NO-REGRET LEARNING IN CONTINUOUS GAMES

Panayotis Mertikopoulos[1,2]

[1]French National Center for Scientific Research (CNRS)

[2]Laboratoire d'Informatique de Grenoble (LIG)

GDO conference – Vienna, March 14, 2018

Outline

Online decision processes

---

repeat

At each epoch $n = 1, 2, \ldots$

Choose **action** x_n

Get **payoff** $u_n(x_n)$

until end

Online decision processes

---

repeat

At each epoch $n = 1, 2, \ldots$

Choose **action** x_n

Get **payoff** $u_n(x_n)$

until end

Main question: *How to choose an action at each epoch in an "optimal" way?*

- *Uncertain world:* no beliefs, feedback, knowledge of future, etc.
- *Obliviousness:* Are payoffs affected by the agent's previous actions?
- *Optimality:* What is "optimal" in this setting?

Regret minimization

Performance often quantified by the agent's **regret**

$$[u_n(x) - u_n(x_n)]$$

Regret minimization

Performance often quantified by the agent's **regret**

$$\sum_{n=1}^{T} [u_n(x) - u_n(x_n)]$$

Regret minimization

Performance often quantified by the agent's **regret**

$$\max_{x \in \mathcal{X}} \sum_{n=1}^{T} [u_n(x) - u_n(x_n)]$$

Regret minimization

Performance often quantified by the agent's **regret**

$$\mathrm{Reg}(T) = \max_{x \in \mathcal{X}} \sum_{n=1}^{T} [u_n(x) - u_n(x_n)]$$

Regret minimization

Performance often quantified by the agent's **regret**

$$\mathrm{Reg}(T) = \max_{x \in \mathcal{X}} \sum_{n=1}^{T} [u_n(x) - u_n(x_n)]$$

No regret: $\mathrm{Reg}(T) = o(T)$

"The sequence of chosen actions is asymptotically as good as the best fixed action in hindsight."

Worst-case guarantee: *in the absense of more information, at least minimize regret*

Regret minimization

Performance often quantified by the agent's **regret**

$$\mathrm{Reg}(T) = \max_{x \in \mathcal{X}} \sum_{n=1}^{T} [u_n(x) - u_n(x_n)]$$

No regret: $\mathrm{Reg}(T) = o(T)$

"The sequence of chosen actions is asymptotically as good as the best fixed action in hindsight."

Worst-case guarantee: *in the absense of more information, at least minimize regret*

Prolific literature on no-regret procedures/algorithms:

- Economics *(Hannan, Blackwell, Erev & Roth, Rustichini, Hart & Mas-Colell,…)*
- Computer science *(Vovk, Littlestone & Warmuth, Freund & Schapire,…)*
- Online learning & Optimization *(Cesa-Bianchi & Lugosi, Zinkevich,…)*

No-regret learning in games

Does no-regret lead to rational play?

No-regret learning in games

Does no-regret lead to rational play?

- What does "rational play" mean?
- How to achieve no regret?
- How are these two notions linked?

Outline

Continuous games

The game

- Finite set of *players* $i \in \mathcal{N} = \{1, \ldots, N\}$
- Each player selects an *action* x_i from a compact, convex set $\mathcal{X}_i$
- Reward of player i determined by *payoff function* $u_i : \mathcal{X} \equiv \prod_i \mathcal{X}_i \to \mathbb{R}$

Assumptions

- $u_i(x_i; x_{-i})$ individually concave in x_i
- $u_i(x)$ continuously differentiable on $\mathcal{X}$ [can be relaxed]

Examples

- Finite games (mixed extensions)
- Atomic routing games (splittable)
- Divisible good auctions
- Cournot oligopolies
- ...

Nash equilibrium

Nash equilibrium

Action profile $x^* = (x_1^*, \ldots, x_n^*) \in \mathcal{X}$ that is *unilaterally stable*

$$u_i(x_i^*; x_{-i}^*) \geq u_i(x_i; x_{-i}^*) \quad \text{for every player } i \in \mathcal{N} \text{ and every deviation } x_i \in \mathcal{X}_i$$

Nash equilibrium

Nash equilibrium

Action profile $x^* = (x_1^*, \ldots, x_n^*) \in \mathcal{X}$ that is *unilaterally stable*

$$u_i(x_i^*; x_{-i}^*) \geq u_i(x_i; x_{-i}^*) \quad \text{for every player } i \in \mathcal{N} \text{ and every deviation } x_i \in \mathcal{X}_i$$

Direction of individual payoff ascent

$$v_i(x) = \nabla_{x_i} u_i(x_i; x_{-i})$$

Intuition: $u_i(x_i; x_{-i}^*)$ non-decreasing when moving away from $x_i^* \implies v_i(x^*)$ forms acute angle with $x_i - x_i^*$

Nash equilibrium

Nash equilibrium

Action profile $x^* = (x_1^*, \ldots, x_n^*) \in \mathcal{X}$ that is *unilaterally stable*

$$u_i(x_i^*; x_{-i}^*) \geq u_i(x_i; x_{-i}^*) \quad \text{for every player } i \in \mathcal{N} \text{ and every deviation } x_i \in \mathcal{X}_i$$

Direction of individual payoff ascent

$$v_i(x) = \nabla_{x_i} u_i(x_i; x_{-i})$$

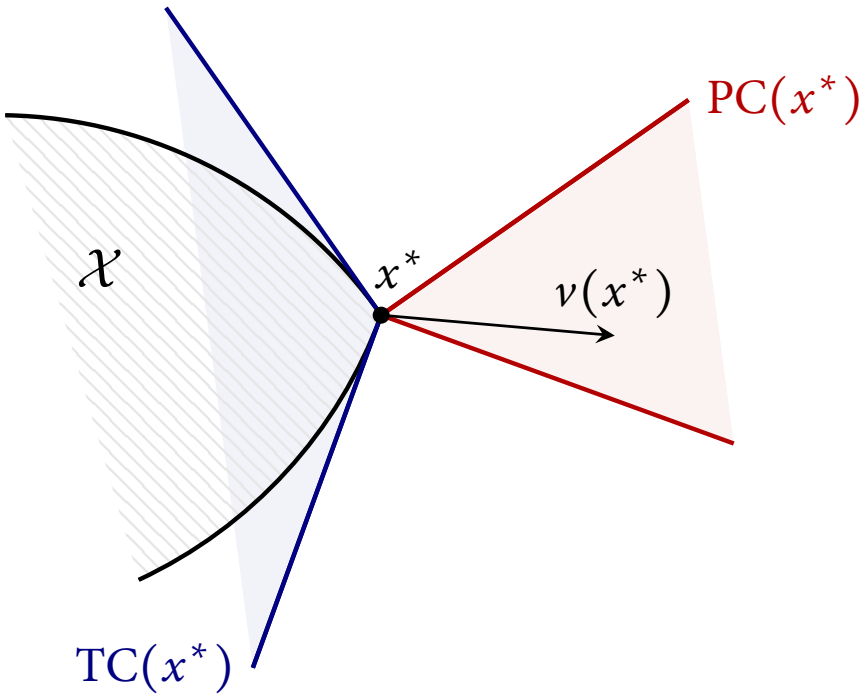
Intuition: $u_i(x_i; x_{-i}^*)$ non-decreasing when moving away from $x_i^* \implies v_i(x^*)$ forms acute angle with $x_i - x_i^*$

Variational characterization

x^* is a Nash equilibrium if and only if

$$\langle v_i(x^*), x_i - x_i^* \rangle \leq 0 \quad \text{for all } i \in \mathcal{N},\ x_i \in \mathcal{X}_i$$

Geometric interpretation



At Nash equilibrium, individual payoff gradients are outward-pointing

Existence and uniqueness

Theorem (Debreu, 1952)

Every concave game admits a Nash equilibrium.

Existence and uniqueness

Theorem (Debreu, 1952)

Every concave game admits a Nash equilibrium.

Theorem (Rosen, 1965)

Suppose that the game is (strictly) monotone*, i.e.*

$$\langle v(x') - v(x), x' - x \rangle < 0 \quad \textit{for all } x \neq x'. \tag{MC}$$

Then, the game admits a unique Nash equilibrium.

Intuition

- Strict concavity for a single player
- "Anti-coordination" for many players

[$-v$ is a monotone operator]

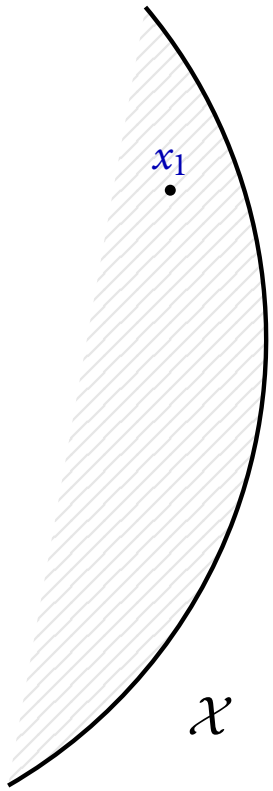
Outline

Achieving no regret

Take a (lazy) gradient step and project (Zinkevich, ICML 2003)

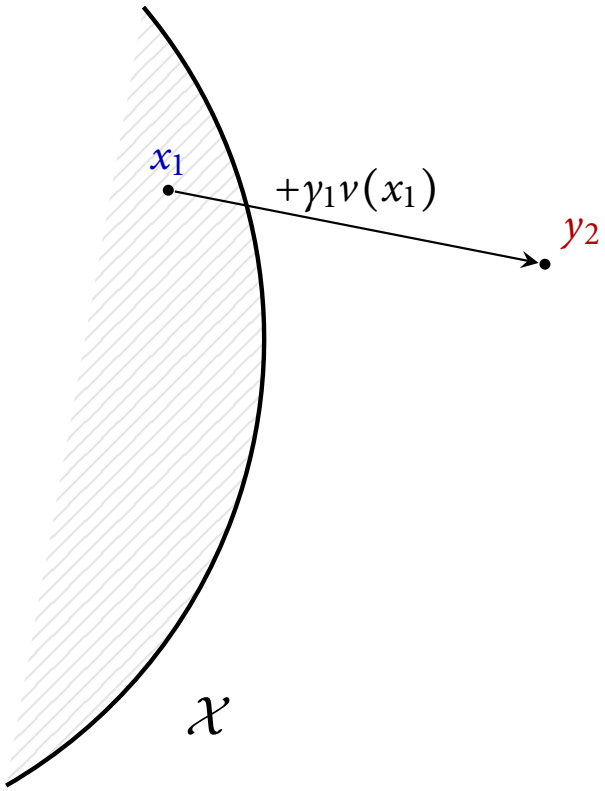
Achieving no regret

Take a (lazy) gradient step and project (Zinkevich, ICML 2003)



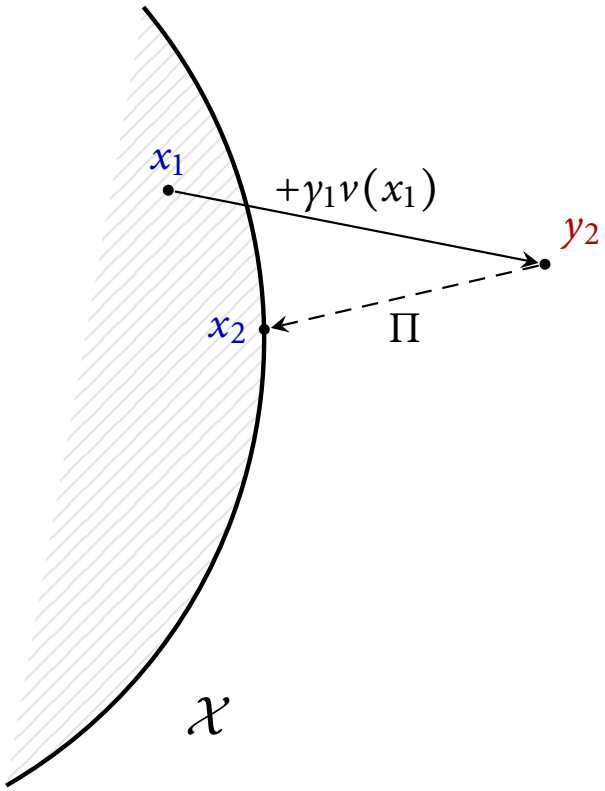
Achieving no regret

Take a (lazy) gradient step and project (Zinkevich, ICML 2003)



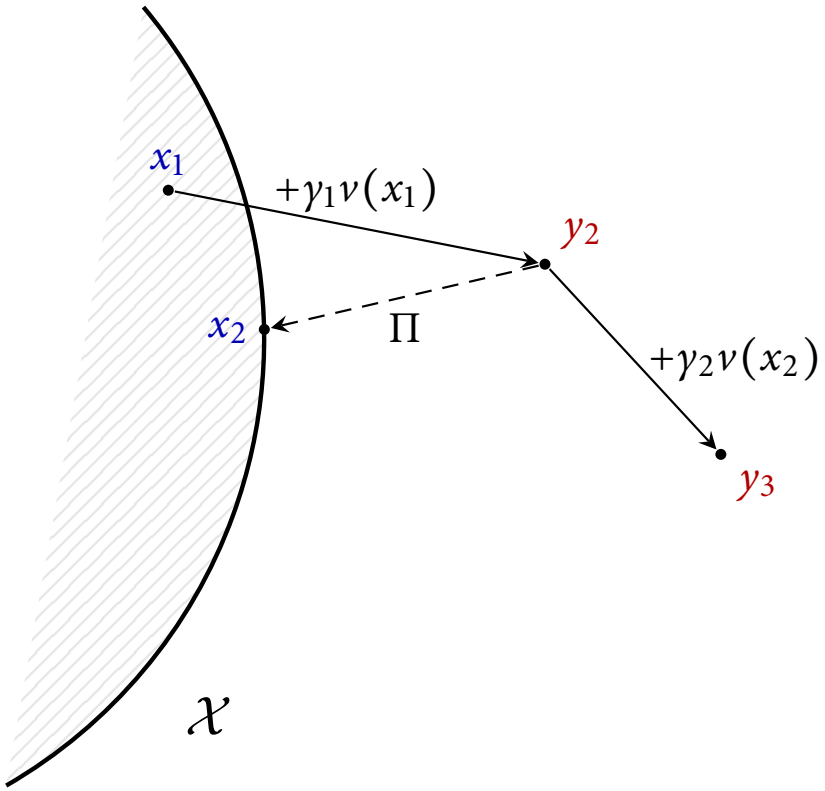
Achieving no regret

Take a (lazy) gradient step and project (Zinkevich, ICML 2003)



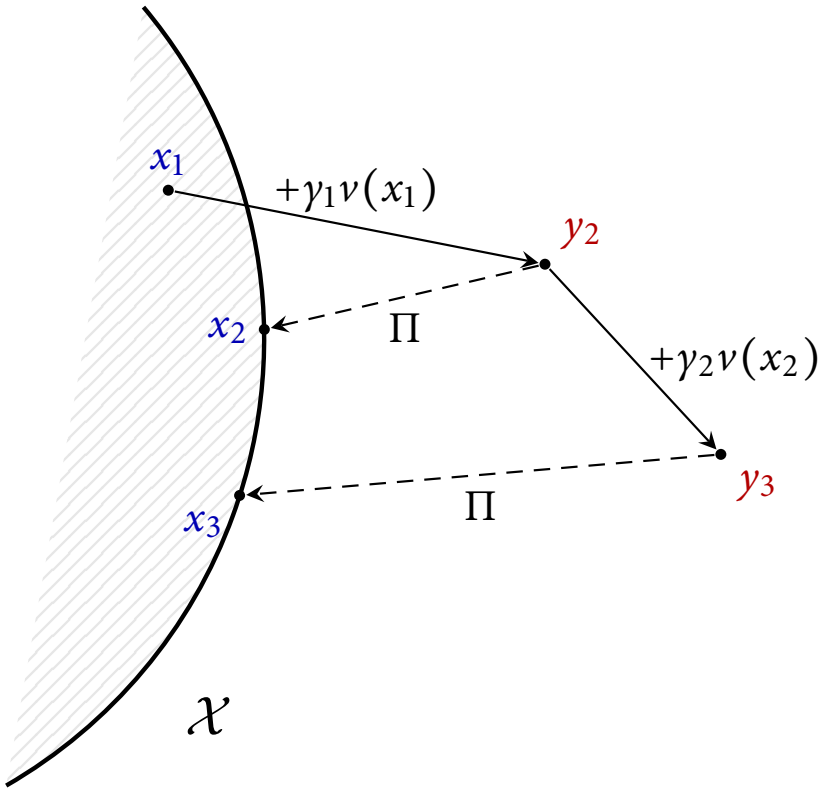
Achieving no regret

Take a (lazy) gradient step and project (Zinkevich, ICML 2003)



Achieving no regret

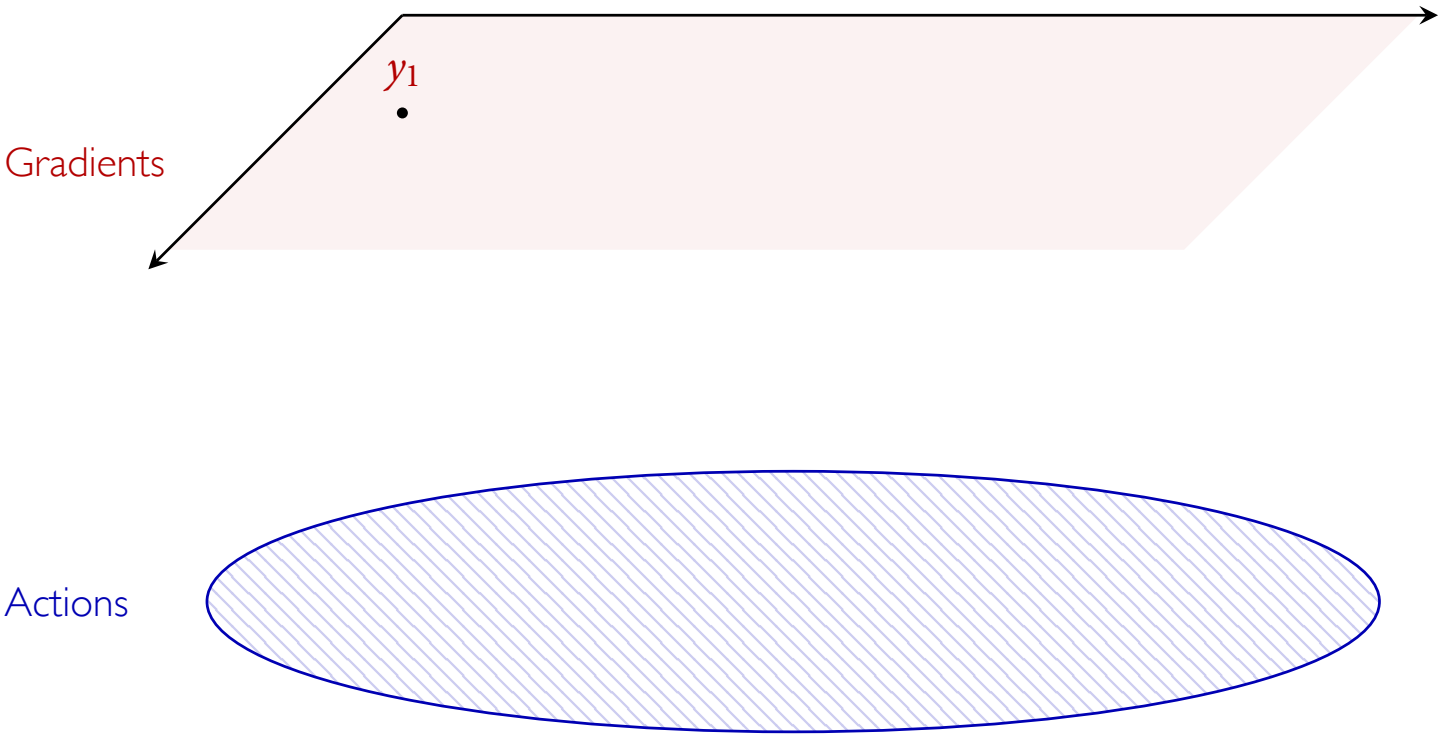
Take a (lazy) gradient step and project (Zinkevich, ICML 2003)



Regret guarantee: $\mathrm{Reg}(T) = \mathcal{O}(T^{1/2})$ if $\gamma_n \propto n^{-1/2}$ *(optimal in T)*

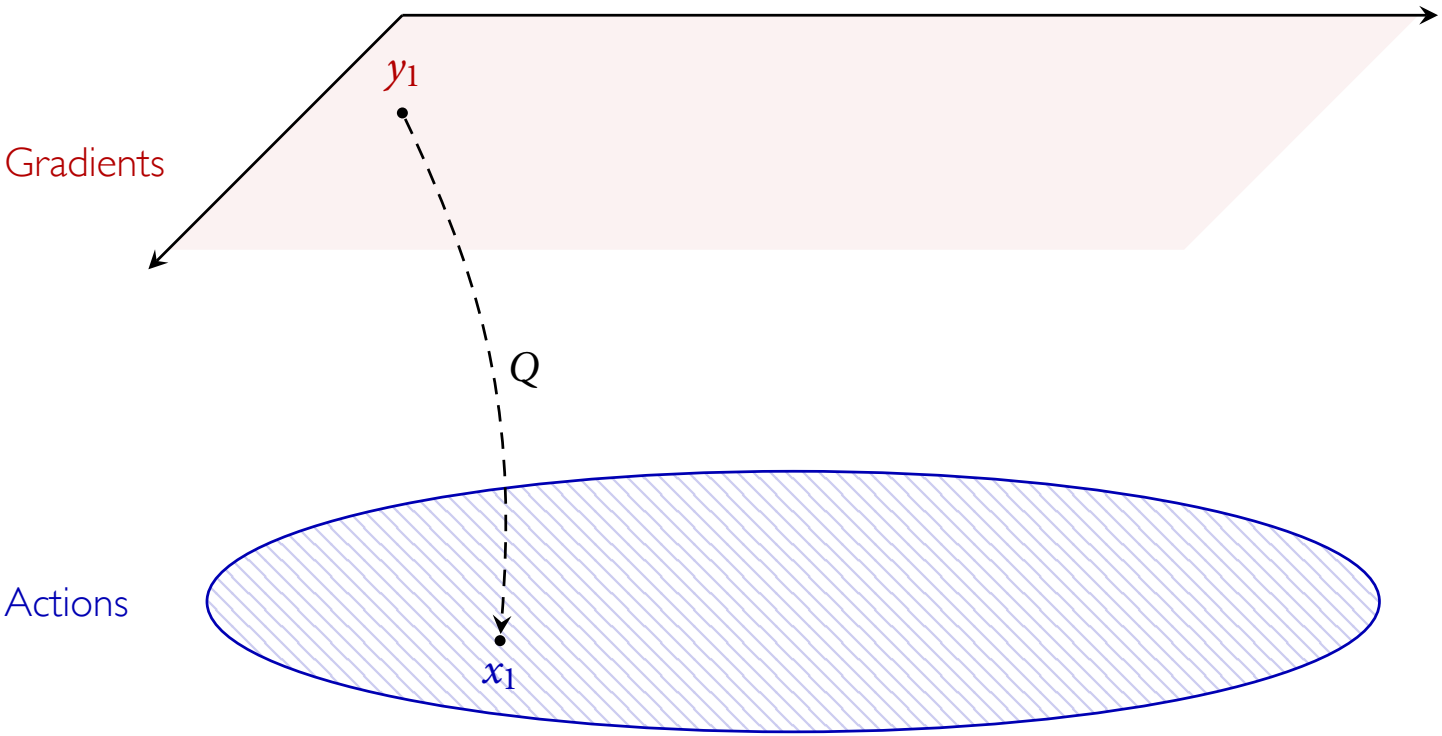
Dual averaging / mirror descent

Take a (lazy) gradient step and *mirror* (too many to list!)



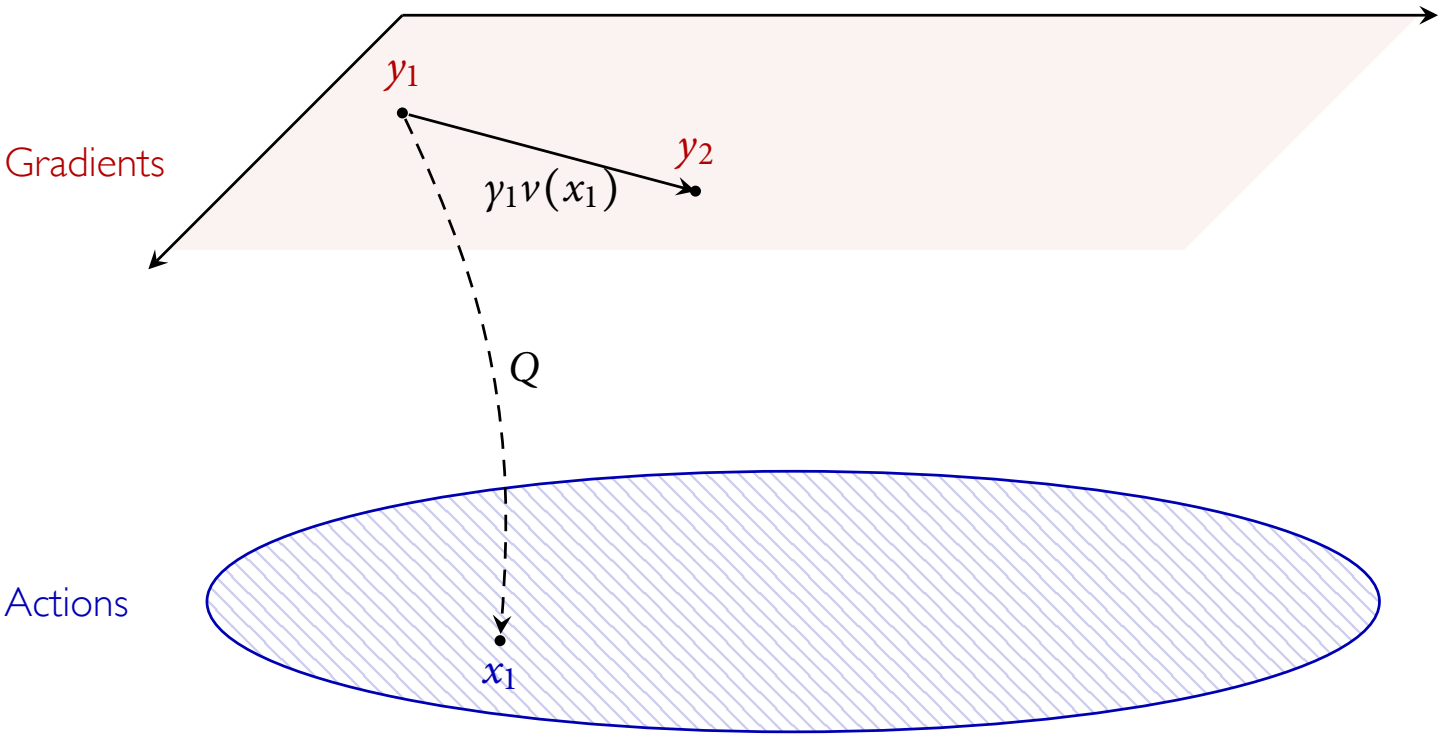
Dual averaging / mirror descent

Take a (lazy) gradient step and *mirror* (too many to list!)



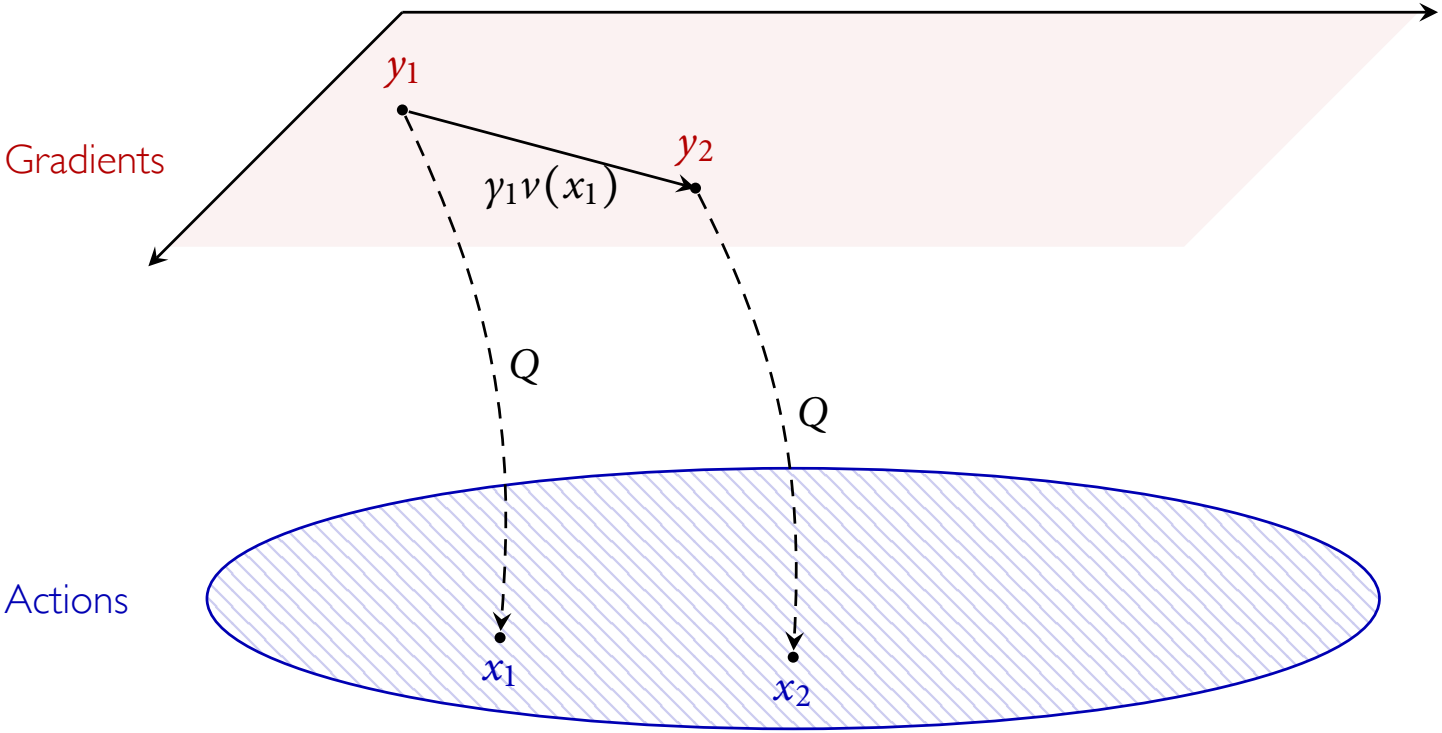
Dual averaging / mirror descent

Take a (lazy) gradient step and *mirror* (too many to list!)



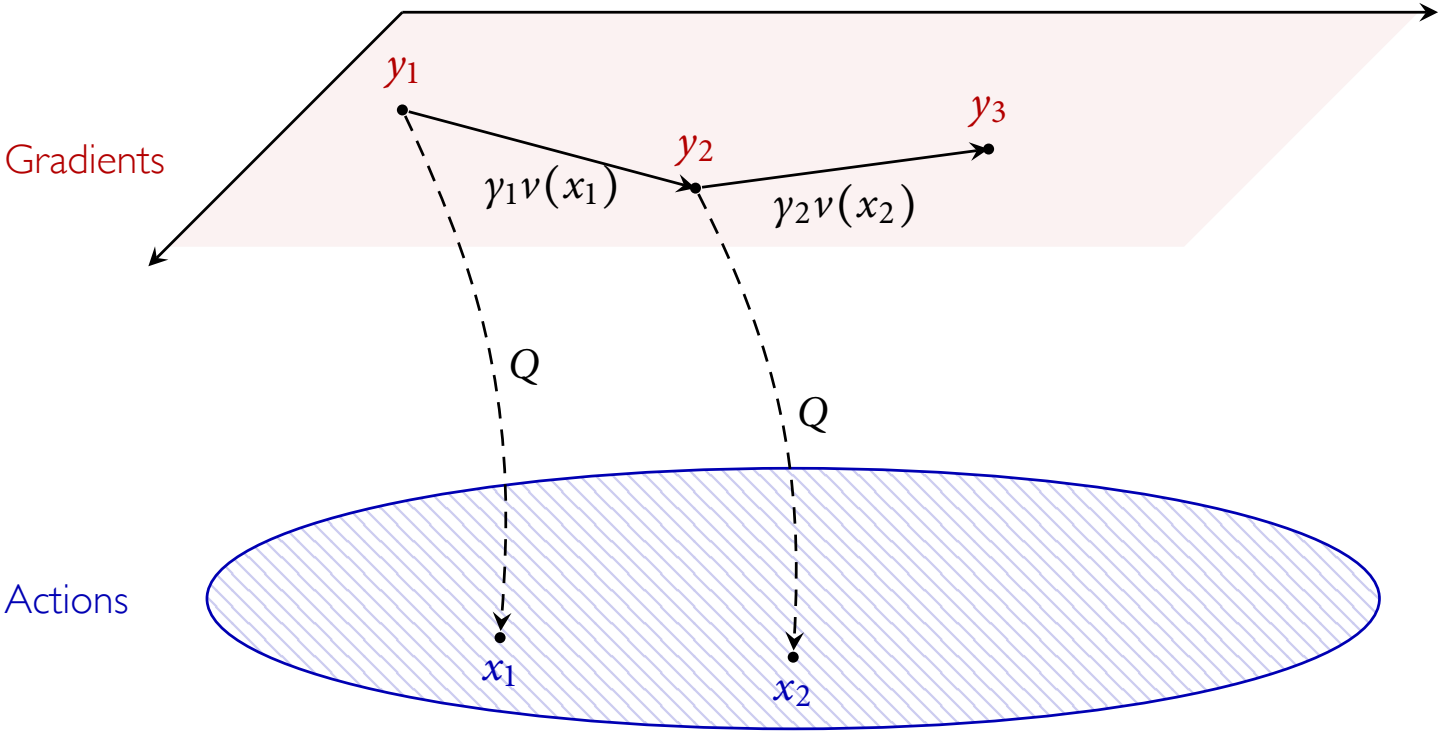
Dual averaging / mirror descent

Take a (lazy) gradient step and *mirror* (too many to list!)



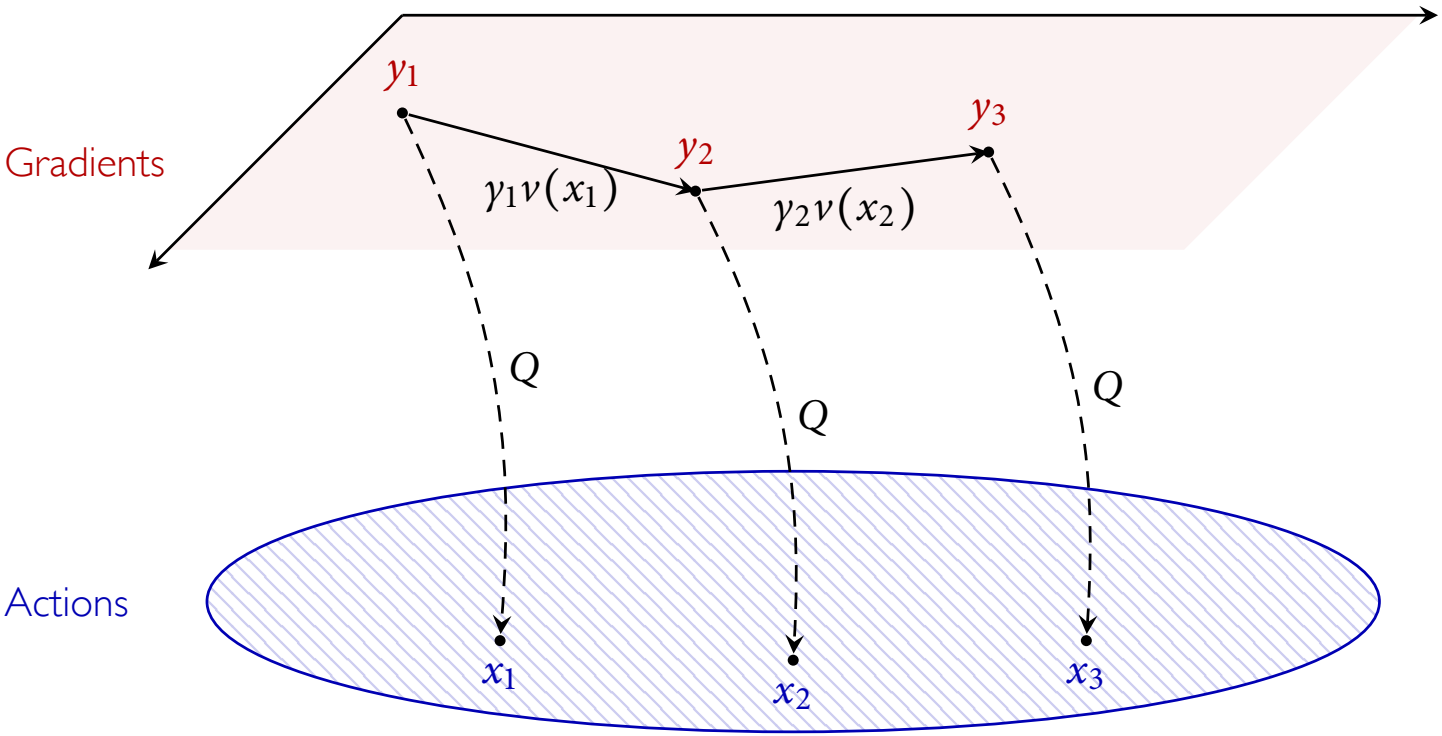
Dual averaging / mirror descent

Take a (lazy) gradient step and *mirror* (too many to list!)



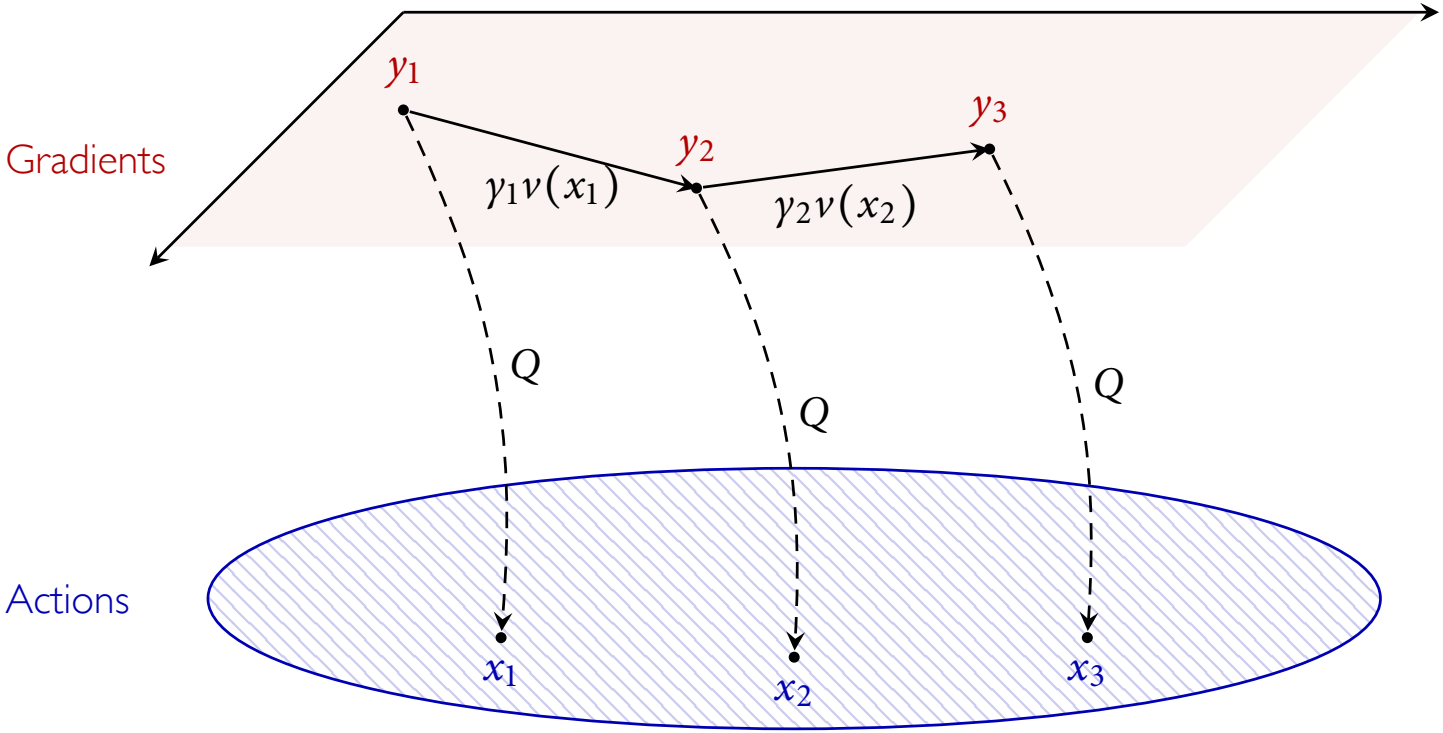
Dual averaging / mirror descent

Take a (lazy) gradient step and *mirror* (too many to list!)



Dual averaging / mirror descent

Take a (lazy) gradient step and *mirror* (too many to list!)



Mirror map:

$$Q(y) = \arg\max\{\langle y, x\rangle - h(x)\}$$

where h is a strongly convex *regularizer* [$h(x) = \|x\|^2/2$ for OGD]

Multi-agent learning via mirror descent

Multi-agent mirror descent:

$$y_{i,n+1} = y_{i,n} + \gamma_n \hat{v}_{i,n}$$
$$x_{i,n+1} = Q_i(y_{i,n+1}) \qquad \text{(MD)}$$

Multi-agent mirror descent

Require: mirror map $Q_i : \mathcal{Y}_i \to \mathcal{X}_i$, step-size sequence $\gamma_n > 0$

Initialize scores y_i, actions $x_i = Q_i(y_i)$
for $n = 1, 2, \ldots$ **do**
  receive reward $u_i(x_1, \ldots, x_N)$ # get payoff
  get estimate $\hat{v}_i$ of payoff gradient $v_i(x)$ # first-order feedback
  $y_i \leftarrow y_i + \gamma_n \hat{v}_i$ # take gradient step
  $x_i \leftarrow Q_i(y_i)$ # update actions
end for

Non-Euclidean example: Exponential weights

Entropic regularization over the simplex ($\mathcal{X} = \Delta(\mathcal{A})$ for some finite set $\mathcal{A}$):

$$h(x) = \sum_{\alpha \in \mathcal{A}} x_\alpha \log x_\alpha$$

Induced mirror map $\rightsquigarrow$ *logit choice:*

$$\Lambda(y) = \arg\max_{x \in \mathcal{X}} \sum_{\alpha \in \mathcal{A}} (x_\alpha y_\alpha - x_\alpha \log x_\alpha) = \frac{(\exp(y_\alpha))_{\alpha \in \mathcal{A}}}{\sum_{\beta \in \mathcal{A}} \exp(y_\beta)}$$

Exponential weights algorithm ***(Freund & Schapire, GEB 1999)***

Require: parameter $\gamma > 0$

set $y_\alpha \leftarrow 0$ for each action $\alpha \in \mathcal{A}$ # initialization
for $n = 1, 2, \ldots$ **do**
  play $x \leftarrow \Lambda(y)$ # logit update
  draw α_n according to x # arm selection
  observe payoff vector v_n and get $\hat{u}_n = v_{n,\alpha_n}$ # payoffs revealed
  set $y \leftarrow y + \gamma v_n$ # score update
end for

Feedback

Perfect gradient information difficult in practice:

- Measurement errors
- Noisy information transmission
- Stochastic utilities (realized vs. expected gradients)
- ...

Imperfect gradient feedback:

$$\hat{v}_n = v(x_n) + U_n$$

with the following hypotheses for U:

(H1) *Zero-mean error:* $\mathbb{E}[U_n \mid \mathcal{F}_{n-1}] = 0$ $\quad [\Longrightarrow \mathbb{E}[\hat{v}_n \mid \mathcal{F}_{n-1}] = v(x_n)]$

(H2) *Finite mean squared error:* $\mathbb{E}[\|U_n\|^2 \mid \mathcal{F}_{n-1}] \leq \sigma^2$ $\quad [\Longrightarrow \mathbb{E}[\|\hat{v}_n\|^2 \mid \mathcal{F}_{n-1}] \leq V^2]$

What is known I: Finite games

- *Optimality/Universality:* MD provides a min-max optimal $\mathcal{O}(T^{1/2})$ regret bound
- Empirical frequency of play converges to *coarse correlated equilibrium*

What is known I: Finite games

- *Optimality/Universality:* MD provides a min-max optimal $\mathcal{O}(T^{1/2})$ regret bound
- Empirical frequency of play converges to *coarse correlated equilibrium*, but:
  - *Actual play may be off-equilibrium*
  - **Dominated strategies may survive in perpetuity** *(Viossat & Zapechelnyuk, JET 2013)*

What is known I: Finite games

- *Optimality/Universality:* MD provides a min-max optimal $\mathcal{O}(T^{1/2})$ regret bound
- Empirical frequency of play converges to *coarse correlated equilibrium*, but:
  - *Actual play may be off-equilibrium*
  - **Dominated strategies may survive in perpetuity** *(Viossat & Zapechelnyuk, JET 2013)*
- In congestion games with perfect information, pure Nash equilibria are *recurrent* *(Kleinberg et al., STOC 2009)*

What is known I: Finite games

- *Optimality/Universality:* MD provides a min-max optimal $\mathcal{O}(T^{1/2})$ regret bound
- Empirical frequency of play converges to *coarse correlated equilibrium*, but:
  - *Actual play may be off-equilibrium*
  - **Dominated strategies may survive in perpetuity** *(Viossat & Zapechelnyuk, JET 2013)*
- In congestion games with perfect information, pure Nash equilibria are *recurrent* *(Kleinberg et al., STOC 2009)*. However:
  - *Non-equilibrium strategies could be played infinitely often*
  - *Can have chaos/aperiodic orbits* *(Palaiopanos et al., NIPS 2017)*

What is known I: Finite games

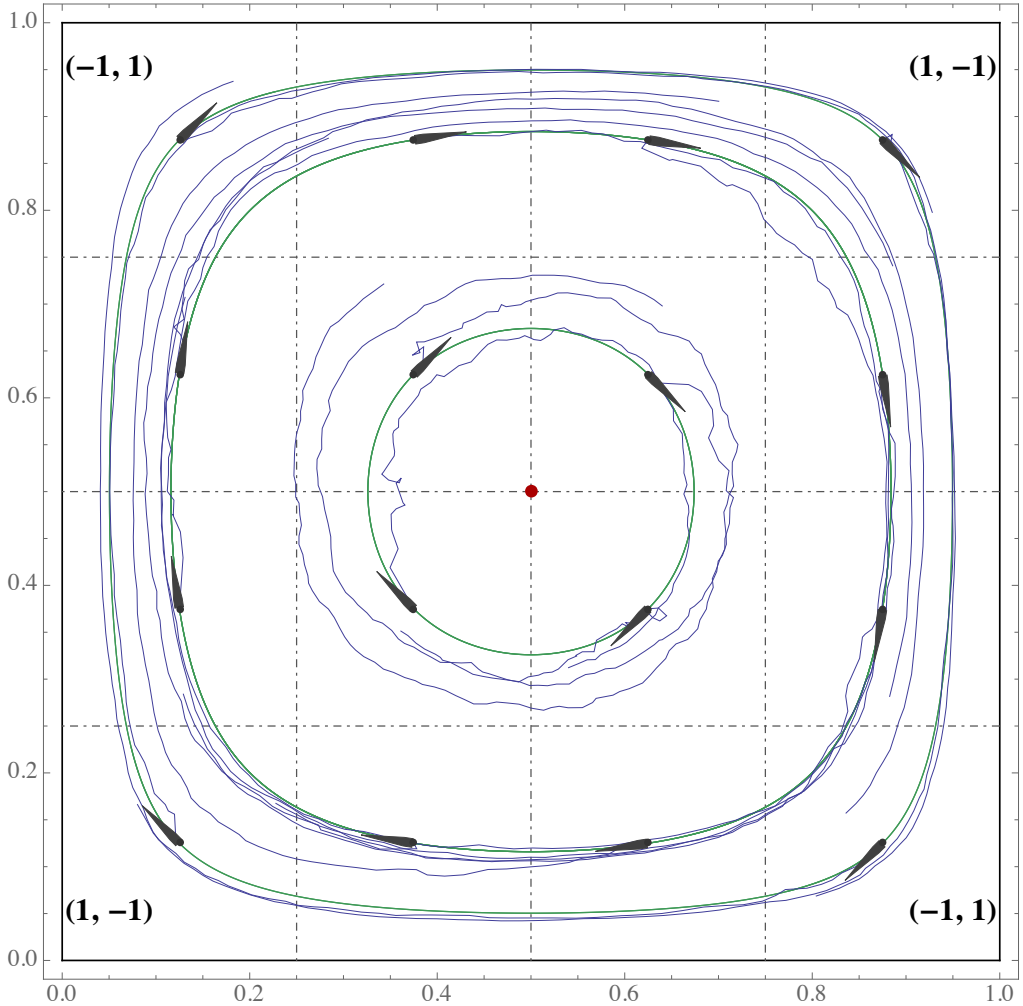
- *Optimality/Universality:* MD provides a min-max optimal $\mathcal{O}(T^{1/2})$ regret bound
- Empirical frequency of play converges to *coarse correlated equilibrium*, but:
  - *Actual play may be off-equilibrium*
  - **Dominated strategies may survive in perpetuity** *(Viossat & Zapechelnyuk, JET 2013)*
- In congestion games with perfect information, pure Nash equilibria are *recurrent* *(Kleinberg et al., STOC 2009)*. However:
  - *Non-equilibrium strategies could be played infinitely often*
  - *Can have chaos/aperiodic orbits* *(Palaiopanos et al., NIPS 2017)*
- In zero-sum games, empirical frequency of play converges

What is known I: Finite games

- *Optimality/Universality:* MD provides a min-max optimal $\mathcal{O}(T^{1/2})$ regret bound
- Empirical frequency of play converges to *coarse correlated equilibrium*, but:
  - *Actual play may be off-equilibrium*
  - **Dominated strategies may survive in perpetuity** *(Viossat & Zapechelnyuk, JET 2013)*
- In congestion games with perfect information, pure Nash equilibria are *recurrent* *(Kleinberg et al., STOC 2009)*. However:
  - *Non-equilibrium strategies could be played infinitely often*
  - *Can have chaos/aperiodic orbits* *(Palaiopanos et al., NIPS 2017)*
- In zero-sum games, empirical frequency of play converges, *but actual play cycles* (Hofbauer, Sorin & Viossat, MOR 2009; M, Piliouras & Papadimitriou, SODA 2018)

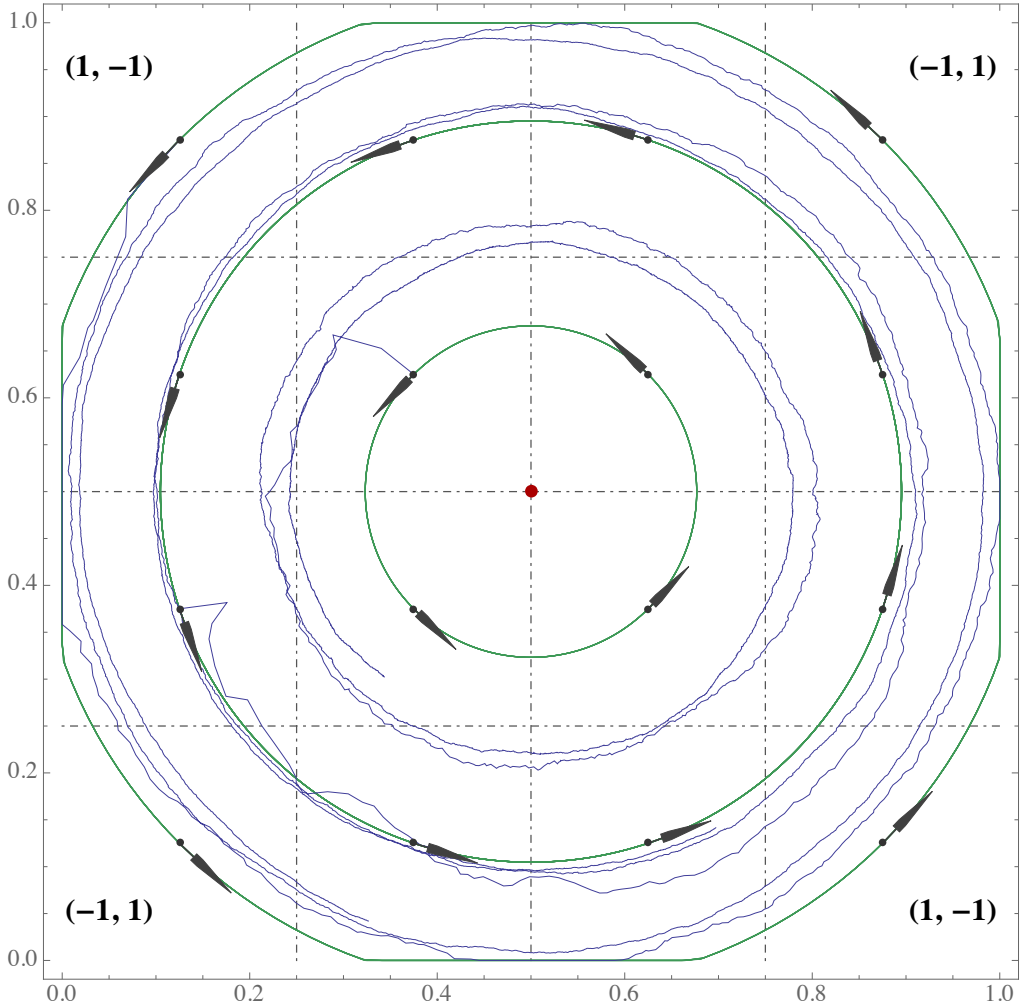
Cycles in adversarial learning

Non-convergence in Matching Pennies



Cycles in adversarial learning

Non-convergence in Matching Pennies



What is known II: Continuous games

- *Optimality/Universality:* MD provides a min-max optimal $\mathcal{O}(T^{1/2})$ regret bound
- In convex-concave zero-sum games, the *ergodic average*

$$\bar{x}_n = \frac{\sum_{k=1}^{n} \gamma_k x_k}{\sum_{k=1}^{n} \gamma_k}$$

converges to Nash equilibrium, *but actual play doesn't* *(Nesterov, MAPR 2009)*

What is known II: Continuous games

- *Optimality/Universality:* MD provides a min-max optimal $\mathcal{O}(T^{1/2})$ regret bound
- In convex-concave zero-sum games, the *ergodic average*

$$\bar{x}_n = \frac{\sum_{k=1}^{n} \gamma_k x_k}{\sum_{k=1}^{n} \gamma_k}$$

converges to Nash equilibrium, *but actual play doesn't* *(Nesterov, MAPR 2009)*

- Similar (ergodic) convergence results in games that admit a *concave potential*, i.e.

$$v_i = \frac{\partial f}{\partial x_i} \quad \text{for some concave function } f: \mathcal{X} \to \mathbb{R}.$$

What is known II: Continuous games

- *Optimality/Universality:* MD provides a min-max optimal $\mathcal{O}(T^{1/2})$ regret bound
- In convex-concave zero-sum games, the *ergodic average*

$$\bar{x}_n = \frac{\sum_{k=1}^{n} \gamma_k x_k}{\sum_{k=1}^{n} \gamma_k}$$

converges to Nash equilibrium, *but actual play doesn't* *(Nesterov, MAPR 2009)*

- Similar (ergodic) convergence results in games that admit a *concave potential*, i.e.

$$v_i = \frac{\partial f}{\partial x_i} \quad \text{for some concave function } f\colon \mathcal{X} \to \mathbb{R}.$$

- *Almost sure convergence of actual play in strictly monotone games (M & Zhou, MAPR 2018)*

Outline

No gradient feedback

In many cases, gradients impossible to compute:

- Multi-armed bandits (clinical trials, …)
- Other players' impact unknown (routing, …)
- Too costly to compute (big data, GANs, …)
- …

No gradient feedback

In many cases, gradients impossible to compute:

- Multi-armed bandits (clinical trials, …)
- Other players' impact unknown (routing, …)
- Too costly to compute (big data, GANs, …)
- …

Possible fixes:

- Two-time-scale approach: fast samples, slow updates [can be slow ☹]

No gradient feedback

In many cases, gradients impossible to compute:

- Multi-armed bandits (clinical trials, …)
- Other players' impact unknown (routing, …)
- Too costly to compute (big data, GANs, …)
- …

Possible fixes:

- Two-time-scale approach: fast samples, slow updates [can be slow ☹]
- Multiple-point estimates *(Bervoets et al., 2016)* [needs synchronization ☹]

No gradient feedback

In many cases, gradients impossible to compute:

- Multi-armed bandits (clinical trials, …)
- Other players' impact unknown (routing, …)
- Too costly to compute (big data, GANs, …)
- …

Possible fixes:

- Two-time-scale approach: fast samples, slow updates [can be slow ☹]
- Multiple-point estimates *(Bervoets et al., 2016)* [needs synchronization ☹]
- *Simultaneous stochastic approximation (Spall, Aut 1997)* [this talk]

Simultaneous stochastic approximation

Estimate $u'(\hat{x})$ at target point $\hat{x}$

$$u'(\hat{x}) \approx \frac{u(\hat{x}+\delta) - u(\hat{x}-\delta)}{2\delta}$$

Simultaneous stochastic approximation

Estimate $u'(\hat{x})$ at target point $\hat{x}$

$$u'(\hat{x}) \approx \frac{u(\hat{x}+\delta) - u(\hat{x}-\delta)}{2\delta}$$

Pick $z = \pm 1$ with probability $1/2$. Then:

$$\mathbb{E}[u(\hat{x}+\delta z)z] = \frac{1}{2}u(\hat{x}+\delta) - \frac{1}{2}u(\hat{x}-\delta)$$

$\Longrightarrow$ Estimate $u'(\hat{x})$ to $\mathcal{O}(\delta)$ by sampling u at $x = \hat{x} + \delta z$ and looking at $\frac{1}{\delta}u(x)z$

Simultaneous stochastic approximation

Estimate $u'(\hat{x})$ at target point $\hat{x}$

$$u'(\hat{x}) \approx \frac{u(\hat{x}+\delta) - u(\hat{x}-\delta)}{2\delta}$$

Pick $z = \pm 1$ with probability $1/2$. Then:

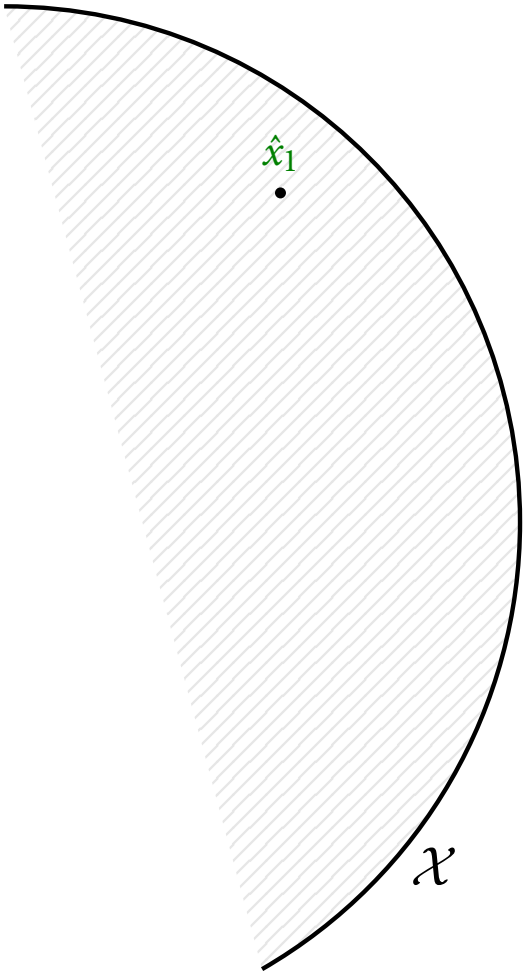
$$\mathbb{E}[u(\hat{x}+\delta z)z] = \frac{1}{2}u(\hat{x}+\delta) - \frac{1}{2}u(\hat{x}-\delta)$$

$\Longrightarrow$ Estimate $u'(\hat{x})$ to $\mathcal{O}(\delta)$ by sampling u at $x = \hat{x} + \delta z$ and looking at $\frac{1}{\delta}u(x)z$

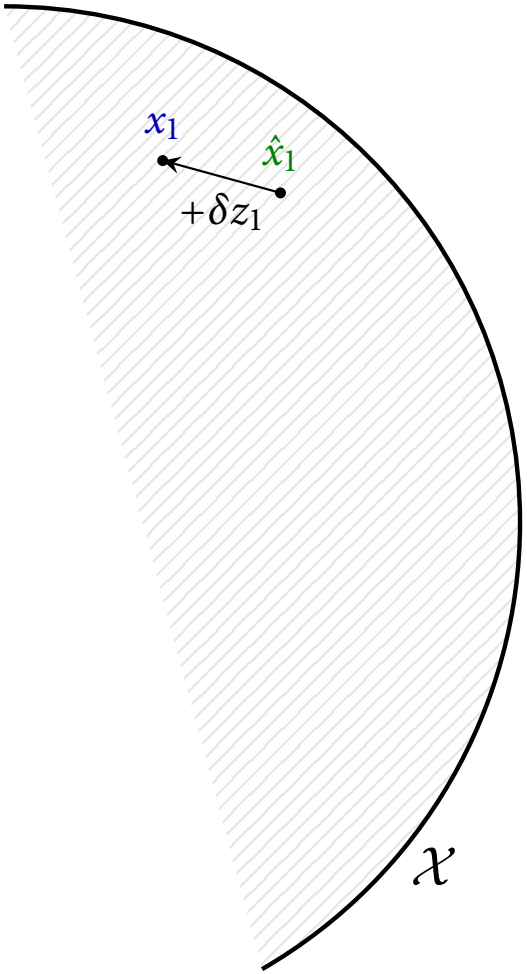
Single-point estimator of ∇u at $\hat{x}$ (general case)

1: Draw z uniformly from $\mathbb{S}^d$
2: Play $x = \hat{x} + \delta z$
3: Get $\hat{u} = u(x)$
4: Set $\hat{v} = \frac{d}{\delta}\hat{u}z$

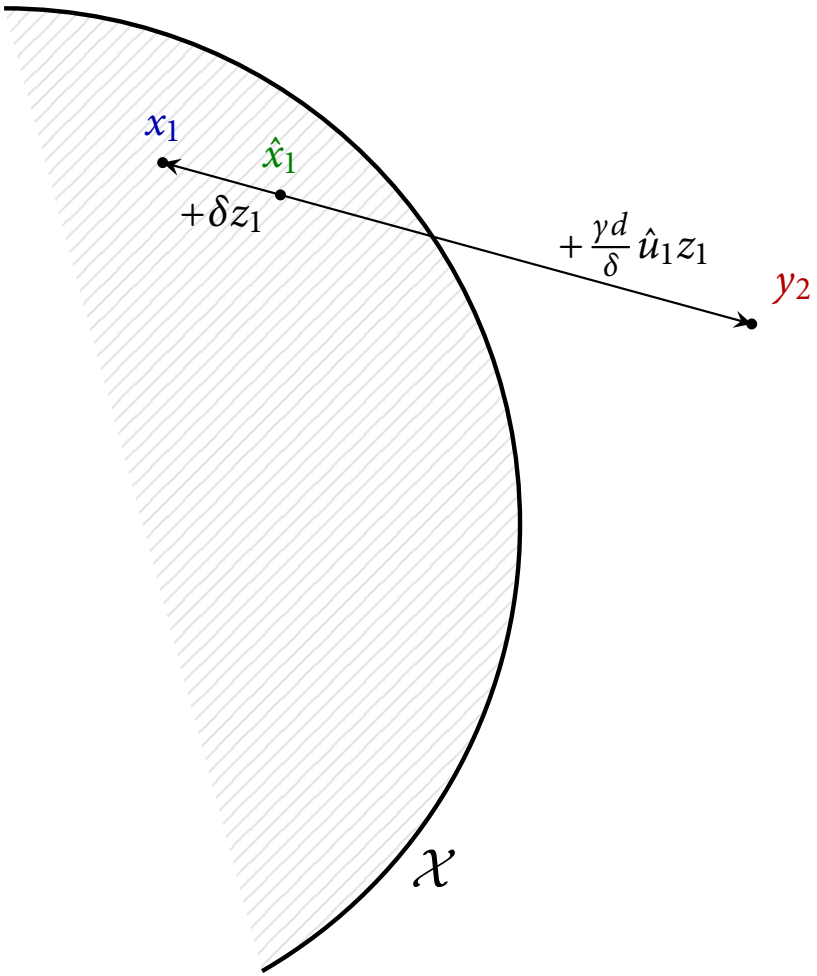
Gradient descent without a gradient



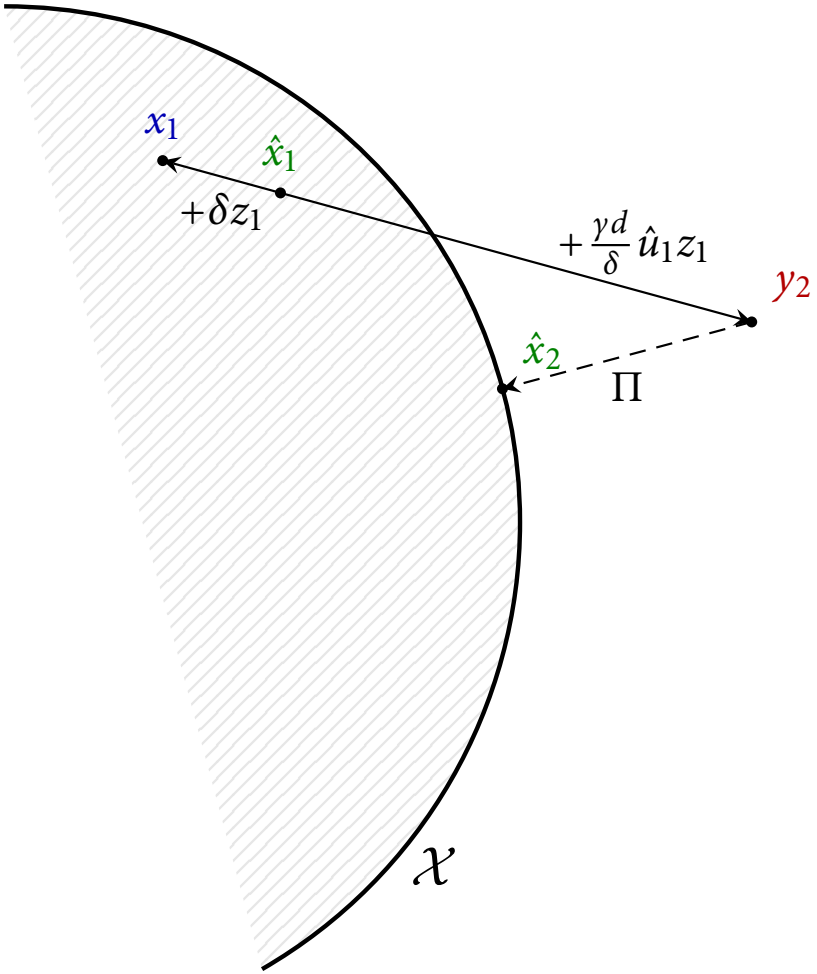
Gradient descent without a gradient



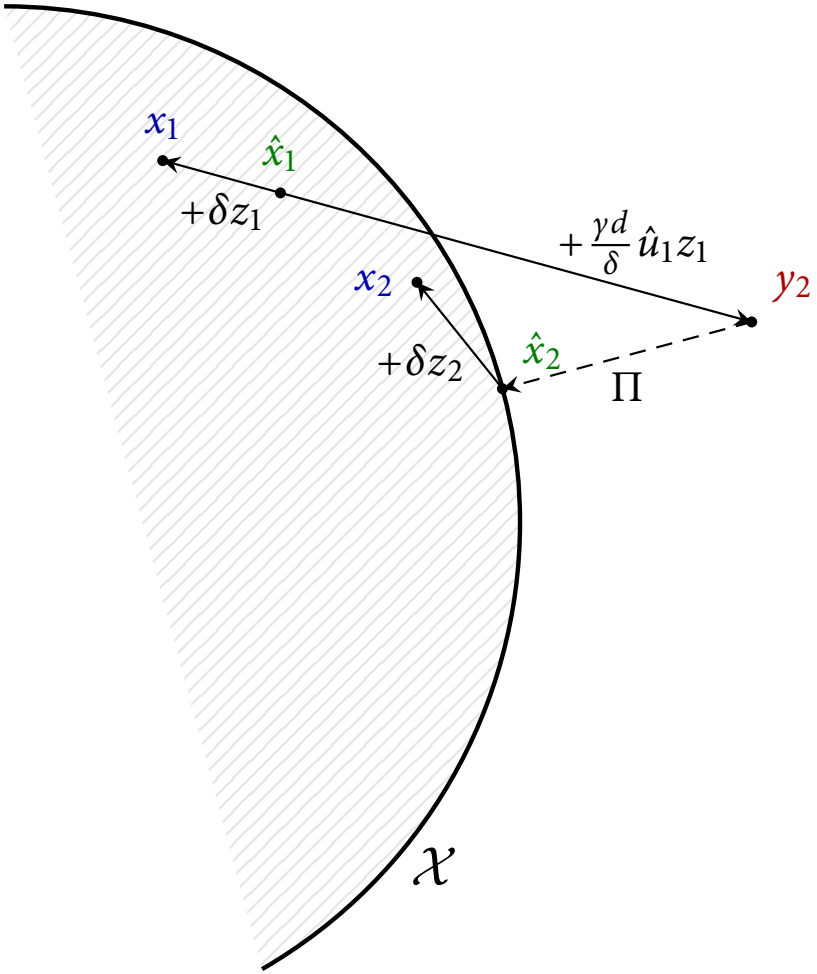
Gradient descent without a gradient



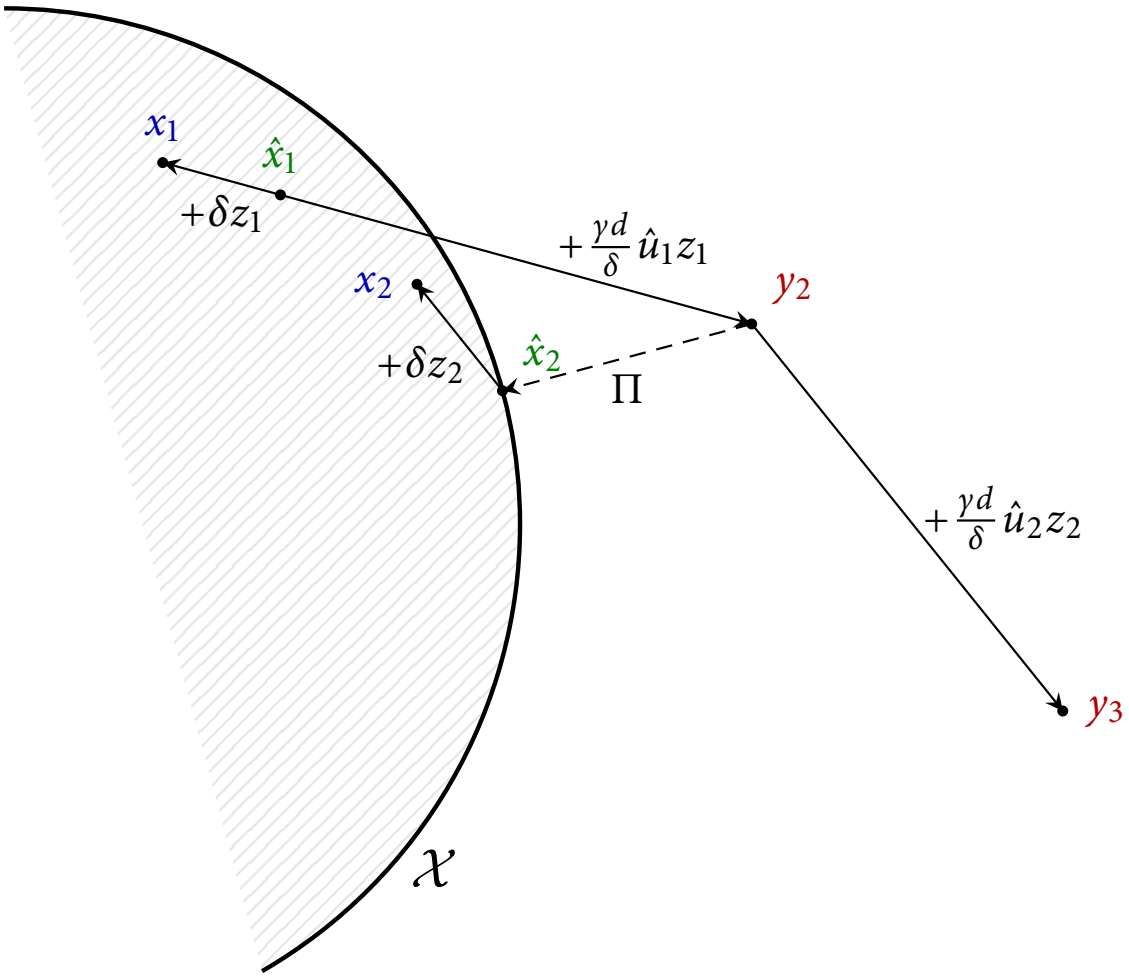
Gradient descent without a gradient



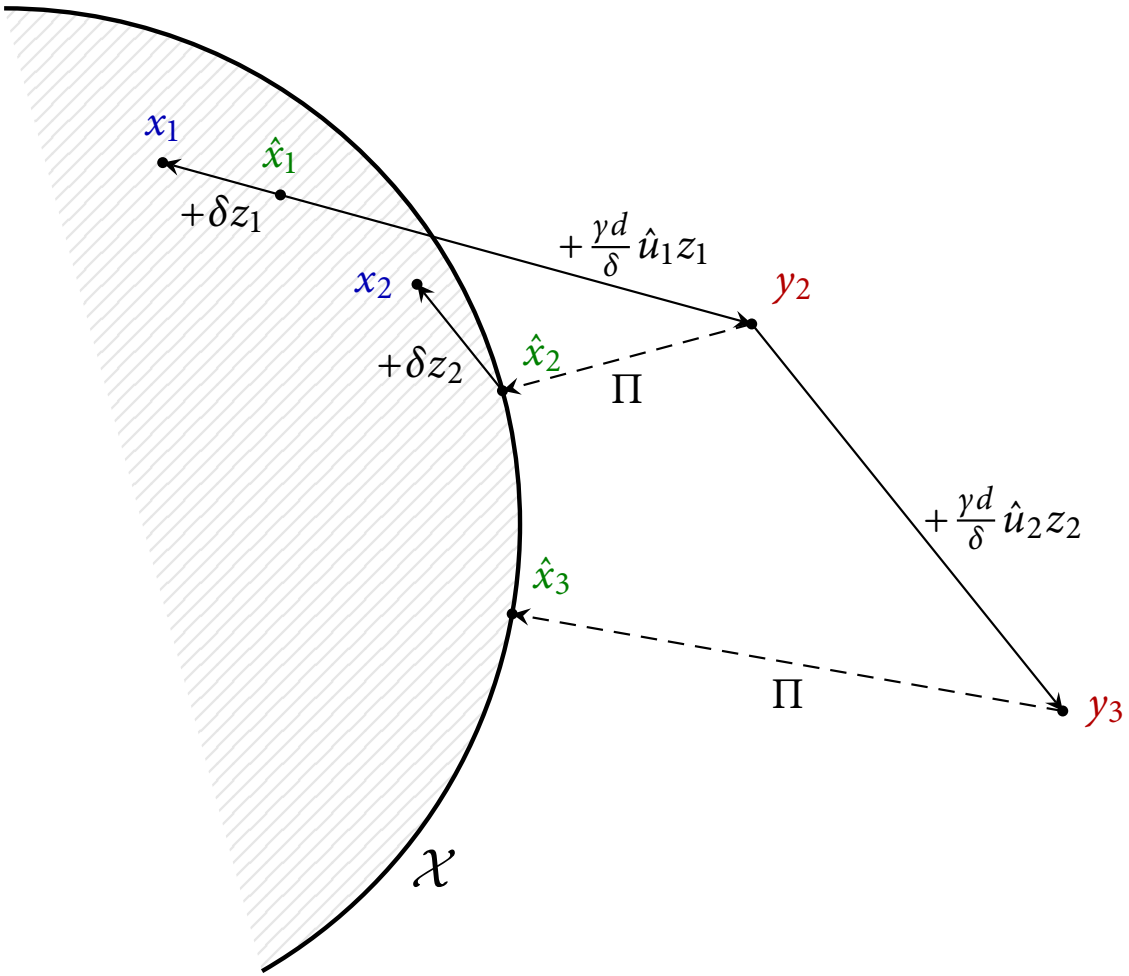
Gradient descent without a gradient



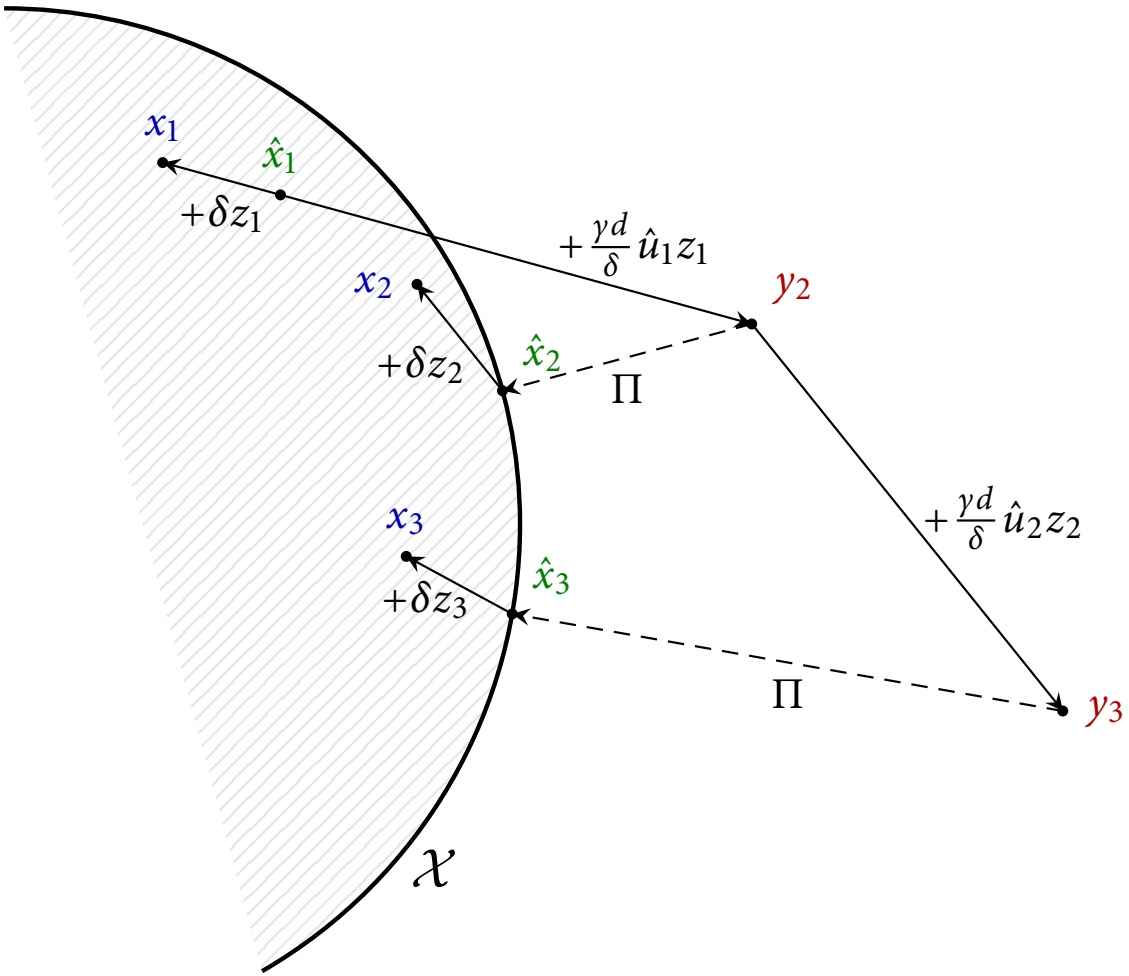
Gradient descent without a gradient



Gradient descent without a gradient



Gradient descent without a gradient



Bandit mirror descent

Mirror descent with bandit feedback

choose pivot point $\hat{x}$ # initialization
repeat At each epoch $n = 1, 2, \dots$
  draw z uniformly from $\mathbb{S}^d$ # exploration direction
  play $x \leftarrow \hat{x} + \delta z$ # choose action
  get $\hat{u} \leftarrow u(x)$ # payoff phase
  set $y \leftarrow y + \frac{\gamma d}{\delta} \hat{u} \cdot z$ # gradient step
  set $\hat{x} \leftarrow Q(y)$ # update pivot
until end

Feasibility adjustment: pivot at

$$\hat{x} - \frac{\delta}{r}(\hat{x} - p)$$

for some interior reference point p such that $\mathbb{B}_r(p) \subseteq \mathcal{X}$

Challenges

Key difficulty:

- One-point estimates are **biased** (no more than $\mathcal{O}(\delta)$ accuracy)

Challenges

Key difficulty:

- One-point estimates are **biased** (no more than $\mathcal{O}(\delta)$ accuracy)
- Can eliminate bias by taking variable $\delta_n \to 0$

Challenges

Key difficulty:

- One-point estimates are **biased** (no more than $\mathcal{O}(\delta)$ accuracy)
- Can eliminate bias by taking variable $\delta_n \to 0$ but in so doing, the **variance explodes**

$$\mathbb{E}[\|\hat{v}\|^2] = \mathcal{O}(1/\delta^2)$$

Challenges

Key difficulty:

- One-point estimates are **biased** (no more than $\mathcal{O}(\delta)$ accuracy)
- Can eliminate bias by taking variable $\delta_n \to 0$ but in so doing, the **variance explodes**

$$\mathbb{E}[\|\hat{v}\|^2] = \mathcal{O}(1/\delta^2)$$

- Stochastic approximation analysis requires *bounded variance*

Challenges

Key difficulty:

- One-point estimates are **biased** (no more than $\mathcal{O}(\delta)$ accuracy)
- Can eliminate bias by taking variable $\delta_n \to 0$ but in so doing, the **variance explodes**

$$\mathbb{E}[\|\hat{v}\|^2] = \mathcal{O}(1/\delta^2)$$

- Stochastic approximation analysis requires *bounded variance*
- The *bias-variance dilemma*: accuracy or stability?

Controlling the variance

Can go for both!

Lemma (Cohen, Héliou & M, 2017)

Consider a Robbins-Monro algorithm of the form

$$x_{n+1} = x_n + \gamma_n [F(x_n) + U_n + b_n], \tag{RM}$$

with bounded Lipschitz F, L^2*-bounded martingale noise* U_n, *and* $b_n \to 0$ (*a.s.*). *If*

$$\sum_{n=1}^{\infty} \gamma_n^{1+q/2} \mathbb{E}[\|U_n\|^q \mid \mathcal{F}_{n-1}] < \infty$$

for some $q \geq 2$, *then* (RM) *is an asymptotic pseudotrajectory* (*APT*) *of the mean dynamics*

$$\dot{x} = F(x). \tag{MD}$$

APT: asymptotically follow (MD) with arbitrary accuracy over windows of arbitrary length.

Convergence analysis

Must balance step-size γ_n against exploration factor δ_n:

- $\lim_{n\to\infty} \gamma_n = \lim_{n\to\infty} \delta_n = 0$ # eliminate noise and bias
- $\sum_{n=1}^{\infty} = \infty$ # the process doesn't stop
- $\sum_{n=1}^{\infty} \gamma_n^2/\delta_n^2 < \infty$ # ensure APT property
- $\lim_{n\to\infty} \gamma_n/\delta_n^2 = 0$ # allow for tracking

Convergence analysis

Must balance step-size γ_n against exploration factor δ_n:

- $\lim_{n\to\infty} \gamma_n = \lim_{n\to\infty} \delta_n = 0$ # eliminate noise and bias
- $\sum_{n=1}^{\infty} = \infty$ # the process doesn't stop
- $\sum_{n=1}^{\infty} \gamma_n^2/\delta_n^2 < \infty$ # ensure APT property
- $\lim_{n\to\infty} \gamma_n/\delta_n^2 = 0$ # allow for tracking

Theorem (Bravo, Leslie & M, 2017)

In strictly monotone games, the sequence of play induced by (MD) *with bandit feedback and step-size/sampling parameters as above converges to Nash equilibrium* (*a.s.*).

Proof

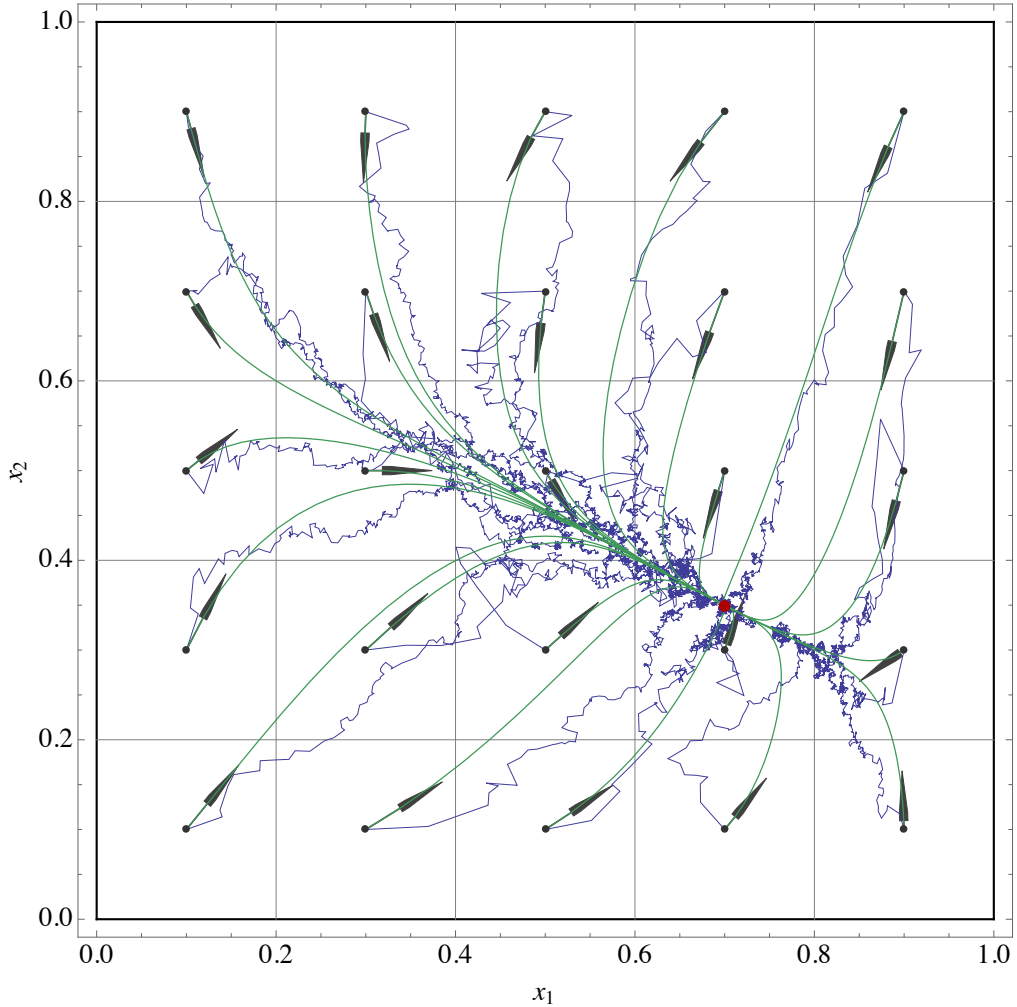
Sketch of proof.

- *Recurrence:* x_n visits any neighborhood of x^* infinitely often (a.s.) *(LLN)*
- Consider continuous-time dynamics $\dot{y} = v(Q(y))$
- *Convergence in continuous time:* $Q(y(t))$ converges to x^* uniformly on compacts
- y_n is (a.s.) an *asymptotic pseudotrajectory* of continuous dynamics *(lemma)*
- Recurrence + uniform convergence + APT $\Longrightarrow$ $\lim_{n\to\infty} x_n = x^*$ **(hard)**

□

Bandit replicator dynamics

Replicator dynamics with bandit feedback in a Cournot duopoly



Conclusions and perspectives

Conclusions

- No-regret learning does not guarantee convergence to equilibrium
- Monotonicity does.
- Convergence does not require gradient information.

Conclusions and perspectives

Conclusions

- No-regret learning does not guarantee convergence to equilibrium
- Monotonicity does.
- Convergence does not require gradient information.

Open questions

- Convergence rate?
- Feedback delays / asynchronous updates? [Partial answer by Zhou et al., NIPS '17]
- **???**