup_{x,y \in X_i} \rho(x,y)$ ).

# UCB in a given domain

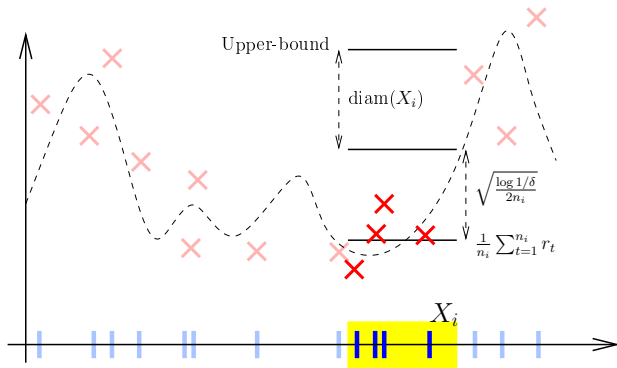


For a **fixed domain**  $X_i \ni x$  containing  $n_i$  points  $\{x_t\} \in X_i$ , we have that  $\sum_{t=1}^{n_i} r_t - f(x_t)$  is a **martingale**. Thus by **Azuma's inequality**,

$$\frac{1}{n_i} \sum_{t=1}^{n_i} r_t + \sqrt{\frac{\log 1/\delta}{2n_i}} \geq \frac{1}{n_i} \sum_{t=1}^{n_i} f(x_t) \geq f(x) - \text{diam}(X_i),$$

since  $f$  is Lipschitz (where  $\text{diam}(X_i) = \sup_{x,y \in X_i} \rho(x,y)$ ).

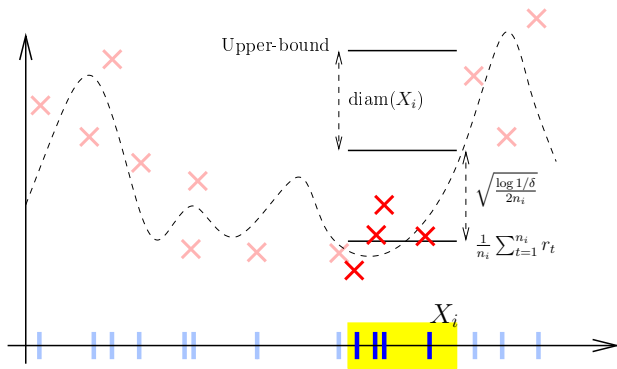
# High probability upper bound



w.p.  $1 - \delta$ , 
$$\frac{1}{n_i} \sum_{t=1}^{n_i} r_t + \sqrt{\frac{\log 1/\delta}{2n_i}} + \text{diam}(X_i) \geq \sup_{x \in X_i} f(x).$$

Tradeoff between number of points in a domain and size of the domain.  
By considering several domains we can derive a tighter upper bound.

# High probability upper bound

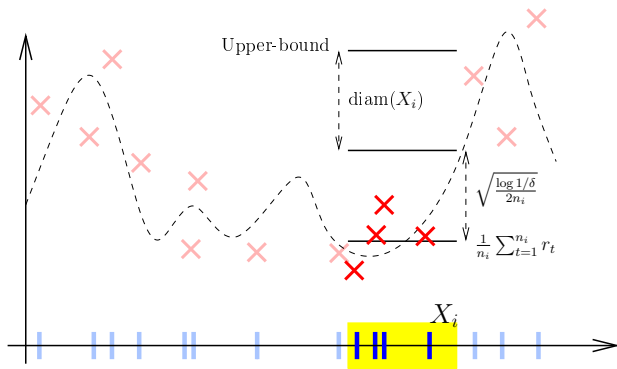


w.p.  $1 - \delta$ ,

$$\frac{1}{n_i} \sum_{t=1}^{n_i} r_t + \sqrt{\frac{\log 1/\delta}{2n_i}} + \text{diam}(X_i) \geq \sup_{x \in X_i} f(x).$$

Tradeoff between number of points in a domain and size of the domain.  
By considering several domains we can derive a tighter upper bound.

# High probability upper bound

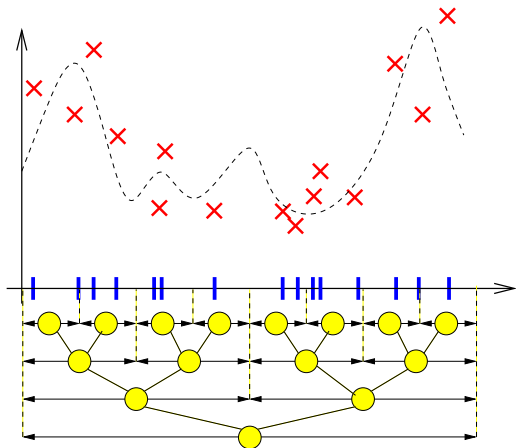


$$\text{w.p. } 1 - \delta, \quad \frac{1}{n_i} \sum_{t=1}^{n_i} r_t + \sqrt{\frac{\log 1/\delta}{2n_i}} + \text{diam}(X_i) \geq \sup_{x \in X_i} f(x).$$

Tradeoff between number of points in a domain and size of the domain.  
By considering several domains we can derive a tighter upper bound.

# A hierarchical decomposition

Use a tree of partitions at all scales:

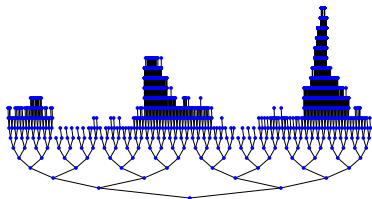
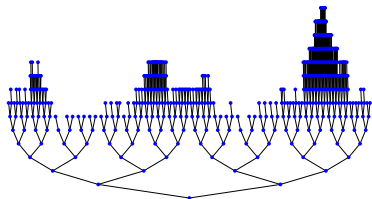
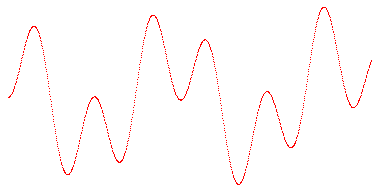
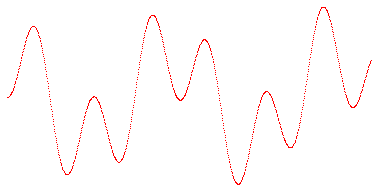


$$B_i(t) \stackrel{\text{def}}{=} \min \left\{ \hat{\mu}_i(t) + \sqrt{\frac{2 \log(t)}{T_i(t)}} + \text{diam}(i), \max_{j \in \mathcal{C}(i)} B_j(t) \right\}$$



# Example in 1d

$r_t \sim \mathcal{B}(f(x_t))$  a Bernoulli distribution with parameter  $f(x_t)$



Resulting tree at time  $n = 1000$  and at  $n = 10000$ .

The **near-optimality dimension**  $d$  of  $f$  is defined as follows: Let

$$\mathcal{X}_\epsilon \stackrel{\text{def}}{=} \{x \in \mathcal{X}, f(x) \geq f^* - \epsilon\}$$

be the set of  $\epsilon$ -optimal points. Then  $\mathcal{X}_\epsilon$  can be covered by  $O(\epsilon^{-d})$  balls of radius  $\epsilon$ . A similar notion was introduced in [Kleinberg, Slivkins, Upfal, 2008].

Theorem (Bubeck, Munos, Stoltz, Szepesvári, 2008)

*HOO satisfies:*

$$R_n = \tilde{O}\left(n^{\frac{d+1}{d+2}}\right).$$

The **near-optimality dimension**  $d$  of  $f$  is defined as follows: Let

$$\mathcal{X}_\epsilon \stackrel{\text{def}}{=} \{x \in \mathcal{X}, f(x) \geq f^* - \epsilon\}$$

be the set of  **$\epsilon$ -optimal points**. Then  $\mathcal{X}_\epsilon$  can be **covered** by  $O(\epsilon^{-d})$  balls of radius  $\epsilon$ . A similar notion was introduced in [Kleinberg, Slivkins, Uppal, 2008].

Theorem (Bubeck, Munos, Stoltz, Szepesvári, 2008)

*HOO satisfies:*

$$R_n = \tilde{O}\left(n^{\frac{d+1}{d+2}}\right).$$

The **near-optimality dimension**  $d$  of  $f$  is defined as follows: Let

$$\mathcal{X}_\epsilon \stackrel{\text{def}}{=} \{x \in \mathcal{X}, f(x) \geq f^* - \epsilon\}$$

be the set of  $\epsilon$ -optimal points. Then  $\mathcal{X}_\epsilon$  can be covered by  $O(\epsilon^{-d})$  balls of radius  $\epsilon$ . A similar notion was introduced in [Kleinberg, Slivkins, Upfal, 2008].

Theorem (Bubeck, Munos, Stoltz, Szepesvári, 2008)

*HOO satisfies:*

$$R_n = \tilde{O}\left(n^{\frac{d+1}{d+2}}\right).$$

The **near-optimality dimension**  $d$  of  $f$  is defined as follows: Let

$$\mathcal{X}_\epsilon \stackrel{\text{def}}{=} \{x \in \mathcal{X}, f(x) \geq f^* - \epsilon\}$$

be the set of  $\epsilon$ -optimal points. Then  $\mathcal{X}_\epsilon$  can be covered by  $O(\epsilon^{-d})$  balls of radius  $\epsilon$ . A similar notion was introduced in [Kleinberg, Slivkins, Upfal, 2008].

Theorem (Bubeck, Munos, Stoltz, Szepesvári, 2008)

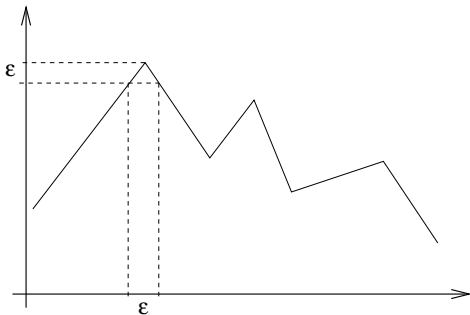
*HOO* satisfies:

$$R_n = \tilde{O}\left(n^{\frac{d+1}{d+2}}\right).$$

## Example 1:

Assume the function is locally peaky around its maximum:

$$f(x^*) - f(x) = \Theta(\|x^* - x\|).$$

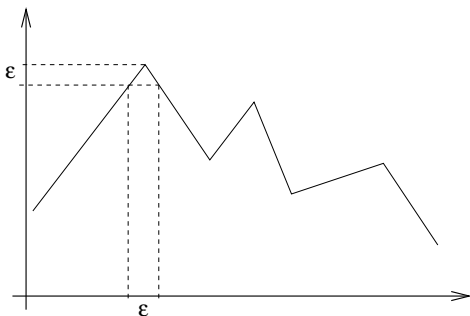


It takes  $O(\epsilon^0)$  balls of radius  $\epsilon$  to cover  $X_\epsilon$  with  $\rho(x, y) = \|x - y\|$ .  
Thus  $d = 0$  and the regret is  $\tilde{O}(\sqrt{n})$ .

## Example 1:

Assume the function is locally peaky around its maximum:

$$f(x^*) - f(x) = \Theta(\|x^* - x\|).$$

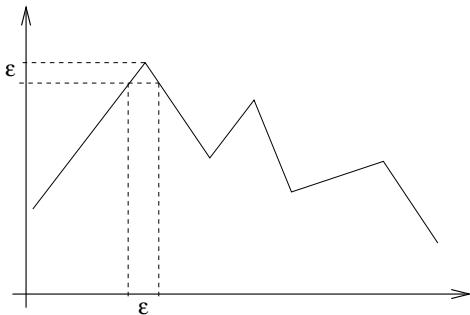


It takes  $O(\epsilon^0)$  balls of radius  $\epsilon$  to cover  $X_\epsilon$  with  $\rho(x, y) = \|x - y\|$ .  
Thus  $d = 0$  and the regret is  $\tilde{O}(\sqrt{n})$ .

## Example 1:

Assume the function is locally peaky around its maximum:

$$f(x^*) - f(x) = \Theta(\|x^* - x\|).$$

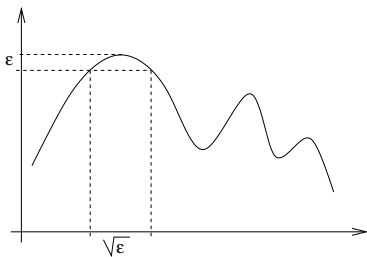


It takes  $O(\epsilon^0)$  balls of radius  $\epsilon$  to cover  $X_\epsilon$  with  $\rho(x, y) = \|x - y\|$ .  
Thus  $d = 0$  and the regret is  $\tilde{O}(\sqrt{n})$ .

## Example 2:

Assume the function is locally quadratic around its maximum:

$$f(x^*) - f(x) = \Theta(\|x^* - x\|^2).$$

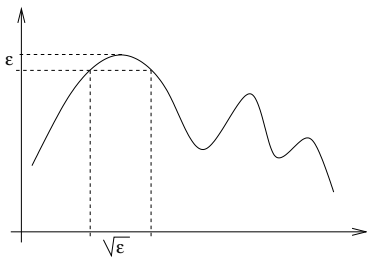


- For  $\rho(x, y) = \|x - y\|$ , it takes  $O(\epsilon^{-D/2})$  balls of radius  $\epsilon$  to cover  $X_\epsilon$ . Thus  $d = D/2$  and  $R_n = \tilde{O}(n^{\frac{D+2}{D+4}})$ .
- For  $\rho(x, y) = \|x - y\|^2$ , it takes  $O(\epsilon^0)$   $\rho$ -balls of radius  $\epsilon$  to cover  $X_\epsilon$ . Thus  $d = 0$  and  $R_n = \tilde{O}(\sqrt{n})$ .

## Example 2:

Assume the function is locally quadratic around its maximum:

$$f(x^*) - f(x) = \Theta(\|x^* - x\|^2).$$

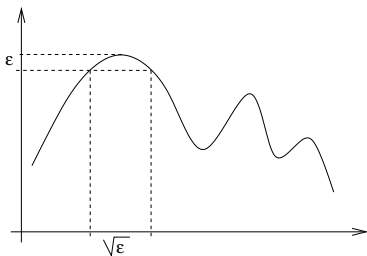


- For  $\rho(x, y) = \|x - y\|$ , it takes  $O(\epsilon^{-D/2})$  balls of radius  $\epsilon$  to cover  $X_\epsilon$ . Thus  $d = D/2$  and  $R_n = \tilde{O}(n^{\frac{D+2}{D+4}})$ .
- For  $\rho(x, y) = \|x - y\|^2$ , it takes  $O(\epsilon^0)$   $\rho$ -balls of radius  $\epsilon$  to cover  $X_\epsilon$ . Thus  $d = 0$  and  $R_n = \tilde{O}(\sqrt{n})$ .

## Example 2:

Assume the function is locally quadratic around its maximum:

$$f(x^*) - f(x) = \Theta(\|x^* - x\|^2).$$



- For  $\rho(x, y) = \|x - y\|$ , it takes  $O(\epsilon^{-D/2})$  balls of radius  $\epsilon$  to cover  $X_\epsilon$ . Thus  $d = D/2$  and  $R_n = \tilde{O}(n^{\frac{D+2}{D+4}})$ .
- For  $\rho(x, y) = \|x - y\|^2$ , it takes  $O(\epsilon^0)$   $\rho$ -balls of radius  $\epsilon$  to cover  $X_\epsilon$ . Thus  $d = 0$  and  $R_n = \tilde{O}(\sqrt{n})$ .

## Example

$\mathcal{X} = [0, 1]^D$ ,  $\alpha \geq 0$  and mean-payoff function  $f$  locally " $\alpha$ -smooth" around (any of) its maximum  $x^*$  (in finite number):

$$f(x^*) - f(x) = \Theta(\|x - x^*\|^\alpha) \text{ as } x \rightarrow x^*.$$

### Theorem

Assume that we run HOO using  $\rho(x, y) = \|x - y\|^\beta$ .

- Known smoothness:  $\beta = \alpha$ .  $R_n = \tilde{O}(\sqrt{n})$ , i.e., the rate is independent of the dimension  $D$ .
- Smoothness underestimated:  $\beta < \alpha$ .  
 $R_n = \tilde{O}(n^{(d+1)/(d+2)})$  where  $d = D \left( \frac{1}{\beta} - \frac{1}{\alpha} \right)$ .
- Smoothness overestimated:  $\beta > \alpha$ . No guarantee. Note: UCT corresponds to  $\beta = +\infty$ .

# Example

$\mathcal{X} = [0, 1]^D$ ,  $\alpha \geq 0$  and mean-payoff function  $f$  locally " $\alpha$ -smooth" around (any of) its maximum  $x^*$  (in finite number):

$$f(x^*) - f(x) = \Theta(\|x - x^*\|^\alpha) \text{ as } x \rightarrow x^*.$$

## Theorem

Assume that we run *HOO* using  $\rho(x, y) = \|x - y\|^\beta$ .

- **Known smoothness:**  $\beta = \alpha$ .  $R_n = \tilde{O}(\sqrt{n})$ , i.e., the rate is independent of the dimension  $D$ .
- **Smoothness underestimated:**  $\beta < \alpha$ .  
 $R_n = \tilde{O}(n^{(d+1)/(d+2)})$  where  $d = D \left( \frac{1}{\beta} - \frac{1}{\alpha} \right)$ .
- **Smoothness overestimated:**  $\beta > \alpha$ . No guarantee. Note: *UCT* corresponds to  $\beta = +\infty$ .

# Example

$\mathcal{X} = [0, 1]^D$ ,  $\alpha \geq 0$  and mean-payoff function  $f$  locally " $\alpha$ -smooth" around (any of) its maximum  $x^*$  (in finite number):

$$f(x^*) - f(x) = \Theta(\|x - x^*\|^\alpha) \text{ as } x \rightarrow x^*.$$

## Theorem

Assume that we run *HOO* using  $\rho(x, y) = \|x - y\|^\beta$ .

- **Known smoothness:**  $\beta = \alpha$ .  $R_n = \tilde{O}(\sqrt{n})$ , i.e., the rate is independent of the dimension  $D$ .
- **Smoothness underestimated:**  $\beta < \alpha$ .  
 $R_n = \tilde{O}(n^{(d+1)/(d+2)})$  where  $d = D \left( \frac{1}{\beta} - \frac{1}{\alpha} \right)$ .
- **Smoothness overestimated:**  $\beta > \alpha$ . No guarantee. Note: *UCT* corresponds to  $\beta = +\infty$ .

# Example

$\mathcal{X} = [0, 1]^D$ ,  $\alpha \geq 0$  and mean-payoff function  $f$  locally " $\alpha$ -smooth" around (any of) its maximum  $x^*$  (in finite number):

$$f(x^*) - f(x) = \Theta(\|x - x^*\|^\alpha) \text{ as } x \rightarrow x^*.$$

## Theorem

Assume that we run *HOO* using  $\rho(x, y) = \|x - y\|^\beta$ .

- **Known smoothness:**  $\beta = \alpha$ .  $R_n = \tilde{O}(\sqrt{n})$ , i.e., the rate is independent of the dimension  $D$ .
- **Smoothness underestimated:**  $\beta < \alpha$ .  
 $R_n = \tilde{O}(n^{(d+1)/(d+2)})$  where  $d = D \left( \frac{1}{\beta} - \frac{1}{\alpha} \right)$ .
- **Smoothness overestimated:**  $\beta > \alpha$ . No guarantee. Note: *UCT* corresponds to  $\beta = +\infty$ .

# Example

$\mathcal{X} = [0, 1]^D$ ,  $\alpha \geq 0$  and mean-payoff function  $f$  locally " $\alpha$ -smooth" around (any of) its maximum  $x^*$  (in finite number):

$$f(x^*) - f(x) = \Theta(\|x - x^*\|^\alpha) \text{ as } x \rightarrow x^*.$$

## Theorem

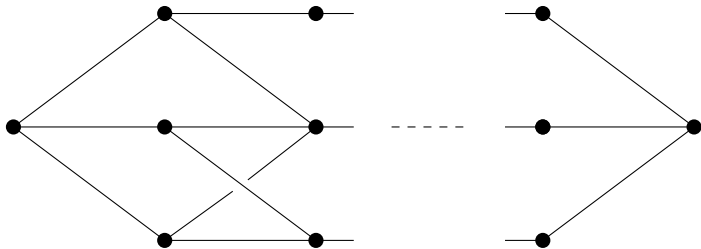
Assume that we run *HOO* using  $\rho(x, y) = \|x - y\|^\beta$ .

- **Known smoothness:**  $\beta = \alpha$ .  $R_n = \tilde{O}(\sqrt{n})$ , i.e., the rate is independent of the dimension  $D$ .
- **Smoothness underestimated:**  $\beta < \alpha$ .  
 $R_n = \tilde{O}(n^{(d+1)/(d+2)})$  where  $d = D \left( \frac{1}{\beta} - \frac{1}{\alpha} \right)$ .
- **Smoothness overestimated:**  $\beta > \alpha$ . No guarantee. Note: *UCT* corresponds to  $\beta = +\infty$ .



# Combinatorial prediction game

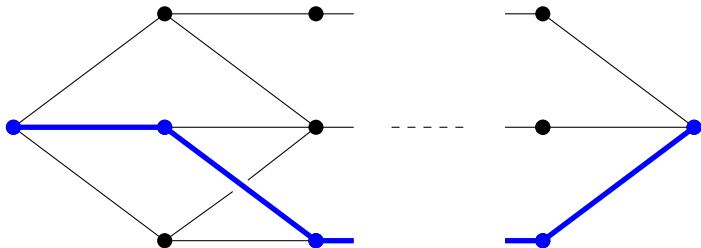
Adversary



Player

# Combinatorial prediction game

Adversary

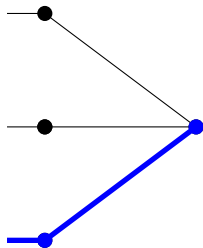
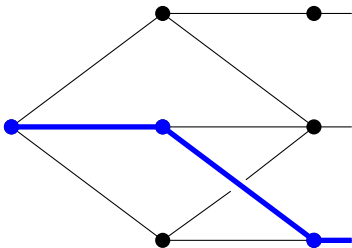


Player →



# Combinatorial prediction game

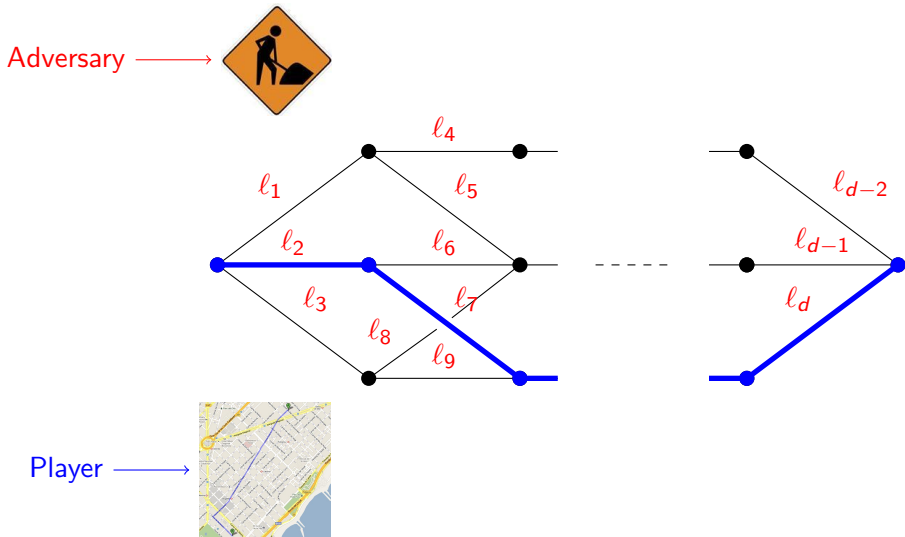
Adversary



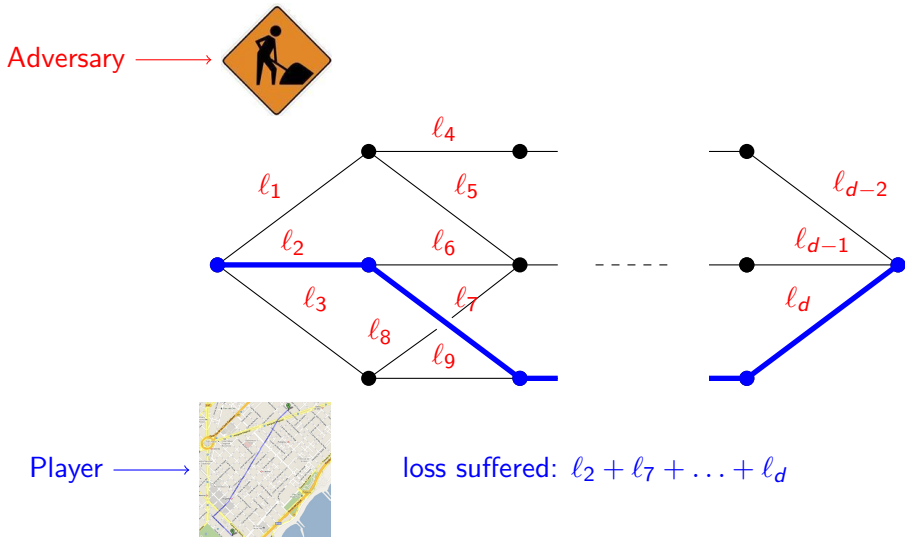
Player



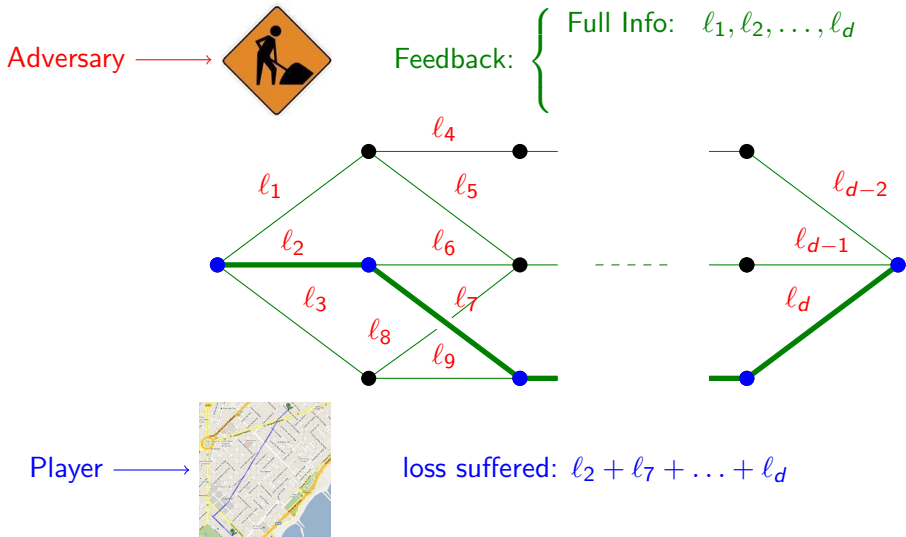
# Combinatorial prediction game



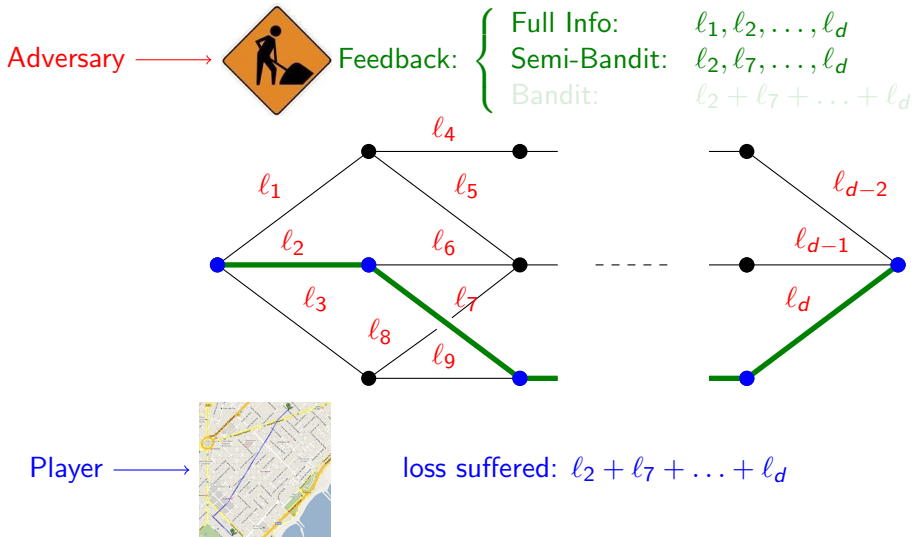
# Combinatorial prediction game



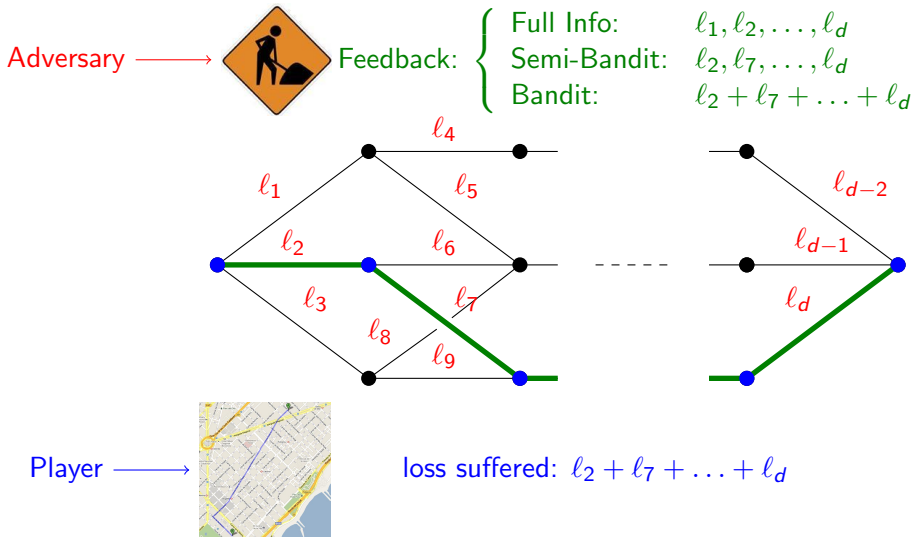
# Combinatorial prediction game



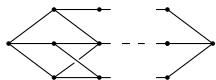
# Combinatorial prediction game



# Combinatorial prediction game



# Notation



$$\longleftrightarrow \mathcal{S} \subset \{0, 1\}^d$$



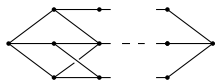
$$\longleftrightarrow l_t \in \mathbb{R}_+^d$$



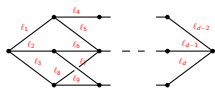
$$\longleftrightarrow V_t \in \mathcal{S}, \text{ loss suffered: } l_t^T V_t$$

$$R_n = \mathbb{E} \sum_{t=1}^n l_t^T V_t - \min_{u \in \mathcal{S}} \mathbb{E} \sum_{t=1}^n l_t^T u$$

# Notation



$$\longleftrightarrow \mathcal{S} \subset \{0, 1\}^d$$



$$\longleftrightarrow l_t \in \mathbb{R}_+^d$$

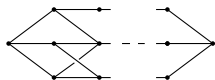


$$\longleftrightarrow V_t \in \mathcal{S}, \text{ loss suffered: } l_t^T V_t$$

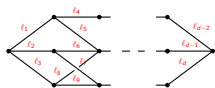
$$R_n = \mathbb{E} \sum_{t=1}^n l_t^T V_t - \min_{u \in \mathcal{S}} \mathbb{E} \sum_{t=1}^n l_t^T u$$



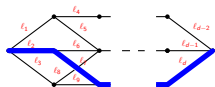
# Notation



$$\longleftrightarrow \mathcal{S} \subset \{0, 1\}^d$$



$$\longleftrightarrow l_t \in \mathbb{R}_+^d$$

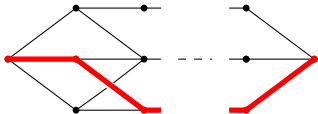


$$\longleftrightarrow V_t \in \mathcal{S}, \text{ loss suffered: } l_t^T V_t$$

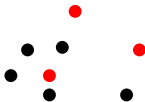
$$R_n = \mathbb{E} \sum_{t=1}^n l_t^T V_t - \min_{u \in \mathcal{S}} \mathbb{E} \sum_{t=1}^n l_t^T u$$

# Set of concepts $S \subset \{0, 1\}^d$

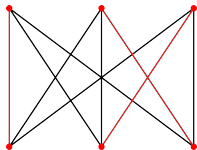
Paths



$k$ -sets



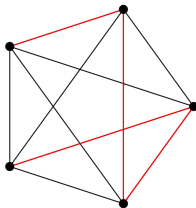
Matchings



$k$ -sized intervals



Spanning trees



Parallel bandits



$$V_t \sim p_t, \quad p_t \in \Delta(\mathcal{S})$$

Then, unbiased estimate  $\tilde{l}_t$  of the loss  $l_t$ :

- $\tilde{l}_t = l_t$  in the full information game,
- $\tilde{l}_{i,t} = \frac{l_{i,t}}{\sum_{V \in \mathcal{S}: V_i=1} p_t(V)} V_{i,t}$  in the semi-bandit game,
- $\tilde{l}_t = P_t^+ V_t V_t^T l_t$ , with  $P_t = \mathbb{E}_{V \sim p_t}(V V^T)$  in the bandit game.

$$V_t \sim p_t, \quad p_t \in \Delta(\mathcal{S})$$

Then, unbiased estimate  $\tilde{l}_t$  of the loss  $l_t$ :

- $\tilde{l}_t = l_t$  in the full information game,
- $\tilde{l}_{i,t} = \frac{l_{i,t}}{\sum_{V \in \mathcal{S}: V_i=1} p_t(V)} V_{i,t}$  in the semi-bandit game,
- $\tilde{l}_t = P_t^+ V_t V_t^T l_t$ , with  $P_t = \mathbb{E}_{V \sim p_t}(V V^T)$  in the bandit game.

$$V_t \sim p_t, \quad p_t \in \Delta(\mathcal{S})$$

Then, unbiased estimate  $\tilde{l}_t$  of the loss  $l_t$ :

- $\tilde{l}_t = l_t$  in the full information game,
- $\tilde{l}_{i,t} = \frac{l_{i,t}}{\sum_{V \in \mathcal{S}: V_i=1} p_t(V)} V_{i,t}$  in the semi-bandit game,
- $\tilde{l}_t = P_t^+ V_t V_t^T l_t$ , with  $P_t = \mathbb{E}_{V \sim p_t}(V V^T)$  in the bandit game.

$$V_t \sim p_t, \quad p_t \in \Delta(S)$$

Then, unbiased estimate  $\tilde{l}_t$  of the loss  $l_t$ :

- $\tilde{l}_t = l_t$  in the full information game,
- $\tilde{l}_{i,t} = \frac{l_{i,t}}{\sum_{V \in S: V_i=1} p_t(V)} V_{i,t}$  in the semi-bandit game,
- $\tilde{l}_t = P_t^+ V_t V_t^T l_t$ , with  $P_t = \mathbb{E}_{V \sim p_t}(V V^T)$  in the bandit game.

$$V_t \sim p_t, \quad p_t \in \Delta(S)$$

Then, unbiased estimate  $\tilde{l}_t$  of the loss  $l_t$ :

- $\tilde{l}_t = l_t$  in the full information game,
- $\tilde{l}_{i,t} = \frac{l_{i,t}}{\sum_{V \in S: V_i=1} p_t(V)} V_{i,t}$  in the semi-bandit game,
- $\tilde{l}_t = P_t^+ V_t V_t^T l_t$ , with  $P_t = \mathbb{E}_{V \sim p_t}(V V^T)$  in the bandit game.

# Loss assumptions

## Definition ( $L_\infty$ )

We say that the adversary satisfies the  $L_\infty$  **assumption**: if  $\|\ell_t\|_\infty \leq 1$  for all  $t = 1, \dots, n$ .

## Definition ( $L_2$ )

We say that the adversary satisfies the  $L_2$  **assumption**: if  $\ell_t^T v \leq 1$  for all  $t = 1, \dots, n$  and  $v \in \mathcal{S}$ .

## Definition ( $L_\infty$ )

We say that the adversary satisfies the  $L_\infty$  **assumption**: if  $\|\ell_t\|_\infty \leq 1$  for all  $t = 1, \dots, n$ .

## Definition ( $L_2$ )

We say that the adversary satisfies the  $L_2$  **assumption**: if  $\ell_t^T v \leq 1$  for all  $t = 1, \dots, n$  and  $v \in \mathcal{S}$ .

## Expanded Exponentially weighted average forecaster (Exp2)

$$p_t(v) = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_s^T v\right)}{\sum_{u \in \mathcal{S}} \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_s^T u\right)}$$

- In the full information game, against  $L_2$  adversaries, we have (for some  $\eta$ )

$$R_n \leq \sqrt{2dn},$$

which is the optimal rate, Dani, Hayes and Kakade [2008].

- Thus against  $L_\infty$  adversaries we have

$$R_n \leq d^{3/2} \sqrt{2n}.$$

But this is suboptimal, Koolen, Warmuth and Kivinen [2010].

- Audibert, Bubeck and Lugosi [2011] showed that, for any  $\eta$ , there exists a subset  $S \subset \{0, 1\}^d$  and an  $L_\infty$  adversary such that:

$$R_n \geq 0.02 d^{3/2} \sqrt{n}.$$

## Expanded Exponentially weighted average forecaster (Exp2)

$$p_t(v) = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_s^T v\right)}{\sum_{u \in \mathcal{S}} \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_s^T u\right)}$$

- In the **full information** game, against  $L_2$  adversaries, we have (for some  $\eta$ )

$$R_n \leq \sqrt{2dn},$$

which is the **optimal** rate, Dani, Hayes and Kakade [2008].

- Thus against  $L_\infty$  adversaries we have

$$R_n \leq d^{3/2} \sqrt{2n}.$$

But this is **suboptimal**, Koolen, Warmuth and Kivinen [2010].

- Audibert, Bubeck and Lugosi [2011] showed that, for any  $\eta$ , there exists a subset  $S \subset \{0, 1\}^d$  and an  $L_\infty$  adversary such that:

$$R_n \geq 0.02 d^{3/2} \sqrt{n}.$$

## Expanded Exponentially weighted average forecaster (Exp2)

$$p_t(v) = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_s^T v\right)}{\sum_{u \in \mathcal{S}} \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_s^T u\right)}$$

- In the **full information** game, against  $L_2$  adversaries, we have (for some  $\eta$ )

$$R_n \leq \sqrt{2dn},$$

which is the **optimal** rate, Dani, Hayes and Kakade [2008].

- Thus against  $L_\infty$  adversaries we have

$$R_n \leq d^{3/2} \sqrt{2n}.$$

But this is **suboptimal**, Koolen, Warmuth and Kivinen [2010].

- Audibert, Bubeck and Lugosi [2011] showed that, for any  $\eta$ , there exists a subset  $S \subset \{0, 1\}^d$  and an  $L_\infty$  adversary such that:

$$R_n \geq 0.02 d^{3/2} \sqrt{n}.$$

## Expanded Exponentially weighted average forecaster (Exp2)

$$p_t(v) = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_s^T v\right)}{\sum_{u \in \mathcal{S}} \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_s^T u\right)}$$

- In the **full information** game, against  $L_2$  adversaries, we have (for some  $\eta$ )

$$R_n \leq \sqrt{2dn},$$

which is the **optimal** rate, Dani, Hayes and Kakade [2008].

- Thus against  $L_\infty$  adversaries we have

$$R_n \leq d^{3/2} \sqrt{2n}.$$

But this is **suboptimal**, Koolen, Warmuth and Kivinen [2010].

- Audibert, Bubeck and Lugosi [2011] showed that, for any  $\eta$ , there exists a subset  $S \subset \{0, 1\}^d$  and an  $L_\infty$  adversary such that:

$$R_n \geq 0.02 d^{3/2} \sqrt{n}.$$

## Definition

Let  $\mathcal{D}$  be a **convex** subset of  $\mathbb{R}^d$  with nonempty interior  $\text{int}(\mathcal{D})$  and boundary  $\partial\mathcal{D}$ . We call **Legendre** any function  $F : \mathcal{D} \rightarrow \mathbb{R}$  such that

- $F$  is **strictly convex** and admits continuous first partial derivatives on  $\text{int}(\mathcal{D})$ ,
- For any  $u \in \partial\mathcal{D}$ , for any  $v \in \text{int}(\mathcal{D})$ , we have

$$\lim_{s \rightarrow 0, s > 0} (u - v)^T \nabla F((1 - s)u + sv) = +\infty.$$

## Definition

Let  $\mathcal{D}$  be a **convex** subset of  $\mathbb{R}^d$  with nonempty interior  $\text{int}(\mathcal{D})$  and boundary  $\partial\mathcal{D}$ . We call **Legendre** any function  $F : \mathcal{D} \rightarrow \mathbb{R}$  such that

- $F$  is **strictly convex** and admits continuous first partial derivatives on  $\text{int}(\mathcal{D})$ ,
- For any  $u \in \partial\mathcal{D}$ , for any  $v \in \text{int}(\mathcal{D})$ , we have

$$\lim_{s \rightarrow 0, s > 0} (u - v)^T \nabla F((1 - s)u + sv) = +\infty.$$

## Definition

Let  $\mathcal{D}$  be a **convex** subset of  $\mathbb{R}^d$  with nonempty interior  $\text{int}(\mathcal{D})$  and boundary  $\partial\mathcal{D}$ . We call **Legendre** any function  $F : \mathcal{D} \rightarrow \mathbb{R}$  such that

- $F$  is **strictly convex** and admits continuous first partial derivatives on  $\text{int}(\mathcal{D})$ ,
- For any  $u \in \partial\mathcal{D}$ , for any  $v \in \text{int}(\mathcal{D})$ , we have

$$\lim_{s \rightarrow 0, s > 0} (u - v)^T \nabla F((1 - s)u + sv) = +\infty.$$

## Definition

The **Bregman divergence**  $D_F : \mathcal{D} \times \text{int}(\mathcal{D})$  associated to a **Legendre** function  $F$  is defined by

$$D_F(u, v) = F(u) - F(v) - (u - v)^T \nabla F(v).$$

## Definition

The **Legendre transform** of  $F$  is defined by

$$F^*(u) = \sup_{x \in \mathcal{D}} x^T u - F(x).$$

Key property for Legendre functions:  $\nabla F^* = (\nabla F)^{-1}$ .

# Online Stochastic Mirror Descent (OSMD)

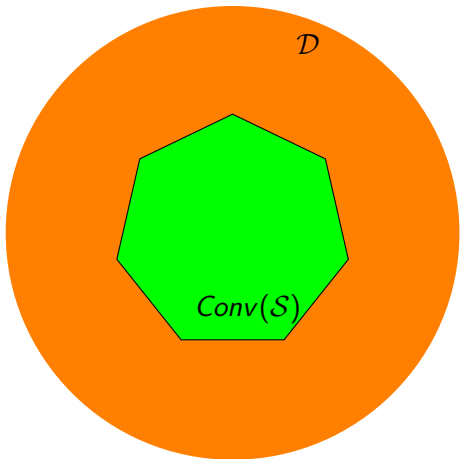
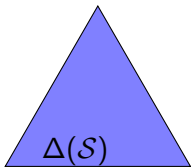
Parameter:  $F$  Legendre on  $\mathcal{D} \supset \text{Conv}(S)$

(1)  $w'_{t+1} \in \mathcal{D}$ :

$$w'_{t+1} = \nabla F^* \left( \nabla F(w_t) - \tilde{\ell}_t \right)$$

(2)  $w_{t+1} \in \underset{w \in \text{Conv}(S)}{\text{argmin}} D_F(w, w'_{t+1})$

(3)  $p_{t+1} \in \Delta(S) : w_{t+1} = \mathbb{E}_{V \sim p_{t+1}} V$



# Online Stochastic Mirror Descent (OSMD)

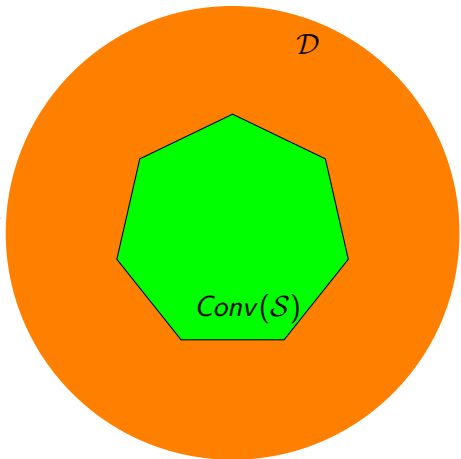
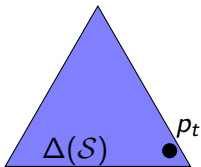
Parameter:  $F$  Legendre on  $\mathcal{D} \supset \text{Conv}(S)$

(1)  $w'_{t+1} \in \mathcal{D}$ :

$$w'_{t+1} = \nabla F^* \left( \nabla F(w_t) - \tilde{\ell}_t \right)$$

(2)  $w_{t+1} \in \underset{w \in \text{Conv}(S)}{\text{argmin}} D_F(w, w'_{t+1})$

(3)  $p_{t+1} \in \Delta(S) : w_{t+1} = \mathbb{E}_{V \sim p_{t+1}} V$



# Online Stochastic Mirror Descent (OSMD)

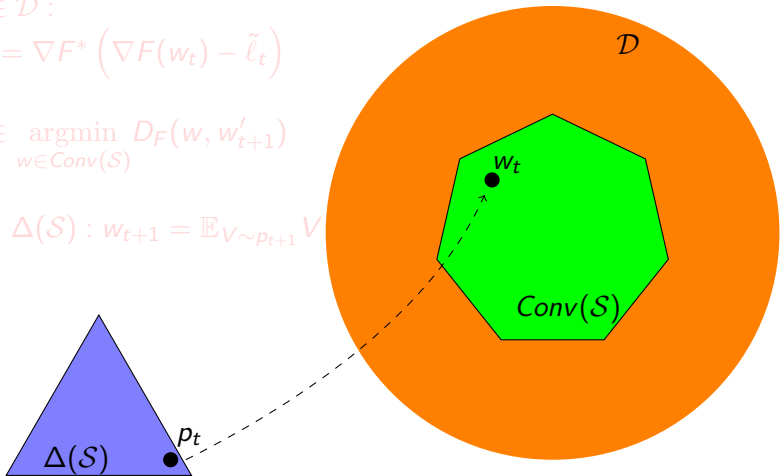
Parameter:  $F$  Legendre on  $\mathcal{D} \supset \text{Conv}(S)$

(1)  $w'_{t+1} \in \mathcal{D}$ :

$$w'_{t+1} = \nabla F^* \left( \nabla F(w_t) - \tilde{\ell}_t \right)$$

(2)  $w_{t+1} \in \underset{w \in \text{Conv}(S)}{\text{argmin}} D_F(w, w'_{t+1})$

(3)  $p_{t+1} \in \Delta(S)$ :  $w_{t+1} = \mathbb{E}_{V \sim p_{t+1}} V$



# Online Stochastic Mirror Descent (OSMD)

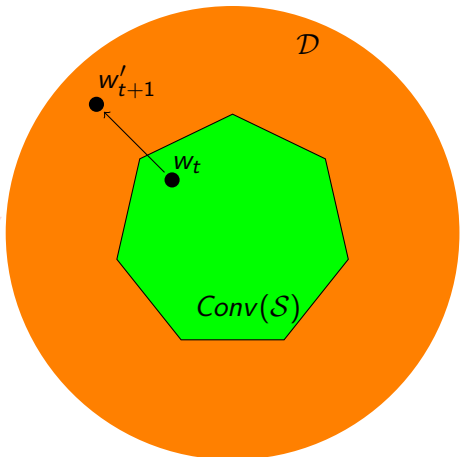
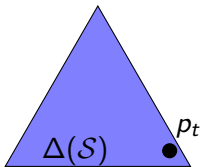
Parameter:  $F$  Legendre on  $\mathcal{D} \supset \text{Conv}(S)$

(1)  $w'_{t+1} \in \mathcal{D}$ :

$$w'_{t+1} = \nabla F^* \left( \nabla F(w_t) - \tilde{\ell}_t \right)$$

(2)  $w_{t+1} \in \underset{w \in \text{Conv}(S)}{\text{argmin}} D_F(w, w'_{t+1})$

(3)  $p_{t+1} \in \Delta(S) : w_{t+1} = \mathbb{E}_{V \sim p_{t+1}} V$



# Online Stochastic Mirror Descent (OSMD)

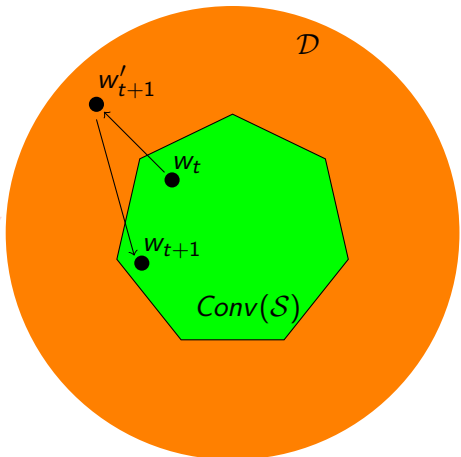
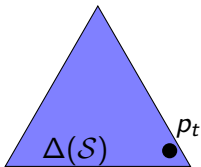
Parameter:  $F$  Legendre on  $\mathcal{D} \supset \text{Conv}(S)$

(1)  $w'_{t+1} \in \mathcal{D}$ :

$$w'_{t+1} = \nabla F^* \left( \nabla F(w_t) - \tilde{\ell}_t \right)$$

(2)  $w_{t+1} \in \underset{w \in \text{Conv}(S)}{\text{argmin}} D_F(w, w'_{t+1})$

(3)  $p_{t+1} \in \Delta(S) : w_{t+1} = \mathbb{E}_{V \sim p_{t+1}} V$



# Online Stochastic Mirror Descent (OSMD)

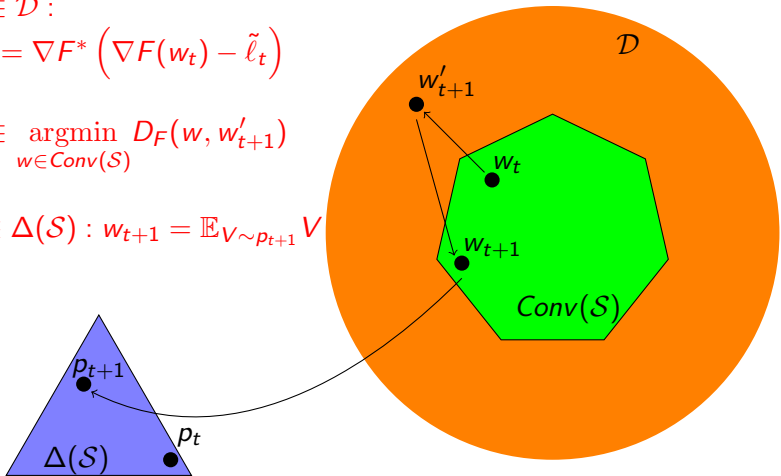
Parameter:  $F$  Legendre on  $\mathcal{D} \supset \text{Conv}(S)$

(1)  $w'_{t+1} \in \mathcal{D}$ :

$$w'_{t+1} = \nabla F^* \left( \nabla F(w_t) - \tilde{\ell}_t \right)$$

(2)  $w_{t+1} \in \underset{w \in \text{Conv}(S)}{\text{argmin}} D_F(w, w'_{t+1})$

(3)  $p_{t+1} \in \Delta(S)$ :  $w_{t+1} = \mathbb{E}_{V \sim p_{t+1}} V$



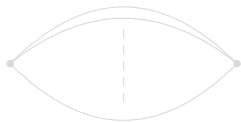
## Theorem

If  $F$  admits a *Hessian*  $\nabla^2 F$  always *invertible* then,

$$R_n \lesssim \text{diam}_{D_F}(\mathcal{S}) + \mathbb{E} \sum_{t=1}^n \tilde{\ell}_t^T (\nabla^2 F(w_t))^{-1} \tilde{\ell}_t.$$

# Different instances of OSMD: LinExp (Entropy Function)

$$\mathcal{D} = [0, +\infty)^d, F(x) = \frac{1}{\eta} \sum_{i=1}^d x_i \log x_i$$



Full Info: Hedge

Semi-Bandit=Bandit: Exp3

Auer et al. [2002]



Full Info: Component Hedge

Koolen, Warmuth and Kivinen [2010]

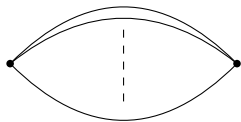
Semi-Bandit: MW

Kale, Reyzin and Schapire [2010]

Bandit: new algorithm

# Different instances of OSMD: LinExp (Entropy Function)

$$\mathcal{D} = [0, +\infty)^d, F(x) = \frac{1}{\eta} \sum_{i=1}^d x_i \log x_i$$



Full Info: Hedge

Semi-Bandit=Bandit: Exp3

Auer et al. [2002]



Full Info: Component Hedge

Koolen, Warmuth and Kivinen [2010]

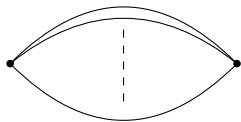
Semi-Bandit: MW

Kale, Reyzin and Schapire [2010]

Bandit: new algorithm

# Different instances of OSMD: LinExp (Entropy Function)

$$\mathcal{D} = [0, +\infty)^d, F(x) = \frac{1}{\eta} \sum_{i=1}^d x_i \log x_i$$



Full Info: Hedge

Semi-Bandit=Bandit: Exp3

Auer et al. [2002]



Full Info: Component Hedge

Koolen, Warmuth and Kivinen [2010]

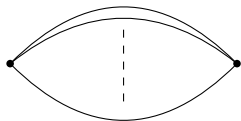
Semi-Bandit: MW

Kale, Reyzin and Schapire [2010]

Bandit: new algorithm

# Different instances of OSMD: LinExp (Entropy Function)

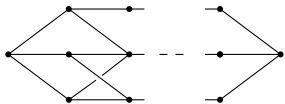
$$\mathcal{D} = [0, +\infty)^d, F(x) = \frac{1}{\eta} \sum_{i=1}^d x_i \log x_i$$



Full Info: Hedge

Semi-Bandit=Bandit: Exp3

Auer et al. [2002]



Full Info: Component Hedge

Koolen, Warmuth and Kivinen [2010]

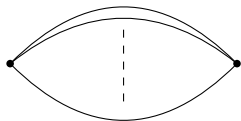
Semi-Bandit: MW

Kale, Reyzin and Schapire [2010]

Bandit: new algorithm

# Different instances of OSMD: LinExp (Entropy Function)

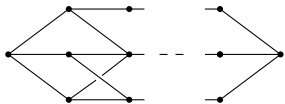
$$\mathcal{D} = [0, +\infty)^d, F(x) = \frac{1}{\eta} \sum_{i=1}^d x_i \log x_i$$



Full Info: Hedge

Semi-Bandit=Bandit: Exp3

Auer et al. [2002]



Full Info: Component Hedge

Koolen, Warmuth and Kivinen [2010]

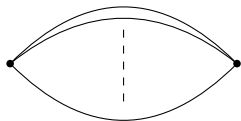
Semi-Bandit: MW

Kale, Reyzin and Schapire [2010]

Bandit: new algorithm

# Different instances of OSMD: LinExp (Entropy Function)

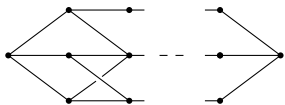
$$\mathcal{D} = [0, +\infty)^d, F(x) = \frac{1}{\eta} \sum_{i=1}^d x_i \log x_i$$



Full Info: Hedge

Semi-Bandit=Bandit: Exp3

Auer et al. [2002]



Full Info: Component Hedge

Koolen, Warmuth and Kivinen [2010]

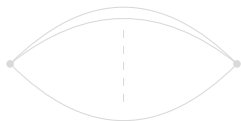
Semi-Bandit: MW

Kale, Reyzin and Schapire [2010]

Bandit: new algorithm

# Different instances of OSMD: LinINF (Exchangeable Hessian)

$$\mathcal{D} = [0, +\infty)^d, F(x) = \sum_{i=1}^d \int_0^{x_i} \psi^{-1}(s) ds$$



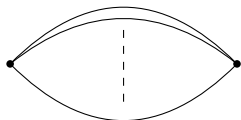
INF, Audibert and Bubeck [2009]



$$\begin{cases} \psi(x) = \exp(\eta x) : \text{LinExp} \\ \psi(x) = (-\eta x)^{-q}, q > 1 : \text{LinPoly} \end{cases}$$

# Different instances of OSMD: LinINF (Exchangeable Hessian)

$$\mathcal{D} = [0, +\infty)^d, F(x) = \sum_{i=1}^d \int_0^{x_i} \psi^{-1}(s) ds$$



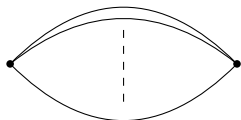
INF, Audibert and Bubeck [2009]



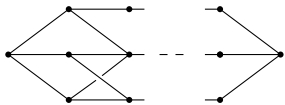
$$\begin{cases} \psi(x) = \exp(\eta x) : \text{LinExp} \\ \psi(x) = (-\eta x)^{-q}, q > 1 : \text{LinPoly} \end{cases}$$

# Different instances of OSMD: LinINF (Exchangeable Hessian)

$$\mathcal{D} = [0, +\infty)^d, F(x) = \sum_{i=1}^d \int_0^{x_i} \psi^{-1}(s) ds$$



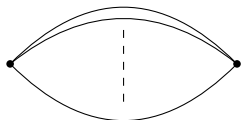
INF, Audibert and Bubeck [2009]



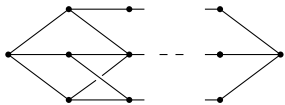
$$\begin{cases} \psi(x) = \exp(\eta x) : \text{LinExp} \\ \psi(x) = (-\eta x)^{-q}, q > 1 : \text{LinPoly} \end{cases}$$

# Different instances of OSMD: LinINF (Exchangeable Hessian)

$$\mathcal{D} = [0, +\infty)^d, F(x) = \sum_{i=1}^d \int_0^{x_i} \psi^{-1}(s) ds$$



INF, Audibert and Bubeck [2009]



$$\begin{cases} \psi(x) = \exp(\eta x) : \text{LinExp} \\ \psi(x) = (-\eta x)^{-q}, q > 1 : \text{LinPoly} \end{cases}$$

$\mathcal{D} = \text{Conv}(S)$ , then

$$w_{t+1} \in \underset{w \in \mathcal{D}}{\text{argmin}} \left( \sum_{s=1}^t \tilde{\ell}_s^T w + F(w) \right)$$

Particularly interesting choice:  $F$  self-concordant barrier function, Abernethy, Hazan and Rakhlin [2008]

$\mathcal{D} = \text{Conv}(S)$ , then

$$w_{t+1} \in \underset{w \in \mathcal{D}}{\operatorname{argmin}} \left( \sum_{s=1}^t \tilde{\ell}_s^T w + F(w) \right)$$

Particularly interesting choice:  $F$  self-concordant barrier function, Abernethy, Hazan and Rakhlin [2008]

## Theorem (Koolen, Warmuth and Kivinen [2010])

In the *full information* game, the *LinExp* strategy (with well-chosen parameters) satisfies for any concept class  $S \subset \{0, 1\}^d$  and any  $L_\infty$ -adversary:

$$R_n \leq d\sqrt{2n}.$$

Moreover for *any strategy*, there exists a subset  $S \subset \{0, 1\}^d$  and an  $L_\infty$ -adversary such that:

$$R_n \geq 0.008 d\sqrt{n}.$$

## Theorem (Audibert, Bubeck and Lugosi [2011])

In the *semi-bandit* game, the *LinExp* strategy (with well-chosen parameters) satisfies for any concept class  $S \subset \{0, 1\}^d$  and any  $L_\infty$ -adversary:

$$R_n \leq d\sqrt{2n}.$$

Moreover for *any strategy*, there exists a subset  $S \subset \{0, 1\}^d$  and an  $L_\infty$ -adversary such that:

$$R_n \geq 0.008 d\sqrt{n}.$$

# Minimax regret for the bandit game

For the **bandit** game the situation becomes trickier.

- First it appears necessary to add some sort of **forced exploration** on  $S$  to control **third order error terms** in the regret bound.
- Second, the control of the quadratic term  $\tilde{\ell}_t^T (\nabla^2 F(w_t))^{-1} \tilde{\ell}_t$  is much more involved than previously.

# Minimax regret for the bandit game

For the **bandit** game the situation becomes trickier.

- First it appears necessary to add some sort of **forced exploration** on  $S$  to control **third order error terms** in the regret bound.

- Second, the control of the quadratic term  $\tilde{\ell}_t^T (\nabla^2 F(w_t))^{-1} \tilde{\ell}_t$  is much more involved than previously.

# Minimax regret for the bandit game

For the **bandit** game the situation becomes trickier.

- First it appears necessary to add some sort of **forced exploration** on  $S$  to control **third order error terms** in the regret bound.
- Second, the control of the quadratic term  $\tilde{\ell}_t^T (\nabla^2 F(w_t))^{-1} \tilde{\ell}_t$  is much more involved than previously.

# Minimax regret for the bandit game

Theorem (Audibert, Bubeck and Lugosi [2011], Bubeck, Cesa-Bianchi and Kakade [2012])

In the *bandit* game, the *Exp2* strategy with *John's exploration* satisfies for any concept class  $S \subset \{0, 1\}^d$  and any  $L_\infty$ -adversary:

$$R_n \leq 4d^2\sqrt{n},$$

and respectively  $R_n \leq 4d\sqrt{n}$  for an  $L_2$ -adversary.

Moreover for *any strategy*, there exists a subset  $S \subset \{0, 1\}^d$  and an  $L_\infty$ -adversary such that:

$$R_n \geq 0.01 d^{3/2}\sqrt{n}.$$

For  $L_2$ -adversaries the lower bound is  $0.05 \min(n, d\sqrt{n})$ .

**Conjecture:** for an  $L_\infty$ -adversary the correct order of magnitude is  $d^{3/2}\sqrt{n}$  and it can be attained with OSMD.