Towards Minimax Policies for Online Linear Optimization with Bandit Feedback

Sébastien Bubeck

joint work with Nicolò Cesa-Bianchi and Sham M. Kakade.

ORFE Operations Research & Financial Engineering

PRINCETON UNIVERSITY

Online Linear Optimization with Bandit Feedback

Parameters available to the player: number of rounds n, action set $\mathcal{A} \subset \mathbb{R}^d$, action set of the adversary $\mathcal{Z} \subset \mathbb{R}^d$.

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

1. Player chooses an action $a_t \in \mathcal{A}$.
2. Simultaneously the adversary chooses an action $z_t \in \mathcal{Z}$.
3. The player incurs and observes the loss $a_t^\top z_t$.

Goal: Minimize the cumulative (pseudo) regret

$$R_n = \mathbb{E} \sum_{t=1}^{n} a_t^\top z_t - \min_{a \in \mathcal{A}} \mathbb{E} \sum_{t=1}^{n} a^\top z_t \, .$$

Online Linear Optimization with Bandit Feedback

Parameters available to the player: number of rounds n, action set $\mathcal{A} \subset \mathbb{R}^d$, action set of the adversary $\mathcal{Z} \subset \mathbb{R}^d$.

At each round $t = 1, 2, \ldots, n$;

1. Player chooses an action $a_t \in \mathcal{A}$.
2. Simultaneously the adversary chooses an action $z_t \in \mathcal{Z}$.
3. The player incurs and observes the loss $a_t^\top z_t$.

Goal: Minimize the cumulative (pseudo) regret

$$R_n = \mathbb{E} \sum_{t=1}^{n} a_t^\top z_t - \min_{a \in \mathcal{A}} \mathbb{E} \sum_{t=1}^{n} a^\top z_t \; .$$

Online Linear Optimization with Bandit Feedback

Parameters available to the player: number of rounds n, action set $\mathcal{A} \subset \mathbb{R}^d$, action set of the adversary $\mathcal{Z} \subset \mathbb{R}^d$.

At each round $t = 1, 2, \ldots, n$;

1. Player chooses an action $a_t \in \mathcal{A}$.
2. Simultaneously the adversary chooses an action $z_t \in \mathcal{Z}$.
3. The player incurs and observes the loss $a_t^\top z_t$.

Goal: Minimize the cumulative (pseudo) regret

$$R_n = \mathbb{E} \sum_{t=1}^{n} a_t^\top z_t - \min_{a \in \mathcal{A}} \mathbb{E} \sum_{t=1}^{n} a^\top z_t \ .$$

Online Linear Optimization with Bandit Feedback

Parameters available to the player: number of rounds n, action set $\mathcal{A} \subset \mathbb{R}^d$, action set of the adversary $\mathcal{Z} \subset \mathbb{R}^d$.

At each round $t = 1, 2, \ldots, n$;

1. Player chooses an action $a_t \in \mathcal{A}$.
2. Simultaneously the adversary chooses an action $z_t \in \mathcal{Z}$.
3. The player incurs and observes the loss $a_t^\top z_t$.

Goal: Minimize the cumulative (pseudo) regret

$$R_n = \mathbb{E} \sum_{t=1}^{n} a_t^\top z_t - \min_{a \in \mathcal{A}} \mathbb{E} \sum_{t=1}^{n} a^\top z_t \ .$$

Online Linear Optimization with Bandit Feedback

Parameters available to the player: number of rounds n, action set $\mathcal{A} \subset \mathbb{R}^d$, action set of the adversary $\mathcal{Z} \subset \mathbb{R}^d$.

At each round $t = 1, 2, \ldots, n$;

1. Player chooses an action $a_t \in \mathcal{A}$.
2. Simultaneously the adversary chooses an action $z_t \in \mathcal{Z}$.
3. The player incurs and observes the loss $a_t^\top z_t$.

Goal: Minimize the cumulative (pseudo) regret

$$R_n = \mathbb{E} \sum_{t=1}^{n} a_t^\top z_t - \min_{a \in \mathcal{A}} \mathbb{E} \sum_{t=1}^{n} a^\top z_t \ .$$

Online Linear Optimization with Bandit Feedback

Parameters available to the player: number of rounds n, action set $\mathcal{A} \subset \mathbb{R}^d$, action set of the adversary $\mathcal{Z} \subset \mathbb{R}^d$.

At each round $t = 1, 2, \ldots, n$;

1. Player chooses an action $a_t \in \mathcal{A}$.
2. Simultaneously the adversary chooses an action $z_t \in \mathcal{Z}$.
3. The player incurs and observes the loss $a_t^\top z_t$.

Goal: Minimize the cumulative (pseudo) regret

$$R_n = \mathbb{E} \sum_{t=1}^{n} a_t^\top z_t - \min_{a \in \mathcal{A}} \mathbb{E} \sum_{t=1}^{n} a^\top z_t \ .$$

Online Linear Optimization with Bandit Feedback

Parameters available to the player: number of rounds n, action set $\mathcal{A} \subset \mathbb{R}^d$, action set of the adversary $\mathcal{Z} \subset \mathbb{R}^d$.

At each round $t = 1, 2, \ldots, n$;

1. Player chooses an action $a_t \in \mathcal{A}$.
2. Simultaneously the adversary chooses an action $z_t \in \mathcal{Z}$.
3. The player incurs and observes the loss $a_t^\top z_t$.

Goal: Minimize the cumulative (pseudo) regret

$$R_n = \mathbb{E} \sum_{t=1}^{n} a_t^\top z_t - \min_{a \in \mathcal{A}} \mathbb{E} \sum_{t=1}^{n} a^\top z_t \ .$$

Minimax policies

We are interested in

$$\sup_{\mathcal{A},\mathcal{Z}} \inf_{strategy} \sup_{adversary} R_n,$$

where the first sup is taken over a suitable class of sets $\mathcal{A}$ and $\mathcal{Z}$. This problem was introduced in McMahan and Blum (2004) and Awerbuch and Kleinberg (2004).

Our goal is to obtain the exact dependency in (n, d) (eventually up to log factors) in the above quantity under the following assumptions:

Assumption

$\mathcal{Z}$ is included in the polar of $\mathcal{A}$, that is $|a^\top z| \leq 1, \forall (a, z) \in \mathcal{A} \times \mathcal{Z}$. $\mathcal{A}$ is bounded and it has a non-empty interior.

Two known algorithms for this task: Exponential weights or Mirror Descent (Nemirovski and Yudin, (1983)). One common difficulty for both methods: how to do an 'optimal' exploration of $\mathcal{A}$?

Minimax policies

We are interested in

$$\sup_{\mathcal{A},\mathcal{Z}} \inf_{strategy} \sup_{adversary} R_n,$$

where the first sup is taken over a suitable class of sets $\mathcal{A}$ and $\mathcal{Z}$. This problem was introduced in McMahan and Blum (2004) and Awerbuch and Kleinberg (2004).

Our goal is to obtain the exact dependency in (n, d) (eventually up to log factors) in the above quantity under the following assumptions:

Assumption

$\mathcal{Z}$ is included in the polar of $\mathcal{A}$, that is $|a^\top z| \leq 1, \forall (a, z) \in \mathcal{A} \times \mathcal{Z}$. $\mathcal{A}$ is bounded and it has a non-empty interior.

Two known algorithms for this task: Exponential weights or Mirror Descent (Nemirovski and Yudin, (1983)). One common difficulty for both methods: how to do an 'optimal' exploration of $\mathcal{A}$?

Minimax policies

We are interested in

$$\sup_{\mathcal{A},\mathcal{Z}} \inf_{strategy} \sup_{adversary} R_n,$$

where the first sup is taken over a suitable class of sets $\mathcal{A}$ and $\mathcal{Z}$. This problem was introduced in McMahan and Blum (2004) and Awerbuch and Kleinberg (2004).

Our goal is to obtain the exact dependency in (n, d) (eventually up to log factors) in the above quantity under the following assumptions:

Assumption

$\mathcal{Z}$ is included in the polar of $\mathcal{A}$, that is $|a^\top z| \leq 1, \forall (a, z) \in \mathcal{A} \times \mathcal{Z}$.
$\mathcal{A}$ is bounded and it has a non-empty interior.

Two known algorithms for this task: Exponential weights or Mirror Descent (Nemirovski and Yudin, (1983)). One common difficulty for both methods: how to do an 'optimal' exploration of $\mathcal{A}$?

Minimax policies

We are interested in

$$\sup_{\mathcal{A},\mathcal{Z}} \inf_{strategy} \sup_{adversary} R_n,$$

where the first sup is taken over a suitable class of sets $\mathcal{A}$ and $\mathcal{Z}$. This problem was introduced in McMahan and Blum (2004) and Awerbuch and Kleinberg (2004).

Our goal is to obtain the exact dependency in (n, d) (eventually up to log factors) in the above quantity under the following assumptions:

Assumption

$\mathcal{Z}$ is included in the polar of $\mathcal{A}$, that is $|a^\top z| \leq 1, \forall (a, z) \in \mathcal{A} \times \mathcal{Z}$.
$\mathcal{A}$ is bounded and it has a non-empty interior.

Two known algorithms for this task: Exponential weights or Mirror Descent (Nemirovski and Yudin, (1983)). One common difficulty for both methods: how to do an 'optimal' exploration of $\mathcal{A}$?

Expanded Exponentially weighted average forecaster (Exp2)

Key observation: if a_t is played at random from some probability distribution $p_t \in \Delta(\mathcal{A})$ (with $p_t(a) > 0, \forall a \in \mathcal{A}$) then one can build an unbiased estimate $\tilde{z}_t$ of the adversary's move z_t:
$\tilde{z}_t = P_t^{-1} a_t a_t^\top z_t$, with $P_t = \mathbb{E}_{a \sim p_t}(aa^\top)$.

Assume that $\mathcal{A}$ is finite. The Exp2 strategy defines the probability distribution p_t with exponential weights, mixed with some exploration distribution $\mu \in \Delta(\mathcal{A})$,

$$p_t(a) = (1-\gamma)\frac{\exp\left(-\eta \sum_{s=1}^{t-1} \tilde{z}_s^\top a\right)}{\sum_{b \in \mathcal{A}} \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{z}_s^\top b\right)} + \gamma\mu.$$

Expanded Exponentially weighted average forecaster (Exp2)

Key observation: if a_t is played at random from some probability distribution $p_t \in \Delta(\mathcal{A})$ (with $p_t(a) > 0, \forall a \in \mathcal{A}$) then one can build an unbiased estimate $\tilde{z}_t$ of the adversary's move z_t:
$\tilde{z}_t = P_t^{-1} a_t a_t^\top z_t$, with $P_t = \mathbb{E}_{a \sim p_t}(a a^\top)$.

Assume that $\mathcal{A}$ is finite. The Exp2 strategy defines the probability distribution p_t with exponential weights, mixed with some exploration distribution $\mu \in \Delta(\mathcal{A})$,

$$p_t(a) = (1-\gamma)\frac{\exp\left(-\eta \sum_{s=1}^{t-1} \tilde{z}_s^\top a\right)}{\sum_{b \in \mathcal{A}} \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{z}_s^\top b\right)} + \gamma\mu.$$

The exploration distribution

- Dani, Hayes and Kakade (2008) used a barycentric spanner for μ and obtained a regret of order $d\sqrt{n \log |\mathcal{A}|}$. Moreover they show that without further assumptions a regret of order $\sqrt{dn \log |\mathcal{A}|}$ is unimprovable (it is tight for $\mathcal{A} = \{-1, 1\}^d$).
- Cesa-Bianchi and Lugosi (2009) used a uniform distribution for μ and obtained for a few specific sets $\mathcal{A}$ a regret of order $\sqrt{dn \log |\mathcal{A}|}$.
- We propose a new distribution, based on John's Theorem from convex geometry, and obtain a regret of order $\sqrt{dn \log |\mathcal{A}|}$ for any finite set $\mathcal{A}$.
  By a discretization argument this also gives a regret of order $d\sqrt{n \log n}$ for any convex body $\mathcal{A}$.

The exploration distribution

- Dani, Hayes and Kakade (2008) used a barycentric spanner for μ and obtained a regret of order $d\sqrt{n \log |\mathcal{A}|}$. Moreover they show that without further assumptions a regret of order $\sqrt{dn \log |\mathcal{A}|}$ is unimprovable (it is tight for $\mathcal{A} = \{-1, 1\}^d$).
- Cesa-Bianchi and Lugosi (2009) used a uniform distribution for μ and obtained for a few specific sets $\mathcal{A}$ a regret of order $\sqrt{dn \log |\mathcal{A}|}$.
- We propose a new distribution, based on John's Theorem from convex geometry, and obtain a regret of order $\sqrt{dn \log |\mathcal{A}|}$ for any finite set $\mathcal{A}$.
  By a discretization argument this also gives a regret of order $d\sqrt{n \log n}$ for any convex body $\mathcal{A}$.

The exploration distribution

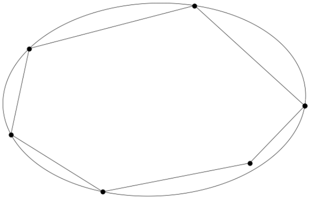
- Dani, Hayes and Kakade (2008) used a barycentric spanner for μ and obtained a regret of order $d\sqrt{n \log |\mathcal{A}|}$. Moreover they show that without further assumptions a regret of order $\sqrt{dn \log |\mathcal{A}|}$ is unimprovable (it is tight for $\mathcal{A} = \{-1, 1\}^d$).
- Cesa-Bianchi and Lugosi (2009) used a uniform distribution for μ and obtained for a few specific sets $\mathcal{A}$ a regret of order $\sqrt{dn \log |\mathcal{A}|}$.
- We propose a new distribution, based on John's Theorem from convex geometry, and obtain a regret of order $\sqrt{dn \log |\mathcal{A}|}$ for any finite set $\mathcal{A}$.
  By a discretization argument this also gives a regret of order $d\sqrt{n \log n}$ for any convex body $\mathcal{A}$.

John's distribution

Theorem (John's Theorem)

Let $\mathcal{K} \subset \mathbb{R}^d$ be a convex set. If the ellipsoid $\mathcal{E}$ of minimal volume enclosing $\mathcal{K}$ is the unit ball in some norm derived from a scalar product $\langle \cdot, \cdot \rangle$, then there exists $M \leq d(d+1)/2 + 1$ contact points $u_1, \ldots, u_M$ between $\mathcal{E}$ and $\mathcal{K}$, and $\mu \in \Delta_M$ (the simplex of dimension $M - 1$), such that

$$x = d \sum_{i=1}^{M} \mu_i \langle x, u_i \rangle u_i, \forall\, x \in \mathbb{R}^d.$$



A few natural questions

1. What about computationally efficient strategies? Abernethy, Hazan and Rakhlin (2008) use Mirror Descent and obtain a regret of order $d\sqrt{\theta n \log n}$ for any $\theta > 0$ such that $Conv(\mathcal{A})$ admits a θ-self concordant barrier (i.e., a suboptimal $d^{3/2}\sqrt{n \log n}$ regret in the worst case).
2. What about optimal regret for specific sets $\mathcal{A}$? A modification of the Mirror Descent strategy described in Abernethy and Rakhlin (2009) attains a regret of order $\sqrt{dn \log n}$ for the Euclidean ball (we provide an alternative strategy and proof for this result).
3. What about the combinatorial setting where $\mathcal{Z} = [-1, 1]^d$ (i.e., $\mathcal{Z}$ is not the polar of $\mathcal{A}$). It was proved in Audibert, Bubeck and Lugosi (2011) that in this setting the Exp2 strategy is provably suboptimal by a factor $\sqrt{d}$ (in the full information setting). In full information (Koolen, Warmuth and Kivinen [2010]) and semi-bandit (ABL11) the key to optimal regret bound is again the Mirror Descent algorithm.

A few natural questions

1. What about computationally efficient strategies? Abernethy, Hazan and Rakhlin (2008) use Mirror Descent and obtain a regret of order $d\sqrt{\theta n \log n}$ for any $\theta > 0$ such that $Conv(\mathcal{A})$ admits a θ-self concordant barrier (i.e., a suboptimal $d^{3/2}\sqrt{n \log n}$ regret in the worst case).
2. What about optimal regret for specific sets $\mathcal{A}$? A modification of the Mirror Descent strategy described in Abernethy and Rakhlin (2009) attains a regret of order $\sqrt{dn \log n}$ for the Euclidean ball (we provide an alternative strategy and proof for this result).
3. What about the combinatorial setting where $\mathcal{Z} = [-1, 1]^d$ (i.e., $\mathcal{Z}$ is not the polar of $\mathcal{A}$). It was proved in Audibert, Bubeck and Lugosi (2011) that in this setting the Exp2 strategy is provably suboptimal by a factor $\sqrt{d}$ (in the full information setting). In full information (Koolen, Warmuth and Kivinen [2010]) and semi-bandit (ABL11) the key to optimal regret bound is again the Mirror Descent algorithm.

A few natural questions

1. What about computationally efficient strategies? Abernethy, Hazan and Rakhlin (2008) use Mirror Descent and obtain a regret of order $d\sqrt{\theta n \log n}$ for any $\theta > 0$ such that $Conv(\mathcal{A})$ admits a θ-self concordant barrier (i.e., a suboptimal $d^{3/2}\sqrt{n \log n}$ regret in the worst case).
2. What about optimal regret for specific sets $\mathcal{A}$? A modification of the Mirror Descent strategy described in Abernethy and Rakhlin (2009) attains a regret of order $\sqrt{dn \log n}$ for the Euclidean ball (we provide an alternative strategy and proof for this result).
3. What about the combinatorial setting where $\mathcal{Z} = [-1, 1]^d$ (i.e., $\mathcal{Z}$ is not the polar of $\mathcal{A}$). It was proved in Audibert, Bubeck and Lugosi (2011) that in this setting the Exp2 strategy is provably suboptimal by a factor $\sqrt{d}$ (in the full information setting). In full information (Koolen, Warmuth and Kivinen [2010]) and semi-bandit (ABL11) the key to optimal regret bound is again the Mirror Descent algorithm.

A few natural questions

1. What about computationally efficient strategies? Abernethy, Hazan and Rakhlin (2008) use Mirror Descent and obtain a regret of order $d\sqrt{\theta n \log n}$ for any $\theta > 0$ such that $Conv(\mathcal{A})$ admits a θ-self concordant barrier (i.e., a suboptimal $d^{3/2}\sqrt{n \log n}$ regret in the worst case).
2. What about optimal regret for specific sets $\mathcal{A}$? A modification of the Mirror Descent strategy described in Abernethy and Rakhlin (2009) attains a regret of order $\sqrt{dn \log n}$ for the Euclidean ball (we provide an alternative strategy and proof for this result).
3. What about the combinatorial setting where $\mathcal{Z} = [-1, 1]^d$ (i.e., $\mathcal{Z}$ is not the polar of $\mathcal{A}$). It was proved in Audibert, Bubeck and Lugosi (2011) that in this setting the Exp2 strategy is provably suboptimal by a factor $\sqrt{d}$ (in the full information setting). In full information (Koolen, Warmuth and Kivinen [2010]) and semi-bandit (ABL11) the key to optimal regret bound is again the Mirror Descent algorithm.

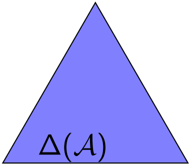
Online Stochastic Mirror Descent (OSMD)

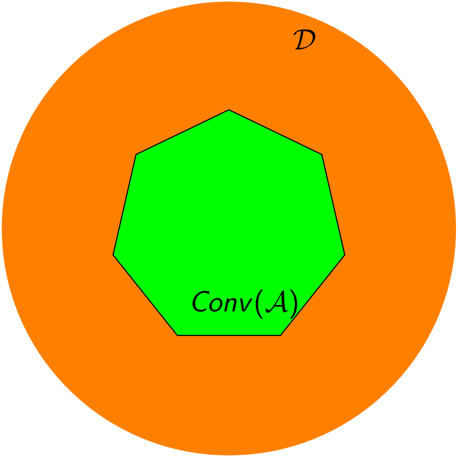
Parameter: F Legendre on $\mathcal{D} \supset Conv(\mathcal{A})$

(1) $w'_{t+1} \in \mathcal{D}$:
$w'_{t+1} = \nabla F^* \left(\nabla F(w_t) - \tilde{z}_t \right)$

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

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





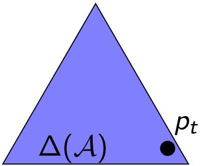
Online Stochastic Mirror Descent (OSMD)

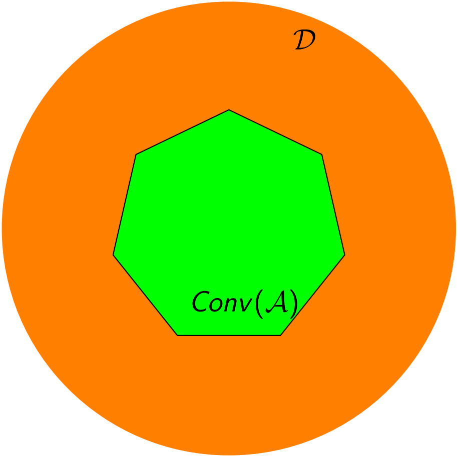
Parameter: F Legendre on $\mathcal{D} \supset Conv(\mathcal{A})$

(1) $w'_{t+1} \in \mathcal{D}$:
$w'_{t+1} = \nabla F^* \left(\nabla F(w_t) - \tilde{z}_t \right)$

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

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





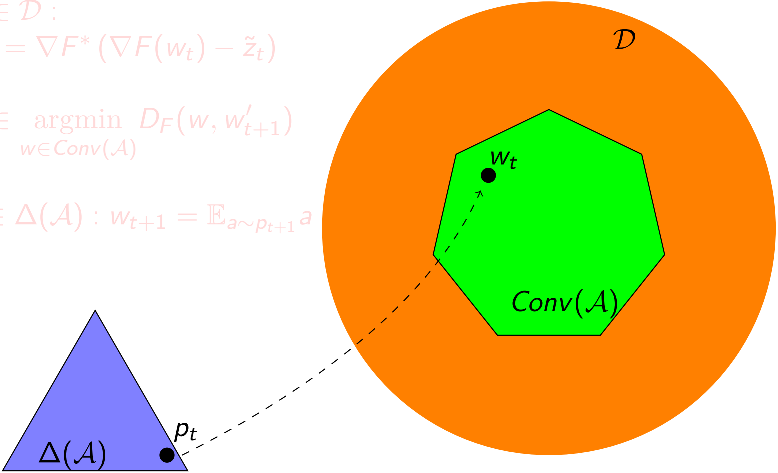
Online Stochastic Mirror Descent (OSMD)

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

(1) $w'_{t+1} \in \mathcal{D}$:
$w'_{t+1} = \nabla F^* \left(\nabla F(w_t) - \tilde{z}_t \right)$

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

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



Online Stochastic Mirror Descent (OSMD)

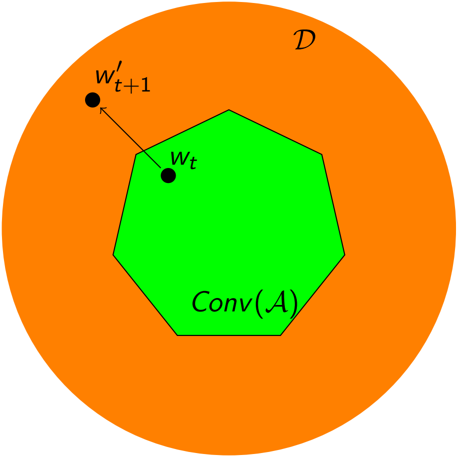
Parameter: F Legendre on $\mathcal{D} \supset Conv(\mathcal{A})$

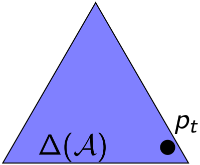
(1) $w'_{t+1} \in \mathcal{D}$:

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

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

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





Online Stochastic Mirror Descent (OSMD)

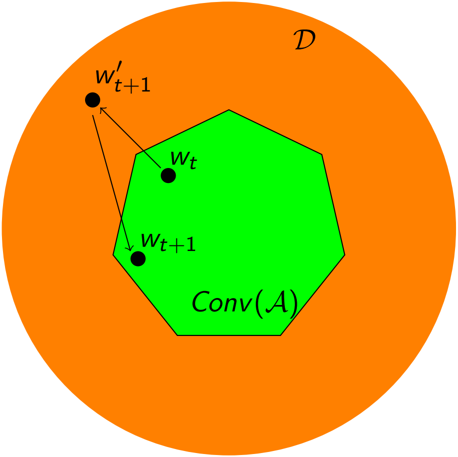
Parameter: F Legendre on $\mathcal{D} \supset Conv(\mathcal{A})$

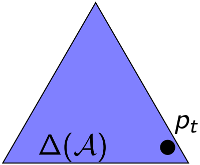
(1) $w'_{t+1} \in \mathcal{D}$:

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

(2) $w_{t+1} \in \underset{w \in Conv(\mathcal{A})}{\operatorname{argmin}} \; D_F(w, w'_{t+1})$

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





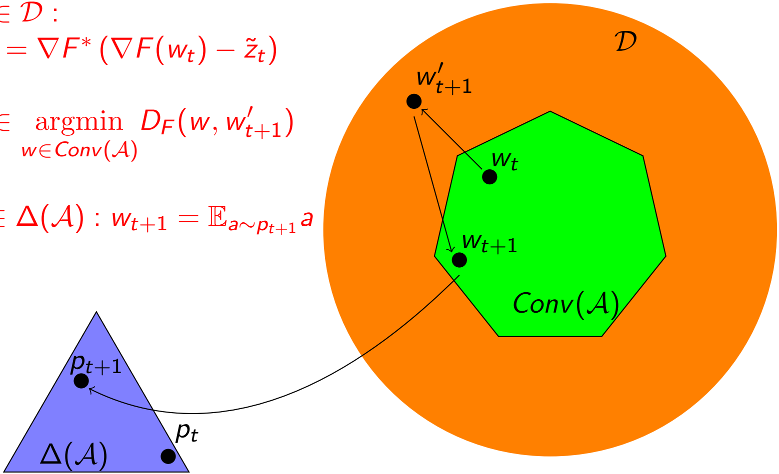
Online Stochastic Mirror Descent (OSMD)

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

(1) $w'_{t+1} \in \mathcal{D}$:
$w'_{t+1} = \nabla F^* \left(\nabla F(w_t) - \tilde{z}_t\right)$

(2) $w_{t+1} \in \underset{w \in Conv(\mathcal{A})}{\operatorname{argmin}} \; D_F(w, w'_{t+1})$

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



An optimal and computationally efficient strategy for the hypercube

We consider here the set $\mathcal{A} = \{-1, 1\}^d$ and we use OSMD with an INF type regularizer, Audibert and Bubeck [2009],

$$F(x) = \sum_{i=1}^{d} \int_{-1}^{x_i} \tanh^{-1}(s) ds$$
$$= \frac{1}{2} \sum_{i=1}^{d} \left((1 + x_i) \log(1 + x_i) + (1 - x_i) \log(1 - x_i) \right) + cst.$$

We choose a very specific distribution p_t to play (approximately) a point $w_t \in Conv(\mathcal{A}) = [-1, 1]^d$:

> *With probability γ, play a_t uniformly at random from the canonical basis (with random sign). With probability $1 - \gamma$, play $a_t = \xi_t$ where $\xi_t(i)$ is drawn from a Rademacher with parameter $\frac{1+w_t(i)}{2}$.*

This strategy has a computational complexity linear in d, and it attains the optimal $d\sqrt{n}$ regret on the hypercube.

An optimal and computationally efficient strategy for the hypercube

We consider here the set $\mathcal{A} = \{-1, 1\}^d$ and we use OSMD with an INF type regularizer, Audibert and Bubeck [2009],

$$F(x) = \sum_{i=1}^{d} \int_{-1}^{x_i} \tanh^{-1}(s) ds$$

$$= \frac{1}{2} \sum_{i=1}^{d} \big((1 + x_i) \log(1 + x_i) + (1 - x_i) \log(1 - x_i) \big) + cst.$$

We choose a very specific distribution p_t to play (approximately) a point $w_t \in Conv(\mathcal{A}) = [-1, 1]^d$:

With probability γ, play a_t uniformly at random from the canonical basis (with random sign). With probability $1 - \gamma$, play $a_t = \xi_t$ where $\xi_t(i)$ is drawn from a Rademacher with parameter $\frac{1+w_t(i)}{2}$.

This strategy has a computational complexity linear in d, and it attains the optimal $d\sqrt{n}$ regret on the hypercube.

An optimal and computationally efficient strategy for the hypercube

We consider here the set $\mathcal{A} = \{-1, 1\}^d$ and we use OSMD with an INF type regularizer, Audibert and Bubeck [2009],

$$F(x) = \sum_{i=1}^{d} \int_{-1}^{x_i} \tanh^{-1}(s) ds$$
$$= \frac{1}{2} \sum_{i=1}^{d} \left((1 + x_i) \log(1 + x_i) + (1 - x_i) \log(1 - x_i) \right) + cst.$$

We choose a very specific distribution p_t to play (approximately) a point $w_t \in Conv(\mathcal{A}) = [-1, 1]^d$:

With probability γ, play a_t uniformly at random from the canonical basis (with random sign). With probability $1 - \gamma$, play $a_t = \xi_t$ where $\xi_t(i)$ is drawn from a Rademacher with parameter $\frac{1+w_t(i)}{2}$.

This strategy has a computational complexity linear in d, and it attains the optimal $d\sqrt{n}$ regret on the hypercube.

An optimal and computationally efficient strategy for the hypercube

We consider here the set $\mathcal{A} = \{-1, 1\}^d$ and we use OSMD with an INF type regularizer, Audibert and Bubeck [2009],

$$\begin{aligned} F(x) &= \sum_{i=1}^{d} \int_{-1}^{x_i} \tanh^{-1}(s) ds \\ &= \frac{1}{2} \sum_{i=1}^{d} \left((1 + x_i) \log(1 + x_i) + (1 - x_i) \log(1 - x_i) \right) + cst. \end{aligned}$$

We choose a very specific distribution p_t to play (approximately) a point $w_t \in Conv(\mathcal{A}) = [-1, 1]^d$:

> *With probability γ, play a_t uniformly at random from the canonical basis (with random sign). With probability $1 - \gamma$, play $a_t = \xi_t$ where $\xi_t(i)$ is drawn from a Rademacher with parameter $\frac{1+w_t(i)}{2}$.*

This strategy has a computational complexity linear in d, and it attains the optimal $d\sqrt{n}$ regret on the hypercube.