Online Learning in Games

DRAFT

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 9: Online learning in games with adaptivity and stochastic feedback

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)

EE-735 (Spring 2024)

License Information for Online Learning in Games Slides

- This work is released under a Creative Commons License with the following terms:
- **Attribution**
  - The licensor permits others to copy, distribute, display, and perform the work. In return, licensees must give the original authors credit.
- **Non-Commercial**
  - The licensor permits others to copy, distribute, display, and perform the work. In return, licensees may not use the work for commercial purposes – unless they get the licensor's permission.
- **Share Alike**
  - The licensor permits others to distribute derivative works only under a license identical to the one that governs the licensor's work.
- Full Text of the License

Acknowledgements

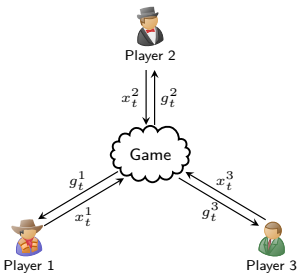
These slides were originally prepared by Wanyun Xie and Luca Viano.

Online Learning in Continuous Games

At each round $t = 1, 2, \ldots$, each player $i \in \mathcal{N}$

- Plays an action $x_t^i \in \mathcal{X}^i$
- Suffers loss $\ell^i(\mathbf{x}_t)$ and receives (first order) feedback $g_t^i \approx \nabla_i \ell^i(\mathbf{x}_t)$

- Each player i has a convex closed action set $\mathcal{X}^i$ and a loss function $\ell^i \colon \mathcal{X}^1 \times \ldots \times \mathcal{X}^N \to \mathbb{R}$
- Joint action of all players $\mathbf{x} = (x^i)_{i \in \mathcal{N}} = (x^i, \mathbf{x}^{-i})$
- $\ell^i(\cdot, \mathbf{x}^{-i})$ is convex and $\nabla_i \ell^i(\mathbf{x}_t)$ is Lipschitz continuous

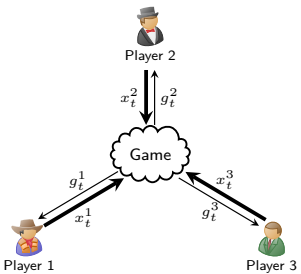


Online Learning in Continuous Games

At each round $t = 1, 2, \ldots$, each player $i \in \mathcal{N}$

- Plays an action $x_t^i \in \mathcal{X}^i$
- Suffers loss $\ell^i(\mathbf{x}_t)$ and receives (first order) feedback $g_t^i \approx \nabla_i \ell^i(\mathbf{x}_t)$

- Each player i has a convex closed action set $\mathcal{X}^i$ and a loss function $\ell^i \colon \mathcal{X}^1 \times \ldots \times \mathcal{X}^N \to \mathbb{R}$
- Joint action of all players $\mathbf{x} = (x^i)_{i \in \mathcal{N}} = (x^i, \mathbf{x}^{-i})$
- $\ell^i(\cdot, \mathbf{x}^{-i})$ is convex and $\nabla_i \ell^i(\mathbf{x}_t)$ is Lipschitz continuous

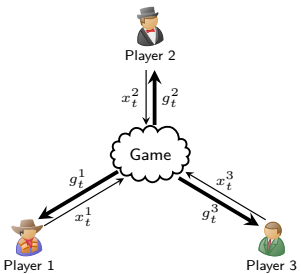


Online Learning in Continuous Games

At each round $t = 1, 2, \ldots$, each player $i \in \mathcal{N}$

- Plays an action $x_t^i \in \mathcal{X}^i$
- Suffers loss $\ell^i(\mathbf{x}_t)$ and receives (first order) feedback $g_t^i \approx \nabla_i \ell^i(\mathbf{x}_t)$

- Each player i has a convex closed action set $\mathcal{X}^i$ and a loss function $\ell^i \colon \mathcal{X}^1 \times \ldots \times \mathcal{X}^N \to \mathbb{R}$
- Joint action of all players $\mathbf{x} = (x^i)_{i \in \mathcal{N}} = (x^i, \mathbf{x}^{-i})$
- $\ell^i(\cdot, \mathbf{x}^{-i})$ is convex and $\nabla_i \ell^i(\mathbf{x}_t)$ is Lipschitz continuous

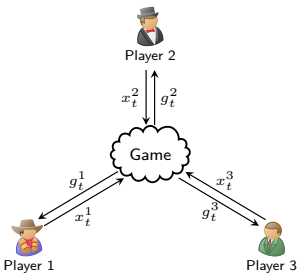


Online Learning in Continuous Games

At each round $t = 1, 2, \ldots$, each player $i \in \mathcal{N}$

- Plays an action $x_t^i \in \mathcal{X}^i$
- Suffers loss $\ell^i(\mathbf{x}_t)$ and receives (first order) feedback $g_t^i \approx \nabla_i \ell^i(\mathbf{x}_t)$

- Each player i has a convex closed action set $\mathcal{X}^i$ and a loss function $\ell^i \colon \mathcal{X}^1 \times \ldots \times \mathcal{X}^N \to \mathbb{R}$
- Joint action of all players $\mathbf{x} = (x^i)_{i \in \mathcal{N}} = (x^i, \mathbf{x}^{-i})$
- $\ell^i(\cdot, \mathbf{x}^{-i})$ is convex and $\nabla_i \ell^i(\mathbf{x}_t)$ is Lipschitz continuous
- Players can be adversarial or optimizing their own benefit



Nash equilibrium and Regret

- Nash equilibrium $\mathbf{x}_\star$: for all $i \in \mathcal{N}$ and all $x^i \in \mathcal{X}^i$, $\ell^i(x^i_\star, \mathbf{x}^{-i}_\star) \leq \ell^i(x^i, \mathbf{x}^{-i}_\star)$
  - Hard to compute in general
  - Players only know the game via gradient feedback
- Individual regret of player i:

$$\mathrm{Reg}^i_T(\mathcal{P}^i) = \max_{p^i \in \mathcal{P}^i} \sum_{t=1}^{T} \Big(\underbrace{\ell^i(x^i_t, \mathbf{x}^{-i}_t) - \ell^i(p^i, \mathbf{x}^{-i}_t)}_{\text{cost of not playing } p^i \text{ in round } t} \Big).$$

  No regret if $\mathrm{Reg}^i_T(\mathcal{P}^i) = o(T)$

- **Nash equilibrium leads to no regret but the converse is more delicate**

Optimistic Gradient

- Standard Gradient Descent Ascent can diverge in bilinear problems.
- We solved the issue with Extragradient but this requires twice the number of oracle calls.
- An alternative is Optimistic gradient (OG) [$\mathbf{x}_t = \mathbf{X}_{t+\frac{1}{2}}$]

$$\mathbf{X}_{t+\frac{1}{2}} = \mathbf{X}_t - \eta_t \hat{\mathbf{V}}_{t-\frac{1}{2}}, \quad \mathbf{X}_{t+1} = \mathbf{X}_t - \eta_{t+1} \hat{\mathbf{V}}_{t+\frac{1}{2}}$$

- OG does not require an intermedite oracle calls.
- It performs the extrapolation step using past gradients.

Problems

- Fast convergence of sequence of play is mostly proved for **suitably tuned** learning rates
  Adaptive Learning in Continuous Games: Optimal Regret Bounds and Convergence to Nash Equilibrium [**?**]

- Nearly constant regret is possible under **only perfect feedback** if all players play some prescribed algorithm
  No-Regret Learning in Games with Noisy Feedback: Faster Rates and Adaptivity via Learning Rate Separation [**?**]

Optimistic Mirror Descent (OptMD)

OptMD class of algorithms has been shown to enjoy optimal regret minimization guarantees.

$$X_t^i = \arg\min_{x \in \mathcal{X}^i} \langle g_{t-1}^i, x \rangle + \frac{D^i(x, X_{t-1}^i)}{\eta_t^i}$$
$$X_{t+\frac{1}{2}}^i = \arg\min_{x \in \mathcal{X}^i} \langle g_{t-1}^i, x \rangle + \frac{D^i(x, X_t^i)}{\eta_t^i} \qquad \text{(OptMD)}$$

Regularizer $h^i : \mathcal{X}^i \to \mathbb{R}$ i.e., a continuous, strongly convex function

Bregman divergence: $D^i(p, x) = h^i(p) - h^i(x) - \langle \nabla h^i(x), p - x \rangle \quad p \in \mathcal{X}^i, x \in \operatorname{dom} \partial h^i$

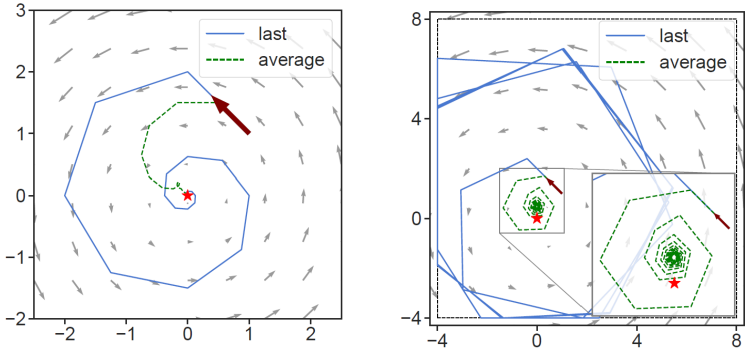
Widely used instances of OptMD:

- Past extra-gradient (PEG): $h^i(x) = \frac{\|x\|_2^2}{2}$
- Optimistic multiplicative weights update (OMWU): $h^i(x) = \sum_{k=1}^{d_i} x_k \log x_k$

Limitation of OptMD

Example: Two-player planar bilinear zero-sum game $\ell^1(\mathbf{x}) = -\ell^2(\mathbf{x}) = x^1 x^2$ where $\mathcal{X}^1 = \mathcal{X}^2 = [-4, 8]$

- Failure with large stepsize



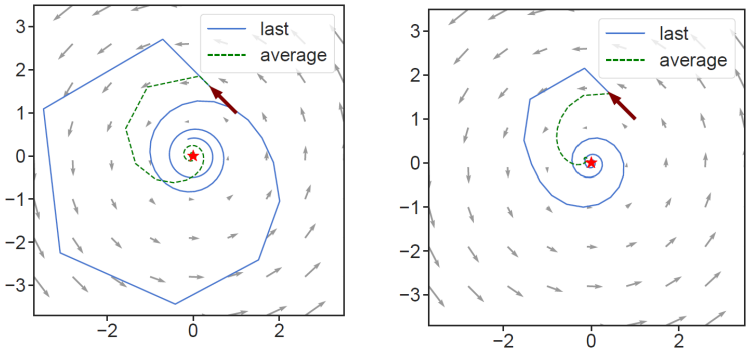
(a) $\eta_t = 0.5$ (b) $\eta_t = 0.7$

Figure: **The trajectories of play**

Limitation of OptMD

Example: Two-player planar bilinear zero-sum game $\ell^1(\mathbf{x}) = -\ell^2(\mathbf{x}) = x^1 x^2$ where $\mathcal{X}^1 = \mathcal{X}^2 = [-4, 8]$

- Decreasing stepsize $\eta_t = \frac{1}{\sqrt{t}}$ $\rightarrow$ slow convergence and slow regret minimization



(a) $\eta_t = \frac{1}{\sqrt{t}}$ (b) Adaptive [?]

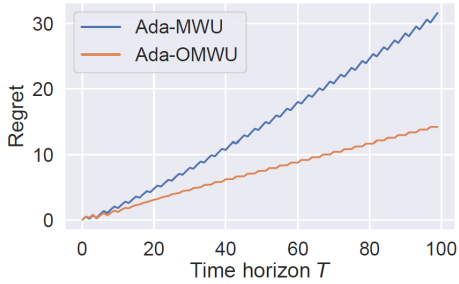
Figure: **The trajectories of play**

Limitation of OptMD

Problem: Mirror descent type methods with adaptive learning rates may lead to superlinear **regret**

Assume that player 1 has a linear loss and simplex-constrained action set.

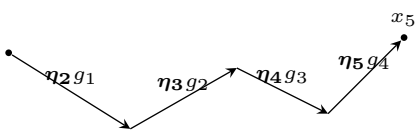
- $\mathcal{X}^1 = \Delta^1 = \{(w_1, w_2) \in \mathbb{R}^2_+, w_1 + w_2 = 1\}$
- Feedback sequence: $\underbrace{[-e_1, \ldots, -e_1}_{[\mathsf{T}/3]}, \underbrace{-e_2, \ldots, -e_2]}_{[2\mathsf{T}/3]}$
- Adaptive (Optimistic) Multiplicative Weight Update

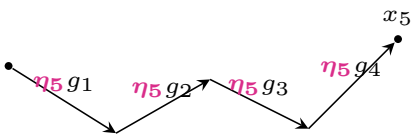


Limitation of OptMD

Cause: New information enters MD with a decreasing weight

Solution: Enter each feedback with equal weight (e.g. Dual averaging or stabilization technique)





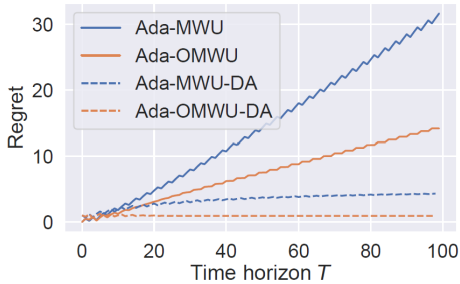
Limitation of OptMD

Cause: New information enters MD with a decreasing weight

Solution: Enter each feedback with equal weight (e.g. Dual averaging or stabilization technique)

Assume that player 1 has a linear loss and simplex-constrained action set.

- $\mathcal{X}^1 = \Delta^1 = \{(w_1, w_2) \in \mathbb{R}^2_+, w_1 + w_2 = 1\}$
- Feedback sequence: $\underbrace{[-e_1, \dots, -e_1}_{[T/3]}, \underbrace{[-e_2, \dots, -e_2]}_{[2T/3]}$
- Adaptive (Optimistic) Multiplicative Weight Update with Dual Averaging



Optimistic Dual Averaging (OptDA)

In contrast to OptMD, OptDA aggregates all feedback received with the same weight.
Each player selects an action $x_t^i = X_{t+\frac{1}{2}}^i$ after taking a "conservatively optimistic" step forward.

$$Y_t^i = -\eta_t^i \sum_{s=1}^{t-1} g_s^i$$

$$X_t^i = \arg\min_{x \in \mathcal{X}^i} \sum_{s=1}^{t-1} \langle g_s^i, x \rangle + \frac{h^i(x)}{\eta_t^i} = Q(Y_t^i)$$

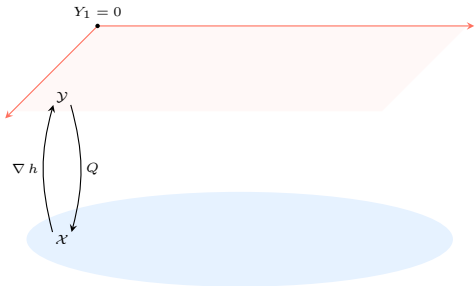
$$X_{t+\frac{1}{2}}^i = \arg\min_{x \in \mathcal{X}^i} \langle g_{t-1}^i, x \rangle + \frac{D^i(x, X_t^i)}{\eta_t^i}$$

Regularizer h^i: 1-strongly convex and C^1

Mirror map: $Q^i(y) = \arg\max_{x \in \mathcal{X}^i} \langle y, x \rangle - h^i(x)$

Bregman divergence:
$D^i(p, x) = h^i(p) - h^i(x) - \langle \nabla h^i(x), p - x \rangle$

- Play $x_t = X_{t+\frac{1}{2}}$ and receive feedback g_t
- Accumulate gradient and compute X_{t+1}, $X_{t+\frac{3}{2}}$



Optimistic Dual Averaging (OptDA)

In contrast to OptMD, OptDA aggregates all feedback received with the same weight.
Each player selects an action $x_t^i = X_{t+\frac{1}{2}}^i$ after taking a "conservatively optimistic" step forward.

$$Y_t^i = -\eta_t^i \sum_{s=1}^{t-1} g_s^i$$

$$X_t^i = \arg\min_{x \in \mathcal{X}^i} \sum_{s=1}^{t-1} \langle g_s^i, x \rangle + \frac{h^i(x)}{\eta_t^i} = Q(Y_t^i)$$

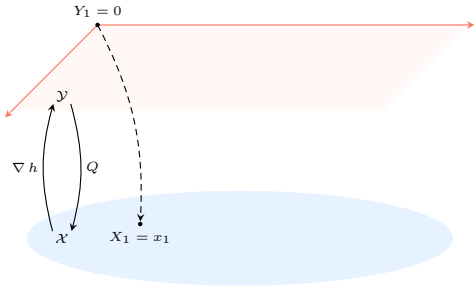
$$X_{t+\frac{1}{2}}^i = \arg\min_{x \in \mathcal{X}^i} \langle g_{t-1}^i, x \rangle + \frac{D^i(x, X_t^i)}{\eta_t^i}$$

Regularizer h^i: 1-strongly convex and C^1

Mirror map: $Q^i(y) = \arg\max_{x \in \mathcal{X}^i} \langle y, x \rangle - h^i(x)$

Bregman divergence:
$D^i(p, x) = h^i(p) - h^i(x) - \langle \nabla h^i(x), p - x \rangle$

- Play $x_t = X_{t+\frac{1}{2}}$ and receive feedback g_t
- Accumulate gradient and compute X_{t+1}, $X_{t+\frac{3}{2}}$



Optimistic Dual Averaging (OptDA)

In contrast to OptMD, OptDA aggregates all feedback received with the same weight.
Each player selects an action $x_t^i = X_{t+\frac{1}{2}}^i$ after taking a "conservatively optimistic" step forward.

$$Y_t^i = -\eta_t^i \sum_{s=1}^{t-1} g_s^i$$

$$X_t^i = \arg\min_{x \in \mathcal{X}^i} \sum_{s=1}^{t-1} \langle g_s^i, x \rangle + \frac{h^i(x)}{\eta_t^i} = Q(Y_t^i)$$

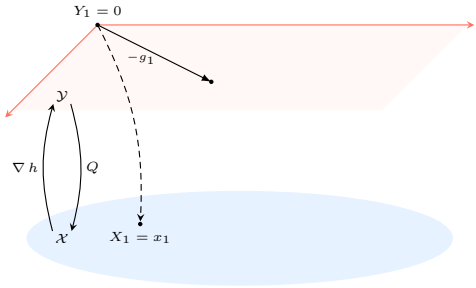
$$X_{t+\frac{1}{2}}^i = \arg\min_{x \in \mathcal{X}^i} \langle g_{t-1}^i, x \rangle + \frac{D^i(x, X_t^i)}{\eta_t^i}$$

Regularizer h^i: 1-strongly convex and C^1

Mirror map: $Q^i(y) = \arg\max_{x \in \mathcal{X}^i} \langle y, x \rangle - h^i(x)$

Bregman divergence:
$D^i(p, x) = h^i(p) - h^i(x) - \langle \nabla h^i(x), p - x \rangle$

- Play $x_t = X_{t+\frac{1}{2}}$ and receive feedback g_t
- Accumulate gradient and compute X_{t+1}, $X_{t+\frac{3}{2}}$



Optimistic Dual Averaging (OptDA)

In contrast to OptMD, OptDA aggregates all feedback received with the same weight.
Each player selects an action $x_t^i = X_{t+\frac{1}{2}}^i$ after taking a "conservatively optimistic" step forward.

$$Y_t^i = -\eta_t^i \sum_{s=1}^{t-1} g_s^i$$

$$X_t^i = \arg\min_{x \in \mathcal{X}^i} \sum_{s=1}^{t-1} \langle g_s^i, x \rangle + \frac{h^i(x)}{\eta_t^i} = Q(Y_t^i)$$

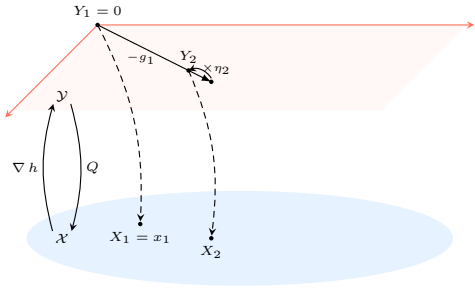
$$X_{t+\frac{1}{2}}^i = \arg\min_{x \in \mathcal{X}^i} \langle g_{t-1}^i, x \rangle + \frac{D^i(x, X_t^i)}{\eta_t^i}$$

Regularizer h^i: 1-strongly convex and C^1

Mirror map: $Q^i(y) = \arg\max_{x \in \mathcal{X}^i} \langle y, x \rangle - h^i(x)$

Bregman divergence:
$D^i(p, x) = h^i(p) - h^i(x) - \langle \nabla h^i(x), p - x \rangle$

- Play $x_t = X_{t+\frac{1}{2}}$ and receive feedback g_t
- Accumulate gradient and compute X_{t+1}, $X_{t+\frac{3}{2}}$



Optimistic Dual Averaging (OptDA)

In contrast to OptMD, OptDA aggregates all feedback received with the same weight.
Each player selects an action $x_t^i = X_{t+\frac{1}{2}}^i$ after taking a "conservatively optimistic" step forward.

$$Y_t^i = -\eta_t^i \sum_{s=1}^{t-1} g_s^i$$

$$X_t^i = \arg\min_{x \in \mathcal{X}^i} \sum_{s=1}^{t-1} \langle g_s^i, x \rangle + \frac{h^i(x)}{\eta_t^i} = Q(Y_t^i)$$

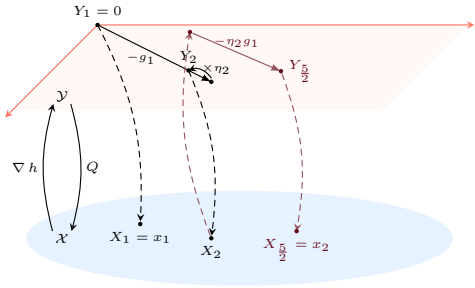
$$X_{t+\frac{1}{2}}^i = \arg\min_{x \in \mathcal{X}^i} \langle g_{t-1}^i, x \rangle + \frac{D^i(x, X_t^i)}{\eta_t^i}$$

Regularizer h^i: 1-strongly convex and C^1

Mirror map: $Q^i(y) = \arg\max_{x \in \mathcal{X}^i} \langle y, x \rangle - h^i(x)$

Bregman divergence:
$D^i(p, x) = h^i(p) - h^i(x) - \langle \nabla h^i(x), p - x \rangle$

- Play $x_t = X_{t+\frac{1}{2}}$ and receive feedback g_t
- Accumulate gradient and compute X_{t+1}, $X_{t+\frac{3}{2}}$



Optimistic Dual Averaging (OptDA)

In contrast to OptMD, OptDA aggregates all feedback received with the same weight.
Each player selects an action $x_t^i = X_{t+\frac{1}{2}}^i$ after taking a "conservatively optimistic" step forward.

$$Y_t^i = -\eta_t^i \sum_{s=1}^{t-1} g_s^i$$

$$X_t^i = \arg\min_{x \in \mathcal{X}^i} \sum_{s=1}^{t-1} \langle g_s^i, x \rangle + \frac{h^i(x)}{\eta_t^i} = Q(Y_t^i)$$

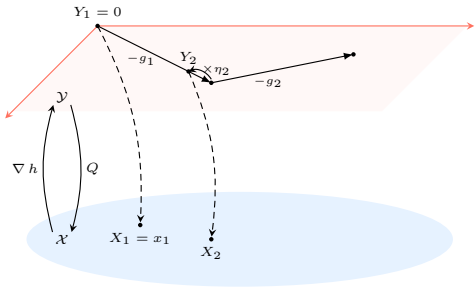
$$X_{t+\frac{1}{2}}^i = \arg\min_{x \in \mathcal{X}^i} \langle g_{t-1}^i, x \rangle + \frac{D^i(x, X_t^i)}{\eta_t^i}$$

Regularizer h^i: 1-strongly convex and C^1

Mirror map: $Q^i(y) = \arg\max_{x \in \mathcal{X}^i} \langle y, x \rangle - h^i(x)$

Bregman divergence:
$D^i(p, x) = h^i(p) - h^i(x) - \langle \nabla h^i(x), p - x \rangle$

- Play $x_t = X_{t+\frac{1}{2}}$ and receive feedback g_t
- Accumulate gradient and compute X_{t+1}, $X_{t+\frac{3}{2}}$



Optimistic Dual Averaging (OptDA)

In contrast to OptMD, OptDA aggregates all feedback received with the same weight.
Each player selects an action $x_t^i = X_{t+\frac{1}{2}}^i$ after taking a "conservatively optimistic" step forward.

$$Y_t^i = -\eta_t^i \sum_{s=1}^{t-1} g_s^i$$

$$X_t^i = \arg\min_{x \in \mathcal{X}^i} \sum_{s=1}^{t-1} \langle g_s^i, x \rangle + \frac{h^i(x)}{\eta_t^i} = Q(Y_t^i)$$

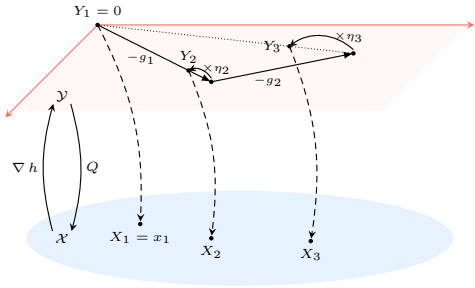
$$X_{t+\frac{1}{2}}^i = \arg\min_{x \in \mathcal{X}^i} \langle g_{t-1}^i, x \rangle + \frac{D^i(x, X_t^i)}{\eta_t^i}$$

Regularizer h^i: 1-strongly convex and C^1

Mirror map: $Q^i(y) = \arg\max_{x \in \mathcal{X}^i} \langle y, x \rangle - h^i(x)$

Bregman divergence:
$D^i(p, x) = h^i(p) - h^i(x) - \langle \nabla h^i(x), p - x \rangle$

- Play $x_t = X_{t+\frac{1}{2}}$ and receive feedback g_t
- Accumulate gradient and compute X_{t+1}, $X_{t+\frac{3}{2}}$



Optimistic Dual Averaging (OptDA)

In contrast to OptMD, OptDA aggregates all feedback received with the same weight.
Each player selects an action $x_t^i = X_{t+\frac{1}{2}}^i$ after taking a "conservatively optimistic" step forward.

$$Y_t^i = -\eta_t^i \sum_{s=1}^{t-1} g_s^i$$

$$X_t^i = \arg\min_{x \in \mathcal{X}^i} \sum_{s=1}^{t-1} \langle g_s^i, x \rangle + \frac{h^i(x)}{\eta_t^i} = Q(Y_t^i)$$

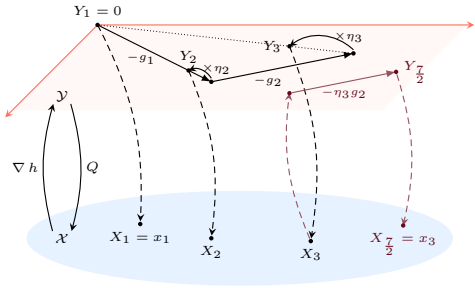
$$X_{t+\frac{1}{2}}^i = \arg\min_{x \in \mathcal{X}^i} \langle g_{t-1}^i, x \rangle + \frac{D^i(x, X_t^i)}{\eta_t^i}$$

Regularizer h^i: 1-strongly convex and C^1

Mirror map: $Q^i(y) = \arg\max_{x \in \mathcal{X}^i} \langle y, x \rangle - h^i(x)$

Bregman divergence:
$D^i(p, x) = h^i(p) - h^i(x) - \langle \nabla h^i(x), p - x \rangle$

- Play $x_t = X_{t+\frac{1}{2}}$ and receive feedback g_t
- Accumulate gradient and compute X_{t+1}, $X_{t+\frac{3}{2}}$



Energy inequality

Lemma

Suppose that player i runs OptDA (or DS-OptMD). For any $p^i \in \mathcal{X}^i$, we have

$$\lambda_{t+1}^i \psi_{t+1}^i\left(p^i\right) \leq \lambda_t^i \psi_t^i\left(p^i\right) - \underbrace{\left\langle g_t^i, X_{t+\frac{1}{2}}^i - p^i \right\rangle}_{\text{linearized regret}} + \left(\lambda_{t+1}^i - \lambda_t^i\right) \varphi^i\left(p^i\right)$$

$$+ \left\langle \underbrace{g_t^i - g_{t-1}^i}_{\text{gradient variation}}, X_{t+\frac{1}{2}}^i - X_{t+1}^i \right\rangle - \underbrace{\lambda_t^i D^i\left(X_{t+1}^i, X_{t+\frac{1}{2}}^i\right) - \lambda_t^i D^i\left(X_{t+\frac{1}{2}}^i, X_t^i\right)}_{\text{distance between successive iterates}}$$

where $(\psi_t^i)_{t\in\mathbb{N}}$ and φ are non-negative, and $\lambda_t^i = 1/\eta_t^i$.

ψ_t^i is a convergence measure (Bregman divergence or Fenchel coupling)

- $\psi_t^i(p^i) \geq \frac{1}{2}\|X_t^i - p_t^i\|^2$
- Reciprocity condition: $X_t^i \to p_t^i$ then $\psi_t^i(p^i) \to 0$

Energy inequality

Lemma

Suppose that player i runs OptDA (or DS-OptMD). For any $p^i \in \mathcal{X}^i$, we have

$$\lambda_{t+1}^i \psi_{t+1}^i \left(p^i\right) \leq \lambda_t^i \psi_t^i \left(p^i\right) - \left\langle g_t^i, X_{t+\frac{1}{2}}^i - p^i \right\rangle + \left(\lambda_{t+1}^i - \lambda_t^i\right) \varphi^i \left(p^i\right)$$
$$+ \left\langle g_t^i - g_{t-1}^i, X_{t+\frac{1}{2}}^i - X_{t+1}^i \right\rangle - \lambda_t^i D^i \left(X_{t+1}^i, X_{t+\frac{1}{2}}^i\right) - \lambda_t^i D^i \left(X_{t+\frac{1}{2}}^i, X_t^i\right)$$

where $(\psi_t^i)_{t\in\mathbb{N}}$ and φ are non-negative, and $\lambda_t^i = 1/\eta_t^i$.

Sum the energy inequality from $t = 1$ to T gives

$$\sum_{t=1}^{T} \left\langle g_t^i, X_{t+\frac{1}{2}}^i - p^i \right\rangle \leq \lambda_{T+1}^i \varphi^i \left(p^i\right) + \sum_{t=1}^{T} \frac{\left\| g_t^i - g_{t-1}^i \right\|_{(i),*}^2}{\lambda_t^i} - \sum_{t=2}^{T} \frac{\lambda_{t-1}^i}{8} \left\| X_{t+\frac{1}{2}}^i - X_{t-\frac{1}{2}}^i \right\|_{(i)}^2$$

Adaptive learning rate

Rearranging, we get

$$\sum_{t=1}^{T} \left\langle g_t^i, X_{t+\frac{1}{2}}^i - p^i \right\rangle \leq \lambda_{T+1}^i \varphi^i\left(p^i\right) + \sum_{t=1}^{T} \frac{\left\| g_t^i - g_{t-1}^i \right\|_{(i),*}^2}{\lambda_t^i} - \sum_{t=2}^{T} \frac{\lambda_{t-1}^i}{8} \left\| X_{t+\frac{1}{2}}^i - X_{t-\frac{1}{2}}^i \right\|_{(i)}^2 \tag{1}$$

Take the adaptive learning rate

$$\eta_t^i = \frac{1}{\sqrt{\tau^i + \sum_{s=1}^{t-1} \left\| g_t^i - g_{t-1}^i \right\|_{(i),*}^2}} \tag{Adapt}$$

- $\tau^i > 0$ can be chosen freely by the player
- η_t^i is computed solely based on local information available to each player

Theoretical guarantees for general convex games

Let player i play OptDA or DS-OptMD with (Adapt):

- No-regret

Theorem

If $\mathcal{P}^i \subseteq \mathcal{X}^i$ is bounded and $G = \sup_t \|g_t^i\|$, the regret incurred by the player is bounded as $\mathrm{Reg}_T^i(\mathcal{P}^i) = \mathcal{O}(G\sqrt{T} + G^2)$

Drop $-\sum_{t=2}^{T} \frac{\lambda_{t-1}^i}{8} \left\| X_{t+\frac{1}{2}}^i - X_{t-\frac{1}{2}}^i \right\|_{(i)}^2$ in (1) gives

$$\sum_{t=1}^{T} \left\langle g_t^i, X_{t+\frac{1}{2}}^i - p^i \right\rangle \leq \lambda_{T+1}^i \varphi^i\left(p^i\right) + \sum_{t=1}^{T} \frac{\left\| g_t^i - g_{t-1}^i \right\|_{(i),*}^2}{\lambda_t^i}$$

- Consistent

  If $\mathcal{X}^i$ is compact and the action profile x_t^{-i} of all other players converges to some limit profile x_∞^{-i}, the trajectory of chosen actions of player i converges to the best response set $\arg\min_{x^i \in \mathcal{X}^i} \ell^i(x^i, \mathbf{x}_\infty^{-i})$.

Variational Stability

Definition (Variationally stable games)

Let $\mathbf{V} = (\nabla_1 \ell^1, \ldots, \nabla_M \ell^M)$. A continuous convex game is variationally stable if the set $\mathcal{X}_\star$ of Nash equilibria of the game is nonempty and

$$\langle \mathbf{V}(\mathbf{x}), \mathbf{x} - \mathbf{x}_\star \rangle = \sum_{i=1}^{N} \langle \nabla_i \ell^i(\mathbf{x}, x^i - x^i_\star) \rangle \geq 0 \text{ for all } \mathbf{x} \in \mathcal{X}, \mathbf{x}_\star \in \mathcal{X}_\star. \tag{2}$$

The game is strictly variationally stable if (2) holds as a strict inequality whenever $\mathbf{x} \notin \mathcal{X}_\star$.

Especially, a game is variationally stable if $\mathbf{V}$ is monotone. E.g.

- Convex-concave zero-sum games
- Zero-sum polymatrix games
- Cournot oligopolies
- Kelly auctions

Theoretical guarantees for variationally stable games

If all players use OptDA or DS-OptMD with (Adapt) in a variationally stable game:

- Constant individual regret For all $i \in \mathcal{N}$ and every bounded comparator set $\mathcal{P}^i \subseteq \mathcal{X}^i$, the individual regret of player i is bounded as $\mathrm{Reg}_T^i(\mathcal{P}^i) = \mathcal{O}(1)$.
- Convergence to Nash equilibrium The induced trajectory of play converges to a Nash equilibrium provided that either of the following is satisfied:
  a. The game is strictly variationally stable.
  b. The game is variationally stable and h^i is (sub)differentiable on all $\mathcal{X}^i$.
  c. The players of a two-player finite zero-sum game follow stabilized OMWU.

Theoretical guarantees for variationally stable games: Proof sketch

- Show that λ_t^i convergences to a finite constant when $t \to +\infty$.
- Under a suitable divergence metric, establish the quasi-Fejér monotonicity of the iterates with respect to any Nash equilibrium $\mathbf{x}_\star$.
- Derive that $\|\mathbf{X}_{t+\frac{1}{2}} - \mathbf{X}_t\| \to 0$ and $\|\mathbf{X}_t - \mathbf{X}_{t-\frac{1}{2}}\| \to 0$ as $t \to +\infty$.
- For general (a and b): Prove that every cluster point of the sequence of play is a NE and conclude.
  For OWMU (c): Prove that the sequence of play has at most one cluster point and subsequently this cluster point must be a NE.

Illustrative experiments

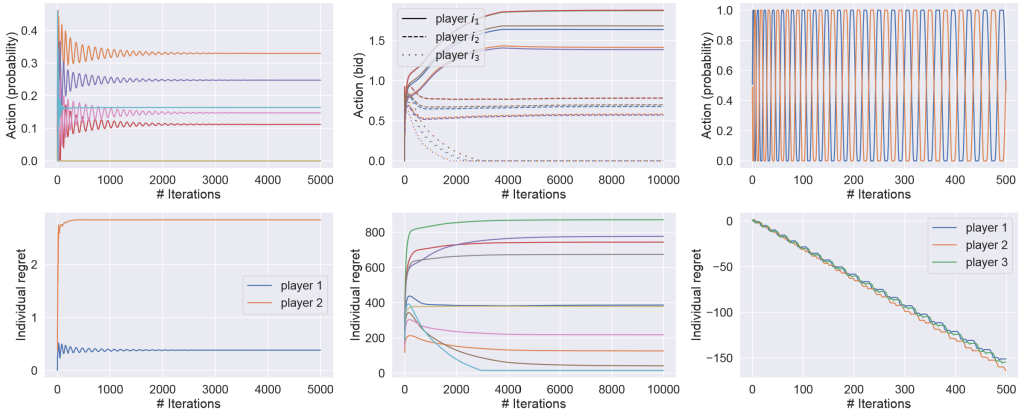


Figure: **Illustrative experiments:** The realized actions (top, each line representing a coordinate of x_t^i) and the individual regret (bottom) of a subset of players in a finite two-player zero-sum game (left), a resource allocation auction (middle), and a three-player matching-pennies game (right). All the players use either adaptive OptDA or adaptive DS-OptMD as their learning strategies. We observe convergence of the realized actions and the regrets in the first two examples.

Summary

A family of algorithms that are

- Adaptive: do not require any prior tuning or knowledge of the game.
- No-regret: achieve $\mathcal{O}(\sqrt{T})$ individual regret against arbitrary opponents.
- Consistent: converge to the best response against convergent opponents.
- Convergent: if employed by all players in a monotone/variationally stable game, the induced sequence of play converges to Nash equilibrium.

Noisy feedback

- Individual regret of player i:

$$\mathrm{Reg}_T^i(\mathcal{P}^i) = \max_{p^i \in \mathcal{P}^i} \sum_{t=1}^{T} \Big(\underbrace{\ell^i(x_t^i, \mathbf{x}_t^{-i}) - \ell^i(p^i, \mathbf{x}_t^{-i})}_{\text{cost of not playing } p^i \text{ in round } t} \Big).$$

  No regret if $\mathrm{Reg}_T^i(\mathcal{P}^i) = o(T)$

  Nearly constant regret is possible under perfect feedback if all players play some prescribed algorithm like OG.

- Stochastic oracle $\mathbb{E}[g_t^i] = \nabla_i \ell^i(\mathbf{x}_t)$
  - Noise: $g_t^i = \nabla_i \ell^i(\mathbf{x}_t) + \xi_t^i$
  - $\mathbb{E}_t\left[\xi_t^i\right] = 0$ (Unbiased)
  - $\mathbb{E}_t\left[\|\xi_t^i\|^2\right] \leq \sigma_A^2 + \sigma_M^2 \|\nabla_i \ell^i(\mathbf{x}_t)\|^2$ (Additive + Multiplicative Noise)

  where ξ_t^i is zero-mean and has finite variance.

- We also use the notation $\widehat{\mathbf{V}}_t = \left[g_t^1, \ldots, g_t^N\right]^T$ and $\mathbf{V}_t = \left[\nabla_1 \ell^1(x_t), \ldots, \nabla_N \ell^N(x_t)\right]^T$.

- Also we use $V^i(x) = \nabla_i \ell^i(x)$.

Constant Regret under Noisy Feedback

Question: OG and others algorithms achieve constant regret in a broad family of games under perfect feedback. Can we achieve the same with noisy feedback?

Answer: Yes if the noise is multiplicative but not with OG !. We need scale separation!
E.g., OG+ $[\mathbf{x}_t = \mathbf{X}_{t+\frac{1}{2}}, \eta_t \leq \gamma_t]$

$$\mathbf{X}_{t+\frac{1}{2}} = \mathbf{X}_t - \gamma_t \hat{\mathbf{V}}_{t-\frac{1}{2}}, \quad \mathbf{X}_{t+1} = \mathbf{X}_t - \eta_t \hat{\mathbf{V}}_{t+\frac{1}{2}}$$

Additional assumption: The game is variationally stable (include monotone games and especially zero-sum polymatrix games). Consider the decision space $\mathcal{X}$ and the solution set $\mathcal{X}^\star$ (assumed non empty)

$$\forall \mathbf{x} \in \mathcal{X}, \forall \mathbf{x}^\star \in \mathcal{X}^\star \, \langle \mathbf{V}(\mathbf{x}), \mathbf{x} - \mathbf{x}^\star \rangle \geq 0$$

Illustrating Example: Failure of Existing Algorithm

- Draw $\mathcal{L}_1(\mathbf{x}) = 3\theta\phi$ or $\mathcal{L}_2(\mathbf{x}) = -\theta\phi$ with equal probability so

$$\ell^1 = -\ell^2 = (\mathcal{L}_1 + \mathcal{L}_2)/2 = \theta\phi$$

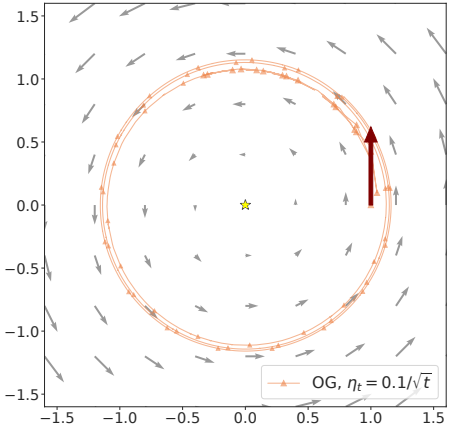
- Stochastic estimate $\mathbb{E}[\hat{\mathbf{V}}_{t+\frac{1}{2}}] = \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})$

$$\hat{\mathbf{V}}_{t+\frac{1}{2}} = \begin{cases} (3\phi_{t+\frac{1}{2}}, -3\theta_{t+\frac{1}{2}}) & \text{with prob. } 1/2 \\ (-\phi_{t+\frac{1}{2}}, \theta_{t+\frac{1}{2}}) & \text{with prob. } 1/2 \end{cases}$$

- Let's compute the variance

$$\begin{aligned} \mathbb{E}_t\left[\hat{\mathbf{V}}_{t+\frac{1}{2}}\right] &= \frac{1}{2}\begin{bmatrix} (3\phi_{t+1/2} - \phi_{t+1/2})^2 \\ (3\theta_{t+1/2} - \theta_{t+1/2})^2 \end{bmatrix} \\ &\quad + \frac{1}{2}\begin{bmatrix} (-\phi_{t+1/2} - \phi_{t+1/2})^2 \\ (-\theta_{t+1/2} - \theta_{t+1/2})^2 \end{bmatrix} \\ &= 4\begin{bmatrix} (\phi_{t+1/2})^2 \\ (-\theta_{t+1/2})^2 \end{bmatrix} = 4\begin{bmatrix} (\nabla_\theta \ell_1)^2 \\ (-\nabla_\phi \ell_2)^2 \end{bmatrix} \end{aligned}$$

- We are in the multiplicative noise case.



Illustrating Example: Scale Separation

- Draw $\mathcal{L}_1(\mathbf{x}) = 3\theta\phi$ or $\mathcal{L}_2(\mathbf{x}) = -\theta\phi$ with equal probability so

$$\ell^1 = -\ell^2 = (\mathcal{L}_1 + \mathcal{L}_2)/2$$

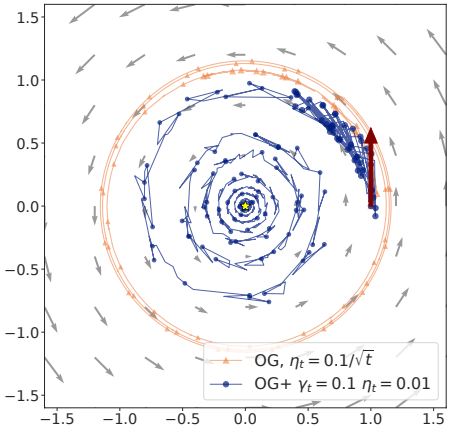
- OG+ $[\mathbf{x}_t = \mathbf{X}_{t+\frac{1}{2}}]$

$$\mathbf{X}_{t+\frac{1}{2}} = \mathbf{X}_t - \gamma_t \hat{\mathbf{V}}_{t-\frac{1}{2}}$$

$$\mathbf{X}_{t+1} = \mathbf{X}_t - \eta_t \hat{\mathbf{V}}_{t+\frac{1}{2}}$$

  With $\gamma_t \geq \eta_t$

- This makes the noise an order smaller than the negative shift in the analysis



Guarantees for OG +

Theorem

Assume all players plays according to OG+ with non-increasing learning rate sequence γ_t and η_t such that

$$\gamma_t = \min\left\{\frac{1}{3L\sqrt{2N(1+\sigma_M^2)}}, \frac{1}{2L(4N+1)\sigma_M^2}\right\}\frac{1}{t^{1/4}\sqrt{\log t}}$$

and

$$\eta_t = \frac{1}{2(1+\sigma_M^2)t^{1/2}\log t}$$

Then, $\mathbb{E}\left[\mathrm{Reg}_{\mathrm{T}}^{\mathrm{i}}(\mathcal{X}^{\mathrm{i}})\right] = \widetilde{\mathcal{O}}(\sqrt{T})$

- The above result is not improvable when both $\sigma_A, \sigma_M > 0$.
- The players' regret is worst then the constant regret achieved under perfect feedback.
- What happens if only $\sigma_M > 0$?

Guarantees for OG +

Theorem

Assume all players plays according to OG+ in a variationally stable game with purely multiplicative noise ($\sigma_A = 0$) with constant learning rate sequence γ_t and η_t such that

$$\gamma_t = \min\left\{\frac{1}{3L\sqrt{2N(1+\sigma_M^2)}}, \frac{1}{2L(4N+1)\sigma_M^2}\right\}$$

and

$$\eta_t = \frac{\gamma_t}{2(1+\sigma_M^2)}$$

Then, $\mathbb{E}\left[\mathrm{Reg}_{\mathrm{T}}^{\mathrm{i}}(\mathcal{X}^{\mathrm{i}})\right] = \mathcal{O}(1)$

- The same regret bound as in the perfect feedback case is achieved.
- Perfect fedback is not necessary to obtain constant regret in variationally stable games.
- However, we need to shift from OG to OG+ to obtain constant regret.

A limiting caveat

- The above guarantees for OG+ requires all the players to adopt the **same** learning rates sequence. That is, in the individual players' updates

$$X_{t+1/2}^i = X_t^i - \gamma_t g_{t-1/2}^i$$
$$X_{t+1}^i = X_t^i - \eta_t g_{t+1/2}^i$$

  the learning rates γ_t, η_t do not depend on the player index i.

- A sad consequence is that the regret guarantees no longer holds if some players plays according to other algorithms or just OG+ with different learning rates.
- A second drawback is that we can not prove guarantees for OG+ in the fully adversarial case.

OG + is very fragile

- The latest gradients are weighted less because the learning rate is decreasing.
- This makes the bound vacuous in the adversarial case in unbounded domain with varying step size [**?**].
- In the case of games where all players update their strategy with OG $+$. We can still prove that

$$\max_{t\in[T]} \mathbb{E}\sqrt{\sum_{i=1}^{N}\left[\|X_t^i - x_\star^i\|^2\right]} \leq \sqrt{\sum_{i=1}^{N}\|X_1^i - x_\star^i\|^2 + \mathcal{O}\left(\log T\right)}$$

therefore it is possible to deploy a time varying learning rate.

The solution is OptDA+

- The solution is to replace the primal update step with its dual counterpart.

$$X_{t+1/2}^i = X_t^i - \gamma_t^i \hat{V}_{t-1/2}^i$$

$$X_{t+1}^i = X_1^i - \eta_{t+1}^i \sum_{s=1}^{t} \hat{V}_{s+1/2}^i$$

- The above updates are dubbed OptDA+.
- The crucial point is that in the update steps all the feedbacks are *post*-multiplied by the same learning rate η_{t+1}^i.
- Each player can now adopt different learning rates !

Formal guarantees for OptDA+

Theorem

Assume all players plays according to OptDA+ with non-increasing learning rate sequence γ_t and η_t such that

$$\gamma_t^i \leq \frac{1}{2L} \min \left\{ \frac{1}{\sqrt{2N(1+\sigma_M^2)}}, \frac{1}{(4N+1)\sigma_M^2} \right\}$$

and $\gamma_t^i = \mathcal{O}\left(t^{-1/4}\right)$ *and*

$$\eta_t^i \leq \frac{\gamma_t^i}{2(1+\sigma_M^2)}$$

and $\eta_t^i = \Theta(t^{-1/2})$. *Then,* $\mathbb{E}\left[\mathrm{Reg}_{\mathrm{T}}^{\mathrm{i}}(\mathcal{X}^{\mathrm{i}})\right] = \mathcal{O}(\sqrt{T})$

- Notice that now players can pick different learning rates.
- We improve over the regret bound achieved by OG + by $\log T$.
- The improvement is possible because OptDA + does not need a bound on $\max_{t\in[T]} \mathbb{E}\left[\|X_t^i - x_\star^i\|^2\right]$.

OptDA+ and multiplicative noise

Theorem

Assume each player $i \in [N]$ plays according to OptDA+ with constant sequences γ_t^i and η_t^i such that

$$\gamma_t^i \leq \frac{1}{2L} \min \left\{ \frac{1}{\sqrt{2N(1+\sigma_M^2)}}, \frac{1}{(4N+1)\sigma_M^2} \right\}$$

and

$$\eta_t^i \leq \frac{\gamma_t^i}{2(1+\sigma_M^2)}.$$

Then, if the feedback satisfies $\sigma_A = 0$, it holds that $\mathbb{E}\left[\mathrm{Reg}_{\mathrm{T}}^{\mathrm{i}}(\mathcal{X}^{\mathrm{i}})\right] = \mathcal{O}(1)$

○ In the pure multiplicative noise case the gain is less evident because also OG+ avoids logarithmic terms under the multiplicative noise setting.

○ The reason is that the step size is constant so the problematic term $\sum_{i=1}^{N} \sum_{t=1}^{T} \left(\frac{1}{\eta_t^i} - \frac{1}{\eta_{t+1}^i} \right) \|X_t^i - x_\star^i\|^2 = 0$ trivially.

Adversarial setting regret bound for OptDA+

- We have proven that OptDA+ is more robust than OG+
- That is, regret bounds hold even if players select different learning rates.
- It turns out that OptDA+ enjoys regret guarantees even in the fully adversarial case.

Theorem

Suppose each player implements OptDA+ with $\gamma_t^i = \mathcal{O}(t^{q-1/2})$ *and* $\eta_t^i = \Theta\left(t^{-1/2}\right)$ *for some* $q \in [0, 1/4]$. *If the perfect feedback is bounded then,* $\mathbb{E}\left[\mathrm{Reg}_{\mathrm{T}}^{\mathrm{i}}(\mathcal{X}^{\mathrm{i}})\right] = \mathcal{O}(T^{1/2+q})$

- We notice that the free parameter q allows to interpolate between the optimal regret guarantees in the adversarial or game play setting.
  - $q = 1/4$ is the optimal setting in the game play case.
  - $q = 0$ ensures the best regret bound in the fully adversarial case.

Proof technique

The proofs for OG+ and OptDA+ are based on the following inequality

$$2\mathbb{E}_{t-1}\left[\left\langle V^i(\mathbf{X}_{t+1/2}), X^i_{t+1/2}\right\rangle - p^i\right] \leq \mathbb{E}_{t-1}\Bigg[\frac{\|X^i_t - p^i\|^2}{\eta^i_t} - \frac{\|X^i_{t+1} - p^i\|^2}{\eta^i_{t+1}} + \left(\frac{1}{\eta^i_{t+1}} - \frac{1}{\eta^i_t}\right)\|u^i_t - p^i\|^2$$

$$- \gamma^i_t(\|V^i(\mathbf{X}_{t+1/2})\|^2 + \|V^i(\mathbf{X}_{t-1/2})\|^2) + \gamma^i_t\|V^i(\mathbf{X}_{t+1/2}) - V^i(\mathbf{X}_{t-1/2})\|^2 - \frac{\|\mathbf{X}_{t+1/2} - \mathbf{X}_{t-1/2}\|^2}{2\eta^i_t}$$

$$+ \sum_{j=1}^{N}(\gamma^j_t + \eta^j_t)^2\|\xi^j_{t-1/2}\|^2 + 2\eta^i_t\|g^i_t\|^2 + (\gamma^i_t)^2 L\|\xi^i_{t-1/2}\|^2\Bigg]$$

- The red term telescopes.
- The blue term requires a different treatment for OptDA+ and OG+.
- In OG+ $u^i_t = X^i_t$ therefore the double dependence on i and t of these term forces to choose same learning rate across players.
- In OptDA+ $u^i_t = X^i_1$ therefore the double dependence on t issue is solved.
- The brown term is bounded by smoothness. At this point the oracle models and the step sizes choices gives the result.

Remaining Problems

- Problem 1: Adaptivity to bypass the need for knowing constants and to cope with adversarial opponents ?
- Problem 2: Last-iterate convergence to Nash Equilibrium ?

Illustrating Example: Adaptivity

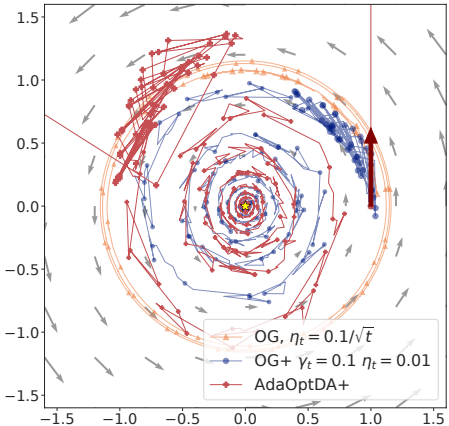
- OptDA+ $[\gamma_t^i \geq \eta_t^i]$

$$X_{t+\frac{1}{2}}^i = X_t^i - \gamma_t^i \, g_{t-1}^i \quad X_{t+1}^i = X_1^i - \eta_{t+1}^i \sum_{s=1}^{t} g_s^i$$

- AdaOptDA+ uses learning rate

$$\gamma_t^i = \frac{1}{\left(1 + \sum_{s=1}^{t-2} \|g_s^i\|^2\right)^{\frac{1}{2}-q}}$$

$$\eta_t^i = \frac{1}{\sqrt{1 + \sum_{s=1}^{t-2} \left(\|g_s^i\|^2 + \|X_s^i - X_{s+1}^i\|^2\right)}}$$



Guarantees for AdaOptDA+

- Using AdaOptDA+ we can prove similar results without knowing the smoothness of the gradient Lipschitz constant L.
- Same applies to the number of players taking part in the game N.
- The price is to pay is the additional assumption that $\forall i \in [N], \forall x^i \in \mathcal{X}^i \ \|V^i(x^i)\| \leq G$ and $\|\xi_t^i\| \leq \bar{\sigma}$ almost surely.

Theorem

Let consider the adversarial case and run AdaOptDA+. Then, it holds that $\mathbb{E}\left[\mathrm{Reg}_{\mathrm{T}}^{\mathrm{i}}(\mathcal{X}^{\mathrm{i}})\right] = \mathcal{O}(T^{1/2+q})$

Theorem

Let consider the game play setting where all players update their strategies according to AdaOptDA+, then we have that

$$\mathbb{E}\left[\mathrm{Reg}_{\mathrm{T}}^{\mathrm{i}}(\mathcal{X}^{\mathrm{i}})\right] = \mathcal{O}(C^{1/q}T^{1/2}) \quad \text{if} \quad \sigma_A > 0.$$

$$\mathbb{E}\left[\mathrm{Reg}_{\mathrm{T}}^{\mathrm{i}}(\mathcal{X}^{\mathrm{i}})\right] = \mathcal{O}(\exp(1/2q)) \quad \text{if} \quad \sigma_A = 0.$$

Benefit of adaptivity and tradeoff of q

- AdaOptDA+ achieves simultaneously the optimal regret bounds with the exact same step sizes.
- This means that we do not even need to know whether $\sigma_A = 0$ or not when we run the algorithm.
- In stark contrast, OptDA+ stepsizes change in the different regimes (additive vs multiplicative noise).
- The q plays an interesting roles. In the adversarial case we have

$$\mathbb{E}\left[\mathrm{Reg}_{\mathrm{T}}^{\mathrm{i}}(\mathcal{X}^{\mathrm{i}})\right] = \mathcal{O}(T^{1/2+q})$$

hence it is minimized for $q = 0$ for which we have the optimal $\mathbb{E}\left[\mathrm{Reg}_{\mathrm{T}}^{\mathrm{i}}(\mathcal{X}^{\mathrm{i}})\right] = \mathcal{O}(T^{1/2})$.
In the game play setting, we have

$$\mathbb{E}\left[\mathrm{Reg}_{\mathrm{T}}^{\mathrm{i}}(\mathcal{X}^{\mathrm{i}})\right] = \mathcal{O}(C^{1/q}T^{1/2}) \quad \text{if} \quad \sigma_A > 0.$$

$$\mathbb{E}\left[\mathrm{Reg}_{\mathrm{T}}^{\mathrm{i}}(\mathcal{X}^{\mathrm{i}})\right] = \mathcal{O}(\exp(1/2q)) \quad \text{if} \quad \sigma_A = 0.$$

Hence, q does not affect the dependence on T but it improves the constants. So we should select q as large as allowed, i.e. $q = 1/4$.

Convergence to Nash equilibrium

Theorem

Consider $\sigma_A = 0$ then AdaOptDA+, OptDA+ and OG+ with aforementioned learning rates produces a sequence $\mathbf{X}_{t+1/2}$ such that it converges to a Nash equilibrium almost surely.

- The key for such prove is to establish the stabilization property $\sum_{t=1}^{T} \|\mathbf{V}(\mathbf{X}_{t+1/2})\|^2 < \infty$ with probability 1.
- Almost sure convergence can be derived also for $\sigma_A > 0$ but only for OG+.
- For OptDA+ and AdaOptDA+ one can prove that the crucial quantity $\sum_{t=1}^{T} \|\mathbf{V}(\mathbf{X}_{t+1/2})\|^2$ grows at a rate determined by q.

Theorem

Let OptDA+ and AdaOptDA+ run with the aforementioned step sizes satisfies $\sum_{t=1}^{T} \|\mathbf{V}(\mathbf{X}_{t+1/2})\|^2 \leq T^{1-q}$.

- For large q, we get more stable trajectory this is consistent with the better regret guarantees achieved for q set as large as allowed $q = 1/4$.

Summary of Results

	Adversarial	All players run the same algorithm			
	Bounded feedback	Additive noise		Multiplicative noise	
	Regret	Regret	Convergence	Regret	Convergence
OG	✗	✗	✗	✗	✗
OG+	✗	$\sqrt{t}\log t$	✓	cst	✓
OptDA+	$\sqrt{t}$	$\sqrt{t}$	–	cst	✓
AdaOptDA+	$\sqrt{t}$	$\sqrt{t}$	–	cst	✓

References I