Multi-agent learning

No-regret learning

Gerard Vreeswijk, Intelligent Software Systems, Computer Science Department, Faculty of Sciences, Utrecht University, The Netherlands.

Wednesday 19th May, 2021

No-regret learning: motivation

No-regret learning: motivation

- **Reinforcement Learning**. *Play those actions that were successful in the past*.

No-regret learning: motivation

- **Reinforcement Learning**. *Play those actions that were successful in the past*.

- **No-regret learning**: might be considered as an extension of reinforcement learning. *Play those actions that would have been successful in the past*.

No-regret learning: motivation

- **Reinforcement Learning**. *Play those actions that were successful in the past.*

- *Similarities*:

- **No-regret learning**: might be considered as an extension of reinforcement learning. *Play those actions that would have been successful in the past.*

No-regret learning: motivation

- **Reinforcement Learning**. *Play those actions that were successful in the past*.

- *Similarities*:

  1. Driven by past payoffs.

- **No-regret learning**: might be considered as an extension of reinforcement learning. *Play those actions that would have been successful in the past*.

No-regret learning: motivation

- **Reinforcement Learning**. *Play those actions that were successful in the past*.
- *Similarities*:
  1. Driven by past payoffs.
  2. Not interested in (the behaviour of) the opponent.
- **No-regret learning**: might be considered as an extension of reinforcement learning. *Play those actions that would have been successful in the past*.

No-regret learning: motivation

- **Reinforcement Learning**. *Play those actions that were successful in the past*.
- *Similarities*:
  1. Driven by past payoffs.
  2. Not interested in (the behaviour of) the opponent.
  3. Probabilistic.
- **No-regret learning**: might be considered as an extension of reinforcement learning. *Play those actions that would have been successful in the past*.

No-regret learning: motivation

- **Reinforcement Learning**. *Play those actions that were successful in the past*.
- *Similarities*:
  1. Driven by past payoffs.
  2. Not interested in (the behaviour of) the opponent.
  3. Probabilistic.
  4. Smooth adaptation.
- **No-regret learning**: might be considered as an extension of reinforcement learning. *Play those actions that would have been successful in the past*.

No-regret learning: motivation

- **Reinforcement Learning**. *Play those actions that were successful in the past*.
- *Similarities*:
  1. Driven by past payoffs.
  2. Not interested in (the behaviour of) the opponent.
  3. Probabilistic.
  4. Smooth adaptation.
  5. Myopic.
- **No-regret learning**: might be considered as an extension of reinforcement learning. *Play those actions that would have been successful in the past*.

No-regret learning: motivation

- **Reinforcement Learning**. *Play those actions that were successful in the past*.
- **No-regret learning**: might be considered as an extension of reinforcement learning. *Play those actions that would have been successful in the past*.
- *Similarities*:
  1. Driven by past payoffs.
  2. Not interested in (the behaviour of) the opponent.
  3. Probabilistic.
  4. Smooth adaptation.
  5. Myopic.
- *Differences*:

No-regret learning: motivation

- **Reinforcement Learning**. *Play those actions that were successful in the past*.

- *Similarities*:

  1. Driven by past payoffs.
  2. Not interested in (the behaviour of) the opponent.
  3. Probabilistic.
  4. Smooth adaptation.
  5. Myopic.

- **No-regret learning**: might be considered as an extension of reinforcement learning. *Play those actions that would have been successful in the past*.

- *Differences*:

  1. Keeping counts of hypothetical actions rests on the assumption that a player is able to estimate payoffs of actions that were actually not played.

No-regret learning: motivation

- **Reinforcement Learning**. *Play those actions that were successful in the past*.

- *Similarities*:

  1. Driven by past payoffs.
  2. Not interested in (the behaviour of) the opponent.
  3. Probabilistic.
  4. Smooth adaptation.
  5. Myopic.

- **No-regret learning**: might be considered as an extension of reinforcement learning. *Play those actions that would have been successful in the past*.

- *Differences*:

  1. Keeping counts of hypothetical actions rests on the assumption that a player is able to estimate payoffs of actions that were actually not played.

     (Knowledge of the payoff matrix helps, but is a stronger assumption.)

No-regret learning: motivation

- **Reinforcement Learning**. *Play those actions that were successful in the past*.
- *Similarities*:
  1. Driven by past payoffs.
  2. Not interested in (the behaviour of) the opponent.
  3. Probabilistic.
  4. Smooth adaptation.
  5. Myopic.
- **No-regret learning**: might be considered as an extension of reinforcement learning. *Play those actions that would have been successful in the past*.
- *Differences*:
  1. Keeping counts of hypothetical actions rests on the assumption that a player is able to estimate payoffs of actions that were actually not played.

     (Knowledge of the payoff matrix helps, but is a stronger assumption.)
  2. It is more easy to obtain results regarding performance.

No-regret learning: motivation

- **Reinforcement Learning**. *Play those actions that were successful in the past*.
- *Similarities*:
  1. Driven by past payoffs.
  2. Not interested in (the behaviour of) the opponent.
  3. Probabilistic.
  4. Smooth adaptation.
  5. Myopic.
- **No-regret learning**: might be considered as an extension of reinforcement learning. *Play those actions that would have been successful in the past*.
- *Differences*:
  1. Keeping counts of hypothetical actions rests on the assumption that a player is able to estimate payoffs of actions that were actually not played.

     (Knowledge of the payoff matrix helps, but is a stronger assumption.)
  2. It is more easy to obtain results regarding performance. (*Correlated equilibrium*.)

Qualitative features of reinforcement and regret

Qualitative features of reinforcement and regret

1. **Probabilistic choice**. A choice of action is never completely determined by history but has a random component.

Qualitative features of reinforcement and regret

1. **Probabilistic choice**. A choice of action is never completely determined by history but has a random component.
   - The randomness ensures exploration.

Qualitative features of reinforcement and regret

1. **Probabilistic choice**. A choice of action is never completely determined by history but has a random component.

   - The randomness ensures exploration.
   - The different magnitudes of the probabilities (arisen through experience) ensures exploitation of past experience.

Qualitative features of reinforcement and regret

1. **Probabilistic choice**. A choice of action is never completely determined by history but has a random component.

   - The randomness ensures exploration.
   - The different magnitudes of the probabilities (arisen through experience) ensures exploitation of past experience.

2. **Smooth adaptation**. The strategy of play changes gradually.

Qualitative features of reinforcement and regret

1. **Probabilistic choice**. A choice of action is never completely determined by history but has a random component.
   - The randomness ensures exploration.
   - The different magnitudes of the probabilities (arisen through experience) ensures exploitation of past experience.
2. **Smooth adaptation**. The strategy of play changes gradually.
   - No-regret learning. Select a pure strategy that would have been most successful, given past play.

Qualitative features of reinforcement and regret

1. **Probabilistic choice**. A choice of action is never completely determined by history but has a random component.
   - The randomness ensures exploration.
   - The different magnitudes of the probabilities (arisen through experience) ensures exploitation of past experience.
2. **Smooth adaptation**. The strategy of play changes gradually.
   - No-regret learning. Select a pure strategy that would have been most successful, given past play.
   - Smoothed fictitious play. Give a soft-max response to the (recent) empirical frequency of opponents' actions.

Qualitative features of reinforcement and regret

1. **Probabilistic choice**. A choice of action is never completely determined by history but has a random component.
   - The randomness ensures exploration.
   - The different magnitudes of the probabilities (arisen through experience) ensures exploitation of past experience.
2. **Smooth adaptation**. The strategy of play changes gradually.
   - No-regret learning. Select a pure strategy that would have been most successful, given past play.
   - Smoothed fictitious play. Give a soft-max response to the (recent) empirical frequency of opponents' actions.
   - Hypothesis testing with smoothed best responses. Give a best response to maintained beliefs about *patterns of play*.

Plan for today

Three parts.

Plan for today

Three parts.

1. **Basic concepts**.

Plan for today

Three parts.

1. **Basic concepts**.

2. **Proportional regret matching**. Hart and Mas-Colell (2000).

Plan for today

Three parts.

1. **Basic concepts**.
2. **Proportional regret matching**. Hart and Mas-Colell (2000).
3. ϵ**-Greedy off-policy regret matching**. Foster and Vohra (1999).

Plan for today

Three parts.

1. **Basic concepts**.
2. **Proportional regret matching**. Hart and Mas-Colell (2000).
3. **ϵ-Greedy off-policy regret matching**. Foster and Vohra (1999).

This presentation almost exclusively follows the second half of Ch. 2 of (Peyton Young, 2004).

Plan for today

Three parts.

1. **Basic concepts**.

2. **Proportional regret matching**. Hart and Mas-Colell (2000).

3. **ϵ-Greedy off-policy regret matching**. Foster and Vohra (1999).

This presentation almost exclusively follows the second half of Ch. 2 of (Peyton Young, 2004).

Peyton Young, H. (2004): *Strategic Learning and it Limits*, Oxford UP. Ch. 2: "Reinforcement and Regret"

Plan for today

Three parts.

1. **Basic concepts**.
2. **Proportional regret matching**. Hart and Mas-Colell (2000).
3. ϵ**-Greedy off-policy regret matching**. Foster and Vohra (1999).

This presentation almost exclusively follows the second half of Ch. 2 of (Peyton Young, 2004).

Peyton Young, H. (2004): *Strategic Learning and it Limits*, Oxford UP. Ch. 2: "Reinforcement and Regret"

Foster, D., and Vohra, R. (1999). "Regret in the on-line decision problem". *Games and Economic Behavior*, **29**, pp. 7-36.

Plan for today

Three parts.

1. **Basic concepts**.
2. **Proportional regret matching**. Hart and Mas-Colell (2000).
3. ϵ**-Greedy off-policy regret matching**. Foster and Vohra (1999).

This presentation almost exclusively follows the second half of Ch. 2 of (Peyton Young, 2004).

Peyton Young, H. (2004): *Strategic Learning and it Limits*, Oxford UP. Ch. 2: "Reinforcement and Regret"

Foster, D., and Vohra, R. (1999). "Regret in the on-line decision problem". *Games and Economic Behavior*, **29**, pp. 7-36.

Hart, S., and Mas-Colell, A. (2000). "A simple adaptive procedure leading to correlated equilibrium". *Econometrica*, **68**, pp. 1127-1150.

Part I: Basic concepts

No-regret: example

Payoffs Player A	0	0	0	1	1	0	0	0	1	0	0	
Actions Player A	L	R	L	L	R	R	L	R	R	R	R	?
Actions Player B	R	L	R	L	R	L	R	L	R	L	L	?

No-regret: example

Payoffs Player A	0	0	0	1	1	0	0	0	1	0	0	
Actions Player A	L	R	L	L	R	R	L	R	R	R	R	?
Actions Player B	R	L	R	L	R	L	R	L	R	L	L	?

- Suppose A is offered to replay the first 11 periods, under the proviso that he must play one pure strategy (i.e., action) throughout.

No-regret: example

Payoffs Player A	0	0	0	1	1	0	0	0	1	0	0	
Actions Player A	L	R	L	L	R	R	L	R	R	R	R	?
Actions Player B	R	L	R	L	R	L	R	L	R	L	L	?

- Suppose A is offered to replay the first 11 periods, under the proviso that he must play one pure strategy (i.e., action) throughout.
-

	Payoff	*Regret*	*Average regret*
Rounds 1-11:	3		
Had L played:	6	$6 - 3$	$(6-3)/11$
Had R played:	5	$5 - 3$	$(5-3)/11$

No-regret: example

Payoffs Player A	0	0	0	1	1	0	0	0	1	0	0	
Actions Player A	L	R	L	L	R	R	L	R	R	R	R	?
Actions Player B	R	L	R	L	R	L	R	L	R	L	L	?

- Suppose A is offered to replay the first 11 periods, under the proviso that he must play one pure strategy (i.e., action) throughout.
-

	Payoff	*Regret*	*Average regret*
Rounds 1-11:	3		
Had L played:	6	$6-3$	$(6-3)/11$
Had R played:	5	$5-3$	$(5-3)/11$

- It is ignored that B likely would have played different if he knew A would have played different.

No-regret: example

Payoffs Player A	0	0	0	1	1	0	0	0	1	0	0	
Actions Player A	L	R	L	L	R	R	L	R	R	R	R	?
Actions Player B	R	L	R	L	R	L	R	L	R	L	L	?

- Suppose A is offered to replay the first 11 periods, under the proviso that he must play one pure strategy (i.e., action) throughout.
-

	Payoff	*Regret*	*Average regret*
Rounds 1-11:	3		
Had L played:	6	$6-3$	$(6-3)/11$
Had R played:	5	$5-3$	$(5-3)/11$

- It is ignored that B likely would have played different if he knew A would have played different.

  So no-regret does not take the interactive nature of play into account.

No-regret: some notation

No-regret: some notation

- The average payoff up to and including round t is

$$\bar{u}^t =_{Def} \frac{1}{t} \sum_{s=1}^{t} u(x^s, y^s).$$

No-regret: some notation

- The average payoff up to and including round t is

$$\bar{u}^t =_{Def} \frac{1}{t} \sum_{s=1}^{t} u(x^s, y^s).$$

- For each action x, the hypothetical average payoff for playing x is

$$\bar{h}_x^t =_{Def} \frac{1}{t} \sum_{s=1}^{t} u(x, y^s).$$

No-regret: some notation

- The average payoff up to and including round t is

$$\bar{u}^t =_{Def} \frac{1}{t} \sum_{s=1}^{t} u(x^s, y^s).$$

- For each action x, the hypothetical average payoff for playing x is

$$\bar{h}_x^t =_{Def} \frac{1}{t} \sum_{s=1}^{t} u(x, y^s).$$

- For each action x, the average regret from not having played x is

$$\bar{r}_x^t =_{Def} \bar{h}_x^t - \bar{u}^t.$$

No-regret: some notation

- The average payoff up to and including round t is

$$\bar{u}^t =_{Def} \frac{1}{t}\sum_{s=1}^{t} u(x^s, y^s).$$

- For each action x, the hypothetical average payoff for playing x is

$$\bar{h}_x^t =_{Def} \frac{1}{t}\sum_{s=1}^{t} u(x, y^s).$$

- For each action x, the average regret from not having played x is

$$\bar{r}_x^t =_{Def} \bar{h}_x^t - \bar{u}^t.$$

- Average regret may be represented as a vector

$$\bar{r}^t =_{Def} (\bar{r}_1^t, \dots, \bar{r}_k^t)^T.$$

No-regret: some notation

- The average payoff up to and including round t is

$$\bar{u}^t =_{Def} \frac{1}{t}\sum_{s=1}^{t} u(x^s, y^s).$$

- For each action x, the hypothetical average payoff for playing x is

$$\bar{h}_x^t =_{Def} \frac{1}{t}\sum_{s=1}^{t} u(x, y^s).$$

- For each action x, the average regret from not having played x is

$$\bar{r}_x^t =_{Def} \bar{h}_x^t - \bar{u}^t.$$

- Average regret may be represented as a vector

$$\bar{r}^t =_{Def} (\bar{r}_1^t, \dots, \bar{r}_k^t)^T.$$

- A given realisation of play

$$\omega = (x_1, y_1), \dots, (x_t, y_t), \dots$$

is said to have no regret if, for all actions x,

No-regret: some notation

- The average payoff up to and including round t is

$$\bar{u}^t =_{Def} \frac{1}{t}\sum_{s=1}^{t} u(x^s, y^s).$$

- For each action x, the hypothetical average payoff for playing x is

$$\bar{h}_x^t =_{Def} \frac{1}{t}\sum_{s=1}^{t} u(x, y^s).$$

- For each action x, the average regret from not having played x is

$$\bar{r}_x^t =_{Def} \bar{h}_x^t - \bar{u}^t.$$

- Average regret may be represented as a vector

$$\bar{r}^t =_{Def} (\bar{r}_1^t, \dots, \bar{r}_k^t)^T.$$

- A given realisation of play

$$\omega = (x_1, y_1), \dots, (x_t, y_t), \dots$$

is said to have no regret if, for all actions x,

$$\limsup_{t\to\infty} \bar{r}_x^t(\omega) \le 0$$

No-regret: some notation

- The average payoff up to and including round t is

$$\bar{u}^t =_{Def} \frac{1}{t}\sum_{s=1}^{t} u(x^s, y^s).$$

- For each action x, the hypothetical average payoff for playing x is

$$\bar{h}_x^t =_{Def} \frac{1}{t}\sum_{s=1}^{t} u(x, y^s).$$

- For each action x, the average regret from not having played x is

$$\bar{r}_x^t =_{Def} \bar{h}_x^t - \bar{u}^t.$$

- Average regret may be represented as a vector

$$\bar{r}^t =_{Def} (\bar{r}_1^t, \dots, \bar{r}_k^t)^T.$$

- A given realisation of play

$$\omega = (x_1, y_1), \dots, (x_t, y_t), \dots$$

is said to have no regret if, for all actions x,

$$\limsup_{t\to\infty} \bar{r}_x^t(\omega) \leq 0$$

i.e. $\lim_{T\to\infty} \sup\{\, \bar{r}_x^t(\omega) \mid T \leq t \,\} \leq 0$

No-regret: some notation

- The average payoff up to and including round t is

$$\bar{u}^t =_{Def} \frac{1}{t}\sum_{s=1}^{t} u(x^s, y^s).$$

- For each action x, the hypothetical average payoff for playing x is

$$\bar{h}^t_x =_{Def} \frac{1}{t}\sum_{s=1}^{t} u(x, y^s).$$

- For each action x, the average regret from not having played x is

$$\bar{r}^t_x =_{Def} \bar{h}^t_x - \bar{u}^t.$$

- Average regret may be represented as a vector

$$\bar{r}^t =_{Def} (\bar{r}^t_1, \dots, \bar{r}^t_k)^T.$$

- A given realisation of play

$$\omega = (x_1, y_1), \dots, (x_t, y_t), \dots$$

is said to have no regret if, for all actions x,

$$\limsup_{t\to\infty} \bar{r}^t_x(\omega) \le 0$$

$$\text{i.e.} \quad \lim_{T\to\infty} \sup\{\, \bar{r}^t_x(\omega) \mid T \le t \,\} \le 0$$

$$\Leftrightarrow \quad \lim_{t\to\infty} [\, \bar{r}^t_x(\omega)\,]_+ = 0.$$

Part II: proportional regret matching

The strategy of proportional regret matching

A strategy $g : H \to \Delta(X)$ is said to have no regret if almost all of its realisations of play have no regret.

The strategy of proportional regret matching

A strategy $g : H \to \Delta(X)$ is said to have no regret if almost all of its realisations of play have no regret. The objective is to formulate a strategy without regret.

The strategy of proportional regret matching

A strategy $g : H \to \Delta(X)$ is said to have no regret if almost all of its realisations of play have no regret. The objective is to formulate a strategy without regret.
One candidate strategy is proposed by Hart and Mas-Colell (2000):

$$q_x^{t+1} =_{Def} \frac{[\bar{r}_x^t]_+}{\sum_{x' \in X} [\bar{r}_{x'}^t]_+}$$

where $[z]_+ =_{Def} z \geq 0 \; ? \; z : 0$.

The strategy of proportional regret matching

A strategy $g : H \to \Delta(X)$ is said to have no regret if almost all of its realisations of play have no regret. The objective is to formulate a strategy without regret.
One candidate strategy is proposed by Hart and Mas-Colell (2000):

$$q_x^{t+1} =_{Def} \frac{[\bar{r}_x^t]_+}{\sum_{x' \in X} [\bar{r}_{x'}^t]_+}$$

where $[z]_+ =_{Def} z \geq 0 \; ? \; z : 0$. This rule is called proportional regret matching, or regret matching (RM for short).

The strategy of proportional regret matching

A strategy $g : H \rightarrow \Delta(X)$ is said to have no regret if almost all of its realisations of play have no regret. The objective is to formulate a strategy without regret.
One candidate strategy is proposed by Hart and Mas-Colell (2000):

$$q_x^{t+1} =_{Def} \frac{[\bar{r}_x^t]_+}{\sum_{x' \in X} [\bar{r}_{x'}^t]_+}$$

where $[z]_+ =_{Def} z \geq 0 \ ? \ z : 0$. This rule is called proportional regret matching, or regret matching (RM for short). Indeed:

Theorem (Hart & Mas-Colell, 2000). *In a finite game, regret matching yields no regret a.s.*

The strategy of proportional regret matching

A strategy $g : H \to \Delta(X)$ is said to have no regret if almost all of its realisations of play have no regret. The objective is to formulate a strategy without regret.
One candidate strategy is proposed by Hart and Mas-Colell (2000):

$$q_x^{t+1} =_{Def} \frac{[\bar{r}_x^t]_+}{\sum_{x' \in X} [\bar{r}_{x'}^t]_+}$$

where $[z]_+ =_{Def} z \geq 0 \ ? \ z : 0$. This rule is called proportional regret matching, or regret matching (RM for short). Indeed:

Theorem (Hart & Mas-Colell, 2000). *In a finite game, regret matching yields no regret a.s.*

Hart & Mas-Colell (2000). "A simple adaptive procedure leading to correlated equilibrium". *Econometrica*, **68**, pp. 1127-1150.

Regret matching differs from reinforcement learning

	0	0	0	1	1	0	0	0	1	0	0	
A	L	R	L	L	R	R	L	R	R	R	R	?
B	R	L	R	L	R	L	R	L	R	L	L	?

Regret matching differs from reinforcement learning

	0	0	0	1	1	0	0	0	1	0	0	
A	L	R	L	L	R	R	L	R	R	R	R	?
B	R	L	R	L	R	L	R	L	R	L	L	?

Proportional regret matching:

	Payoff	*Average regret*	*Regret matching*
Rounds 1-11:	3		
Had L been played:	6	$(6-3)/11$	$3/5$
Had R been played:	5	$(5-3)/11$	$2/5$

Regret matching differs from reinforcement learning

	0	0	0	1	1	0	0	0	1	0	0	
A	L	R	L	L	R	R	L	R	R	R	R	?
B	R	L	R	L	R	L	R	L	R	L	L	?

Proportional regret matching:

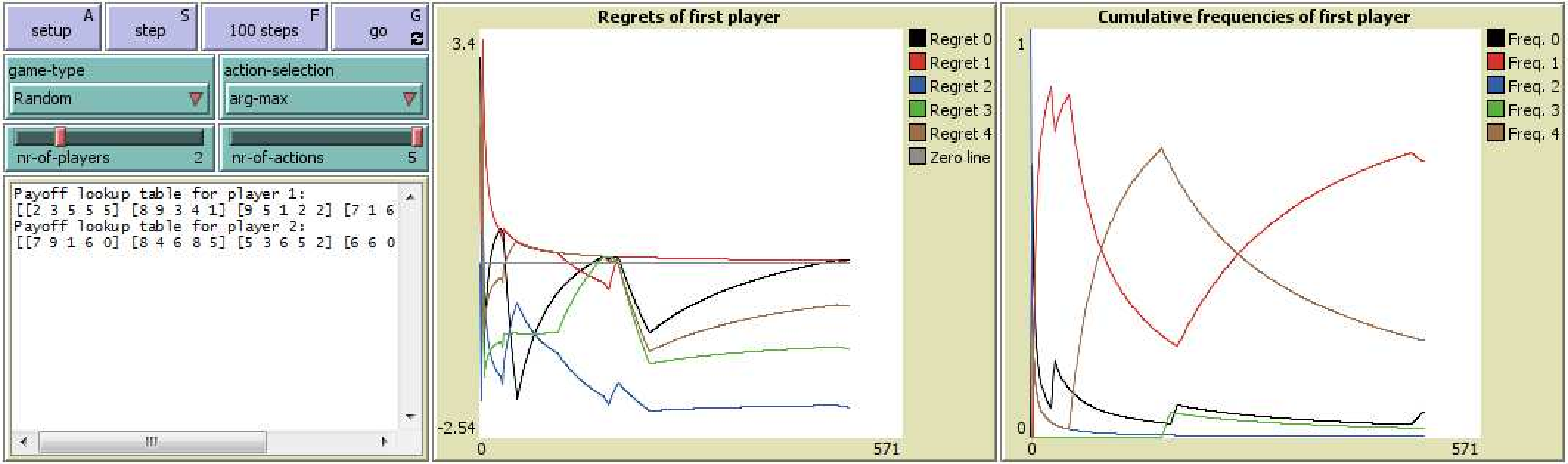
	Payoff	*Average regret*	*Regret matching*
Rounds 1-11:	3		
Had L been played:	6	$(6-3)/11$	$3/5$
Had R been played:	5	$(5-3)/11$	$2/5$

Cumulative payoff matching:

	Accumulated payoff	*Mixed strategy*
Action L:	1	$1/3$
Action R:	2	$2/3$

Regret matching in a 5-person 5-action game

Payoff matrix uninformative. Omitted . . .



Netlogo simulation of regret matching in a 5-person 5-action game.

Regret matching in Shapley's game

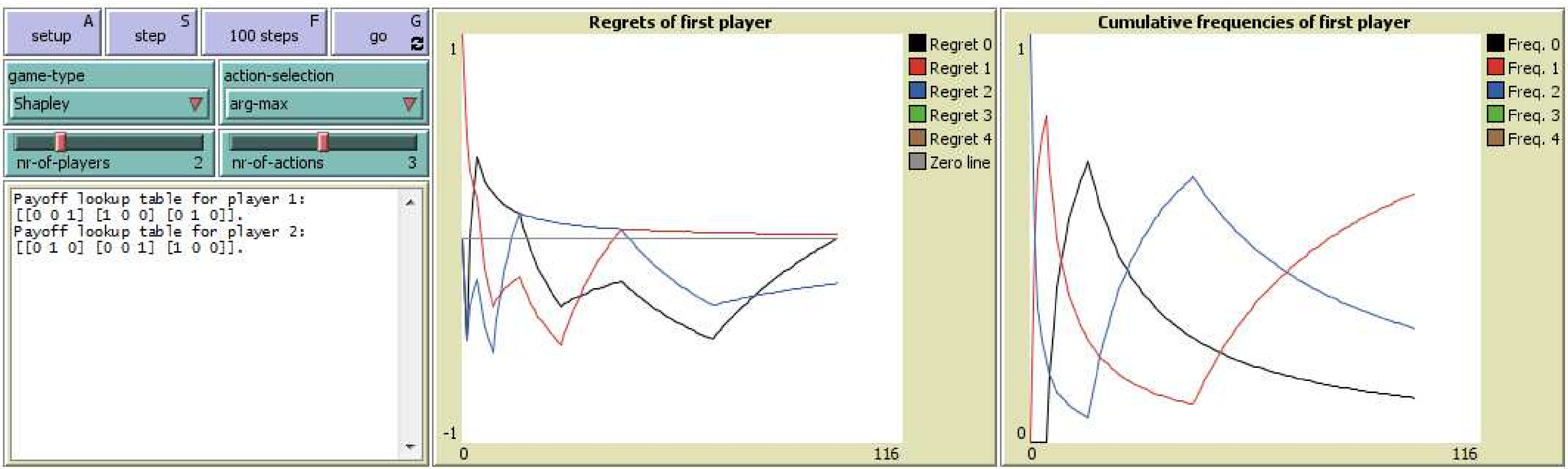
	R	Y	B
R	$(1, 0)$	$(0, 0)$	$(0, 1)$
Y	$(0, 1)$	$(1, 0)$	$(0, 0)$
B	$(0, 0)$	$(0, 1)$	$(1, 0)$

Column is "fashion leader", row is "fashion follower". Column wants to wear a different color than row.

Regret matching in Shapley's game

	R	Y	B
R	$(1,0)$	$(0,0)$	$(0,1)$
Y	$(0,1)$	$(1,0)$	$(0,0)$
B	$(0,0)$	$(0,1)$	$(1,0)$

Column is "fashion leader", row is "fashion follower". Column wants to wear a different color than row.



Netlogo simulation of regret matching in Shapley's game.

Means and ends of regret matching: summary

■

■

Means and ends of regret matching: summary

- Quantities:

-

Means and ends of regret matching: summary

- Quantities:

$$r_x^t =_{Def} \text{ total regret for not playing } x\text{, up to and including } t$$

-

Means and ends of regret matching: summary

- Quantities:

$$r_x^t =_{Def} \text{ total regret for not playing } x\text{, up to and including } t$$
$$\bar{r}_x^t =_{Def} \text{ average regret for not playing } x\text{, up to and including } t$$

-

Means and ends of regret matching: summary

- Quantities:

$$
\begin{aligned}
r_x^t &=_{Def} \text{ total regret for not playing } x\text{, up to and including } t \\
\bar{r}_x^t &=_{Def} \text{ average regret for not playing } x\text{, up to and including } t \\
[\bar{r}_x^t]_+ &=_{Def} \text{ positive average regret for not playing } x
\end{aligned}
$$

-

Means and ends of regret matching: summary

- Quantities:

$$
\begin{aligned}
r_x^t &=_{Def} \text{ total regret for not playing } x\text{, up to and including } t \\
\bar{r}_x^t &=_{Def} \text{ average regret for not playing } x\text{, up to and including } t \\
[\bar{r}_x^t]_+ &=_{Def} \text{ positive average regret for not playing } x \\
\Delta r_x^t &=_{Def} \text{ incremental regret for not playing } x : r_x^t \ - \ r_x^{t-1}
\end{aligned}
$$

-

Means and ends of regret matching: summary

- Quantities:

$$
\begin{aligned}
r_x^t &=_{Def} \text{ total regret for not playing } x\text{, up to and including } t \\
\bar{r}_x^t &=_{Def} \text{ average regret for not playing } x\text{, up to and including } t \\
[\bar{r}_x^t]_+ &=_{Def} \text{ positive average regret for not playing } x \\
\Delta r_x^t &=_{Def} \text{ incremental regret for not playing } x : r_x^t \;-\; r_x^{t-1} \\
E[\Delta r_x^t] &= \text{ expected incremental regret for not playing } x
\end{aligned}
$$

-

Means and ends of regret matching: summary

- Quantities:

$$
\begin{aligned}
r_x^t &=_{Def} \text{ total regret for not playing } x\text{, up to and including } t \\
\bar{r}_x^t &=_{Def} \text{ average regret for not playing } x\text{, up to and including } t \\
[\bar{r}_x^t]_+ &=_{Def} \text{ positive average regret for not playing } x \\
\Delta r_x^t &=_{Def} \text{ incremental regret for not playing } x : r_x^t \; - \; r_x^{t-1} \\
E[\Delta r_x^t] &= \text{ expected incremental regret for not playing } x
\end{aligned}
$$

  Vector versions: r^t, $\bar{r}^t$, $[\bar{r}^t]_+, \dots, E[r^t], E[\Delta r^t]$.

-

Means and ends of regret matching: summary

- Quantities:

$$
\begin{aligned}
r_x^t &=_{Def} \text{ total regret for not playing } x\text{, up to and including } t \\
\bar{r}_x^t &=_{Def} \text{ average regret for not playing } x\text{, up to and including } t \\
[\bar{r}_x^t]_+ &=_{Def} \text{ positive average regret for not playing } x \\
\Delta r_x^t &=_{Def} \text{ incremental regret for not playing } x : r_x^t \; - \; r_x^{t-1} \\
E[\Delta r_x^t] &= \text{ expected incremental regret for not playing } x
\end{aligned}
$$

  Vector versions: r^t, $\bar{r}^t$, $[\bar{r}^t]_+$, $\dots$, $E[r^t]$, $E[\Delta r^t]$.

-

Means and ends of regret matching: summary

- Quantities:

$$
\begin{aligned}
r_x^t &=_{Def} \text{ total regret for not playing } x\text{, up to and including } t \\
\bar{r}_x^t &=_{Def} \text{ average regret for not playing } x\text{, up to and including } t \\
[\bar{r}_x^t]_+ &=_{Def} \text{ positive average regret for not playing } x \\
\Delta r_x^t &=_{Def} \text{ incremental regret for not playing } x : r_x^t \; - \; r_x^{t-1} \\
E[\Delta r_x^t] &= \text{ expected incremental regret for not playing } x
\end{aligned}
$$

  Vector versions: r^t, $\bar{r}^t$, $[\bar{r}^t]_+$, $\dots$, $E[r^t]$, $E[\Delta r^t]$.

- Objective:

$$\lim_{t \to \infty} [\bar{r}^t]_+ = 0 \;\; \text{a.s.}$$

Means and ends of regret matching: summary

- Quantities:

$$
\begin{aligned}
r_x^t &=_{Def} \text{ total regret for not playing } x\text{, up to and including } t \\
\bar{r}_x^t &=_{Def} \text{ average regret for not playing } x\text{, up to and including } t \\
[\bar{r}_x^t]_+ &=_{Def} \text{ positive average regret for not playing } x \\
\Delta r_x^t &=_{Def} \text{ incremental regret for not playing } x : r_x^t \; - \; r_x^{t-1} \\
E[\Delta r_x^t] &= \text{ expected incremental regret for not playing } x
\end{aligned}
$$

  Vector versions: r^t, $\bar{r}^t$, $[\bar{r}^t]_+, \dots, E[r^t]$, $E[\Delta r^t]$.

- Objective:

$$\lim_{t\to\infty} [\bar{r}^t]_+ = 0 \;\text{ a.s.}$$

  i.e., the regret vector must approach the negative orthant with probability one.

Incremental regret and expected incremental regret

Suppose there are only two actions, “1” and “2,” say.

Incremental regret and expected incremental regret

Suppose there are only two actions, “1” and “2,” say.

1. If 1 is executed at $t + 1$ then

Incremental regret and expected incremental regret

Suppose there are only two actions, "1" and "2," say.

1. If 1 is executed at $t+1$ then
   - r_1^{t+1} will not change.

Incremental regret and expected incremental regret

Suppose there are only two actions, "1" and "2," say.

1. If 1 is executed at $t+1$ then
   - r_1^{t+1} will not change.
   - r_2^{t+1} changes with $u(2, y^{t+1}) - u(1, y^{t+1})$.

Incremental regret and expected incremental regret

Suppose there are only two actions, "1" and "2," say.

1. If 1 is executed at $t+1$ then
   - r_1^{t+1} will not change.
   - r_2^{t+1} changes with $u(2, y^{t+1}) - u(1, y^{t+1})$.
2. If 2 is executed at $t+1$ then

Incremental regret and expected incremental regret

Suppose there are only two actions, “1” and “2,” say.

1. If 1 is executed at $t+1$ then
   - r_1^{t+1} will not change.
   - r_2^{t+1} changes with $u(2, y^{t+1}) - u(1, y^{t+1})$.
2. If 2 is executed at $t+1$ then
   - r_1^{t+1} changes with $u(1, y^{t+1}) - u(2, y^{t+1})$.

Incremental regret and expected incremental regret

Suppose there are only two actions, "1" and "2," say.

1. If 1 is executed at $t+1$ then
   - r_1^{t+1} will not change.
   - r_2^{t+1} changes with $u(2, y^{t+1}) - u(1, y^{t+1})$.
2. If 2 is executed at $t+1$ then
   - r_1^{t+1} changes with $u(1, y^{t+1}) - u(2, y^{t+1})$.
   - r_2^{t+1} will not change.

Incremental regret and expected incremental regret

Suppose there are only two actions, "1" and "2," say.

1. If 1 is executed at $t+1$ then
   - r_1^{t+1} will not change.
   - r_2^{t+1} changes with $u(2, y^{t+1}) - u(1, y^{t+1})$.
2. If 2 is executed at $t+1$ then
   - r_1^{t+1} changes with $u(1, y^{t+1}) - u(2, y^{t+1})$.
   - r_2^{t+1} will not change.

If $\alpha^{t+1} =_{Def} u(1, y^{t+1}) - u(2, y^{t+1})$ then incremental regret will be either $(0, -\alpha^{t+1})$

Incremental regret and expected incremental regret

Suppose there are only two actions, “1” and “2,” say.

1. If 1 is executed at $t+1$ then
   - r_1^{t+1} will not change.
   - r_2^{t+1} changes with $u(2, y^{t+1}) - u(1, y^{t+1})$.
2. If 2 is executed at $t+1$ then
   - r_1^{t+1} changes with $u(1, y^{t+1}) - u(2, y^{t+1})$.
   - r_2^{t+1} will not change.

If $\alpha^{t+1} =_{Def} u(1, y^{t+1}) - u(2, y^{t+1})$ then incremental regret will be either $(0, -\alpha^{t+1})$ or $(\alpha^{t+1}, 0)$.

Incremental regret and expected incremental regret

Suppose there are only two actions, “1” and “2,” say.

1. If 1 is executed at $t+1$ then
   - r_1^{t+1} will not change.
   - r_2^{t+1} changes with $u(2, y^{t+1}) - u(1, y^{t+1})$.
2. If 2 is executed at $t+1$ then
   - r_1^{t+1} changes with $u(1, y^{t+1}) - u(2, y^{t+1})$.
   - r_2^{t+1} will not change.

If $\alpha^{t+1} =_{Def} u(1, y^{t+1}) - u(2, y^{t+1})$ then incremental regret will be either $(0, -\alpha^{t+1})$ or $(\alpha^{t+1}, 0)$.

Suppose in round $t+1$ a mixed strategy $q^{t+1} = (q_1^{t+1}, q_2^{t+1})$ is played.

Incremental regret and expected incremental regret

Suppose there are only two actions, "1" and "2," say.

1. If 1 is executed at $t+1$ then
   - r_1^{t+1} will not change.
   - r_2^{t+1} changes with $u(2, y^{t+1}) - u(1, y^{t+1})$.
2. If 2 is executed at $t+1$ then
   - r_1^{t+1} changes with $u(1, y^{t+1}) - u(2, y^{t+1})$.
   - r_2^{t+1} will not change.

If $\alpha^{t+1} =_{Def} u(1, y^{t+1}) - u(2, y^{t+1})$ then incremental regret will be either $(0, -\alpha^{t+1})$ or $(\alpha^{t+1}, 0)$.

Suppose in round $t+1$ a mixed strategy $q^{t+1} = (q_1^{t+1}, q_2^{t+1})$ is played. Then the expected incremental regret is

$$E[\Delta r^{t+1}] = (\qquad\qquad , \qquad\qquad)$$

Incremental regret and expected incremental regret

Suppose there are only two actions, "1" and "2," say.

1. If 1 is executed at $t+1$ then
   - r_1^{t+1} will not change.
   - r_2^{t+1} changes with $u(2, y^{t+1}) - u(1, y^{t+1})$.
2. If 2 is executed at $t+1$ then
   - r_1^{t+1} changes with $u(1, y^{t+1}) - u(2, y^{t+1})$.
   - r_2^{t+1} will not change.

If $\alpha^{t+1} =_{Def} u(1, y^{t+1}) - u(2, y^{t+1})$ then incremental regret will be either $(0, -\alpha^{t+1})$ or $(\alpha^{t+1}, 0)$.

Suppose in round $t+1$ a mixed strategy $q^{t+1} = (q_1^{t+1}, q_2^{t+1})$ is played. Then the expected incremental regret is

$$E[\Delta r^{t+1}] = (\ q_1^{t+1} \cdot 0 \qquad\qquad\qquad , \qquad\qquad\qquad)$$

Incremental regret and expected incremental regret

Suppose there are only two actions, "1" and "2," say.

1. If 1 is executed at $t+1$ then
   - r_1^{t+1} will not change.
   - r_2^{t+1} changes with $u(2, y^{t+1}) - u(1, y^{t+1})$.
2. If 2 is executed at $t+1$ then
   - r_1^{t+1} changes with $u(1, y^{t+1}) - u(2, y^{t+1})$.
   - r_2^{t+1} will not change.

If $\alpha^{t+1} =_{Def} u(1, y^{t+1}) - u(2, y^{t+1})$ then incremental regret will be either $(0, -\alpha^{t+1})$ or $(\alpha^{t+1}, 0)$.

Suppose in round $t+1$ a mixed strategy $q^{t+1} = (q_1^{t+1}, q_2^{t+1})$ is played. Then the expected incremental regret is

$$E[\Delta r^{t+1}] = (\; q_1^{t+1} \cdot 0 + q_2^{t+1} \cdot \alpha^{t+1} \; , \qquad\qquad)$$

Incremental regret and expected incremental regret

Suppose there are only two actions, "1" and "2," say.

1. If 1 is executed at $t+1$ then
   - r_1^{t+1} will not change.
   - r_2^{t+1} changes with $u(2, y^{t+1}) - u(1, y^{t+1})$.
2. If 2 is executed at $t+1$ then
   - r_1^{t+1} changes with $u(1, y^{t+1}) - u(2, y^{t+1})$.
   - r_2^{t+1} will not change.

If $\alpha^{t+1} =_{Def} u(1, y^{t+1}) - u(2, y^{t+1})$ then incremental regret will be either $(0, -\alpha^{t+1})$ or $(\alpha^{t+1}, 0)$.

Suppose in round $t+1$ a mixed strategy $q^{t+1} = (q_1^{t+1}, q_2^{t+1})$ is played. Then the expected incremental regret is

$$E[\Delta r^{t+1}] = (\ q_1^{t+1} \cdot 0 + q_2^{t+1} \cdot \alpha^{t+1}\ ,\ \ q_1^{t+1} \cdot -\alpha^{t+1} \qquad\qquad)$$

Incremental regret and expected incremental regret

Suppose there are only two actions, "1" and "2," say.

1. If 1 is executed at $t+1$ then
   - r_1^{t+1} will not change.
   - r_2^{t+1} changes with $u(2, y^{t+1}) - u(1, y^{t+1})$.
2. If 2 is executed at $t+1$ then
   - r_1^{t+1} changes with $u(1, y^{t+1}) - u(2, y^{t+1})$.
   - r_2^{t+1} will not change.

If $\alpha^{t+1} =_{Def} u(1, y^{t+1}) - u(2, y^{t+1})$ then incremental regret will be either $(0, -\alpha^{t+1})$ or $(\alpha^{t+1}, 0)$.

Suppose in round $t+1$ a mixed strategy $q^{t+1} = (q_1^{t+1}, q_2^{t+1})$ is played. Then the expected incremental regret is

$$E[\Delta r^{t+1}] = (\ q_1^{t+1} \cdot 0 + q_2^{t+1} \cdot \alpha^{t+1}\ ,\ \ q_1^{t+1} \cdot -\alpha^{t+1} + q_2^{t+1} \cdot 0\ )$$

Incremental regret and expected incremental regret

Suppose there are only two actions, “1” and “2,” say.

1. If 1 is executed at $t+1$ then
   - r_1^{t+1} will not change.
   - r_2^{t+1} changes with $u(2, y^{t+1}) - u(1, y^{t+1})$.
2. If 2 is executed at $t+1$ then
   - r_1^{t+1} changes with $u(1, y^{t+1}) - u(2, y^{t+1})$.
   - r_2^{t+1} will not change.

If $\alpha^{t+1} =_{Def} u(1, y^{t+1}) - u(2, y^{t+1})$ then incremental regret will be either $(0, -\alpha^{t+1})$ or $(\alpha^{t+1}, 0)$.

Suppose in round $t+1$ a mixed strategy $q^{t+1} = (q_1^{t+1}, q_2^{t+1})$ is played. Then the expected incremental regret is

$$
\begin{aligned}
E[\Delta r^{t+1}] &= (\; q_1^{t+1} \cdot 0 + q_2^{t+1} \cdot \alpha^{t+1} \;,\; q_1^{t+1} \cdot -\alpha^{t+1} + q_2^{t+1} \cdot 0 \;) \\
&= \alpha^{t+1} (\; q_2^{t+1} \;,\; -q_1^{t+1} \;).
\end{aligned}
$$

Why does regret matching work?

A strategy q such that $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$: 1st attempt

Take $q_1^t = q_2^t = 1/2$ for all t.

A strategy q such that $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$: 1st attempt

Take $q_1^t = q_2^t = 1/2$ for all t. Then

A strategy q such that $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$: 1st attempt

Take $q_1^t = q_2^t = 1/2$ for all t. Then

$$E[\bar{r}_1^t + \bar{r}_2^t] = E[\frac{r_1^{t-1} + \Delta r_1^t}{t} + \frac{r_2^{t-1} + \Delta r_2^t}{t}]$$

A strategy q such that $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$: 1st attempt

Take $q_1^t = q_2^t = 1/2$ for all t. Then

$$E[\bar{r}_1^t + \bar{r}_2^t] = E[\frac{r_1^{t-1} + \Delta r_1^t}{t} + \frac{r_2^{t-1} + \Delta r_2^t}{t}]$$

$$= E[\frac{r_1^{t-1} - \alpha^t/2}{t} + \frac{r_2^{t-1} + \alpha^t/2}{t}] \qquad \text{w.l.o.g.}$$

A strategy q such that $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$: 1st attempt

Take $q_1^t = q_2^t = 1/2$ for all t. Then

$$
\begin{aligned}
E[\bar{r}_1^t + \bar{r}_2^t] &= E[\frac{r_1^{t-1} + \Delta r_1^t}{t} + \frac{r_2^{t-1} + \Delta r_2^t}{t}] \\
&= E[\frac{r_1^{t-1} - \alpha^t/2}{t} + \frac{r_2^{t-1} + \alpha^t/2}{t}] && \text{w.l.o.g.} \\
&= E[\frac{t-1}{t}\left(\bar{r}_1^{t-1} + \bar{r}_2^{t-1}\right)]
\end{aligned}
$$

A strategy q such that $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$: 1st attempt

Take $q_1^t = q_2^t = 1/2$ for all t. Then

$$
\begin{aligned}
E[\bar{r}_1^t + \bar{r}_2^t] &= E[\frac{r_1^{t-1} + \Delta r_1^t}{t} + \frac{r_2^{t-1} + \Delta r_2^t}{t}] \\
&= E[\frac{r_1^{t-1} - \alpha^t/2}{t} + \frac{r_2^{t-1} + \alpha^t/2}{t}] \qquad \text{w.l.o.g.} \\
&= E[\frac{t-1}{t}\left(\bar{r}_1^{t-1} + \bar{r}_2^{t-1}\right)] \\
&= \frac{t-1}{t}E[\bar{r}_1^{t-1} + \bar{r}_2^{t-1}]
\end{aligned}
$$

A strategy q such that $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$: 1st attempt

Take $q_1^t = q_2^t = 1/2$ for all t. Then

$$
\begin{aligned}
E[\bar{r}_1^t + \bar{r}_2^t] &= E[\frac{r_1^{t-1} + \Delta r_1^t}{t} + \frac{r_2^{t-1} + \Delta r_2^t}{t}] \\
&= E[\frac{r_1^{t-1} - \alpha^t/2}{t} + \frac{r_2^{t-1} + \alpha^t/2}{t}] \qquad \text{w.l.o.g.} \\
&= E[\frac{t-1}{t}\left(\bar{r}_1^{t-1} + \bar{r}_2^{t-1}\right)] \\
&= \frac{t-1}{t} E[\bar{r}_1^{t-1} + \bar{r}_2^{t-1}] = \frac{t-1}{t}\left(\bar{r}_1^{t-1} + \bar{r}_2^{t-1}\right)
\end{aligned}
$$

A strategy q such that $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$: 1st attempt

Take $q_1^t = q_2^t = 1/2$ for all t. Then

$$\begin{aligned}
E[\bar{r}_1^t + \bar{r}_2^t] &= E[\frac{r_1^{t-1} + \Delta r_1^t}{t} + \frac{r_2^{t-1} + \Delta r_2^t}{t}] \\
&= E[\frac{r_1^{t-1} - \alpha^t/2}{t} + \frac{r_2^{t-1} + \alpha^t/2}{t}] \qquad \text{w.l.o.g.} \\
&= E[\frac{t-1}{t}\left(\bar{r}_1^{t-1} + \bar{r}_2^{t-1}\right)] \\
&= \frac{t-1}{t} E[\bar{r}_1^{t-1} + \bar{r}_2^{t-1}] = \frac{t-1}{t}\left(\bar{r}_1^{t-1} + \bar{r}_2^{t-1}\right)
\end{aligned}$$

Inductively then $E[\bar{r}_1^t + \bar{r}_2^t] \to 0$

A strategy q such that $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$: 1st attempt

Take $q_1^t = q_2^t = 1/2$ for all t. Then

$$
\begin{aligned}
E[\bar{r}_1^t + \bar{r}_2^t] &= E[\frac{r_1^{t-1} + \Delta r_1^t}{t} + \frac{r_2^{t-1} + \Delta r_2^t}{t}] \\
&= E[\frac{r_1^{t-1} - \alpha^t/2}{t} + \frac{r_2^{t-1} + \alpha^t/2}{t}] \qquad \text{w.l.o.g.} \\
&= E[\frac{t-1}{t}\left(\bar{r}_1^{t-1} + \bar{r}_2^{t-1}\right)] \\
&= \frac{t-1}{t} E[\bar{r}_1^{t-1} + \bar{r}_2^{t-1}] = \frac{t-1}{t}\left(\bar{r}_1^{t-1} + \bar{r}_2^{t-1}\right)
\end{aligned}
$$

Inductively then $E[\bar{r}_1^t + \bar{r}_2^t] \to 0$, so that $\lim_{t\to\infty} \bar{r}_1^t + \bar{r}_2^t = 0$ with probability one.

A strategy q such that $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$: 1st attempt

Take $q_1^t = q_2^t = 1/2$ for all t. Then

$$\begin{aligned} E[\bar{r}_1^t + \bar{r}_2^t] &= E[\frac{r_1^{t-1} + \Delta r_1^t}{t} + \frac{r_2^{t-1} + \Delta r_2^t}{t}] \\ &= E[\frac{r_1^{t-1} - \alpha^t/2}{t} + \frac{r_2^{t-1} + \alpha^t/2}{t}] \qquad \text{w.l.o.g.} \\ &= E[\frac{t-1}{t}\left(\bar{r}_1^{t-1} + \bar{r}_2^{t-1}\right)] \\ &= \frac{t-1}{t} E[\bar{r}_1^{t-1} + \bar{r}_2^{t-1}] = \frac{t-1}{t}\left(\bar{r}_1^{t-1} + \bar{r}_2^{t-1}\right) \end{aligned}$$

Inductively then $E[\bar{r}_1^t + \bar{r}_2^t] \to 0$, so that $\lim_{t\to\infty} \bar{r}_1^t + \bar{r}_2^t = 0$ with probability one.

However, the two terms merely neutralise each other—

A strategy q such that $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$: 1st attempt

Take $q_1^t = q_2^t = 1/2$ for all t. Then

$$\begin{aligned} E[\bar{r}_1^t + \bar{r}_2^t] &= E[\frac{r_1^{t-1} + \Delta r_1^t}{t} + \frac{r_2^{t-1} + \Delta r_2^t}{t}] \\ &= E[\frac{r_1^{t-1} - \alpha^t/2}{t} + \frac{r_2^{t-1} + \alpha^t/2}{t}] \qquad \text{w.l.o.g.} \\ &= E[\frac{t-1}{t}\left(\bar{r}_1^{t-1} + \bar{r}_2^{t-1}\right)] \\ &= \frac{t-1}{t}E[\bar{r}_1^{t-1} + \bar{r}_2^{t-1}] = \frac{t-1}{t}\left(\bar{r}_1^{t-1} + \bar{r}_2^{t-1}\right) \end{aligned}$$

Inductively then $E[\bar{r}_1^t + \bar{r}_2^t] \to 0$, so that $\lim_{t\to\infty} \bar{r}_1^t + \bar{r}_2^t = 0$ with probability one.

However, the two terms merely neutralise each other—which is not what we want: we want *all* regrets to be non-positive.

A strategy q such that $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$: 2nd attempt

A strategy q such that $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$: 2nd attempt

- Each round t, choose an action that would have minimised regret in the previous round.

A strategy q such that $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$: 2nd attempt

- Each round t, choose an action that would have minimised regret in the previous round.
- However: Matching Pennies.

A strategy q such that $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$: 2nd attempt

- Each round t, choose an action that would have minimised regret in the previous round.
- However: Matching Pennies.

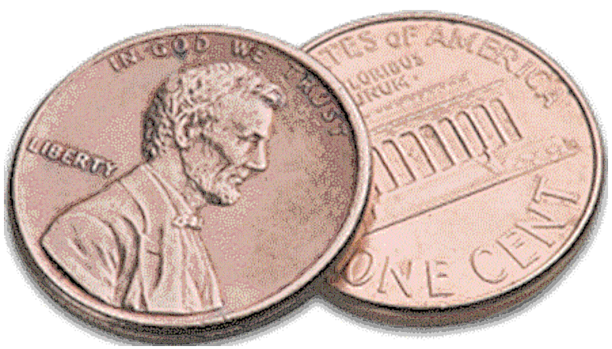
A strategy q such that $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$: 2nd attempt

- Each round t, choose an action that would have minimised regret in the previous round.
- However: Matching Pennies.

  - Switch actions if regret in previous round; else stay.

A strategy q such that $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$: 2nd attempt

- Each round t, choose an action that would have minimised regret in the previous round.
- However: Matching Pennies.

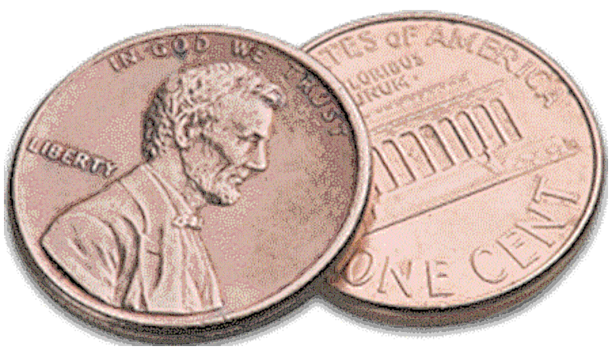


- Switch actions if regret in previous round; else stay.

- Won't work: suppose you meet an opponent who happens to switch every round as well . . .

A strategy q such that $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$: 2nd attempt

- Each round t, choose an action that would have minimised regret in the previous round.
- However: Matching Pennies.



  - Switch actions if regret in previous round; else stay.
  - Won't work: suppose you meet an opponent who happens to switch every round as well . . .
- Won't work in general: your corrections may by coincidence be out of phase with the path of play of your opponent. Peyton Young:

> "Recall that no-regret must hold even when Nature is malevolent."
>
> (p. 26)

Decrease of expected regret

The objective is to find a (mixed) strategy $g : H \to \Delta(\{1,2\})$ such that

$$E[\,\bar{r}^{t+1} \mid r^t, \dots, r^1] < \bar{r}^t \tag{1}$$

Decrease of expected regret

The objective is to find a (mixed) strategy $g : H \to \Delta(\{1,2\})$ such that

$$E[\,\bar{r}^{t+1} \mid r^t, \ldots, r^1] < \bar{r}^t \tag{1}$$

because then Blackwell's approachability theorem can be applied to conclude $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$.

Decrease of expected regret

The objective is to find a (mixed) strategy $g : H \to \Delta(\{1, 2\})$ such that

$$E[\,\bar{r}^{t+1} \mid r^t, \ldots, r^1] \;<\; \bar{r}^t \tag{1}$$

because then Blackwell's approachability theorem can be applied to conclude $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$. Since $\Delta E[r^{t+1}]$ is known, we have

$$E[\,\bar{r}^{t+1} \mid r^t, \ldots, r^1] = E[\,\frac{r^t + \Delta r^{t+1}}{t+1} \mid r^t, \ldots, r^1]$$

Decrease of expected regret

The objective is to find a (mixed) strategy $g : H \to \Delta(\{1,2\})$ such that

$$E[\,\bar{r}^{t+1} \mid r^t, \dots, r^1] \, < \, \bar{r}^t \tag{1}$$

because then Blackwell's approachability theorem can be applied to conclude $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$. Since $\Delta E[r^{t+1}]$ is known, we have

$$\begin{aligned} E[\,\bar{r}^{t+1} \mid r^t, \dots, r^1] &= E[\,\frac{r^t + \Delta r^{t+1}}{t+1} \mid r^t, \dots, r^1] \\ &= \frac{t}{t+1} E[\frac{r^t}{t} \mid r^t, \dots, r^1] + \frac{1}{t+1} E[\Delta r^{t+1} \mid r^t, \dots, r^1] \end{aligned}$$

Decrease of expected regret

The objective is to find a (mixed) strategy $g : H \to \Delta(\{1, 2\})$ such that

$$E[\,\bar{r}^{t+1} \mid r^t, \dots, r^1] \, < \, \bar{r}^t \tag{1}$$

because then Blackwell's approachability theorem can be applied to conclude $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$. Since $\Delta E[r^{t+1}]$ is known, we have

$$\begin{aligned}
E[\,\bar{r}^{t+1} \mid r^t, \dots, r^1] &= E[\,\frac{r^t + \Delta r^{t+1}}{t+1} \mid r^t, \dots, r^1] \\
&= \frac{t}{t+1} E[\frac{r^t}{t} \mid r^t, \dots, r^1] + \frac{1}{t+1} E[\Delta r^{t+1} \mid r^t, \dots, r^1] \\
&= \frac{t}{t+1} \bar{r}^t + \frac{1}{t+1} E[\Delta r^{t+1} \mid r^t, \dots, r^1]
\end{aligned}$$

Decrease of expected regret

The objective is to find a (mixed) strategy $g : H \to \Delta(\{1,2\})$ such that

$$E[\,\bar{r}^{t+1} \mid r^t, \ldots, r^1] \,<\, \bar{r}^t \tag{1}$$

because then Blackwell's approachability theorem can be applied to conclude $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$. Since $\Delta E[r^{t+1}]$ is known, we have

$$\begin{aligned}
E[\,\bar{r}^{t+1} \mid r^t, \ldots, r^1] &= E[\,\frac{r^t + \Delta r^{t+1}}{t+1} \mid r^t, \ldots, r^1] \\
&= \frac{t}{t+1} E[\frac{r^t}{t} \mid r^t, \ldots, r^1] + \frac{1}{t+1} E[\Delta r^{t+1} \mid r^t, \ldots, r^1] \\
&= \frac{t}{t+1} \bar{r}^t + \frac{1}{t+1} E[\Delta r^{t+1} \mid r^t, \ldots, r^1] \\
&= \frac{t}{t+1} \bar{r}^t + \frac{1}{t+1} (-\alpha^{t+1} q_2^{t+1},\ \alpha^{t+1} q_1^{t+1})
\end{aligned}$$

Decrease of expected regret

The objective is to find a (mixed) strategy $g : H \to \Delta(\{1,2\})$ such that

$$E[\,\bar{r}^{t+1} \mid r^t, \dots, r^1] \; < \; \bar{r}^t \tag{1}$$

because then Blackwell's approachability theorem can be applied to conclude $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$. Since $\Delta E[r^{t+1}]$ is known, we have

$$\begin{aligned}
E[\,\bar{r}^{t+1} \mid r^t, \dots, r^1] &= E[\,\frac{r^t + \Delta r^{t+1}}{t+1} \mid r^t, \dots, r^1] \\
&= \frac{t}{t+1} E[\frac{r^t}{t} \mid r^t, \dots, r^1] + \frac{1}{t+1} E[\Delta r^{t+1} \mid r^t, \dots, r^1] \\
&= \frac{t}{t+1} \bar{r}^t + \frac{1}{t+1} E[\Delta r^{t+1} \mid r^t, \dots, r^1] \\
&= \frac{t}{t+1} \bar{r}^t + \frac{1}{t+1} (-\alpha^{t+1} q_2^{t+1} ,\, \alpha^{t+1} q_1^{t+1})
\end{aligned}$$

So, the objective is to find a strategy such that $\alpha^{t+1}(-q_2^{t+1} ,\, q_1^{t+1}) < \bar{r}^t$.

A strategy q such that $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$: 3rd attempt

A strategy q such that $\lim_{t \to \infty} [\bar{r}^t]_+ = 0$: 3rd attempt

- Recall: our objective is $[\bar{r}^t]_+ \to 0$.

A strategy q such that $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$: 3rd attempt

- Recall: our objective is $[\bar{r}^t]_+ \to 0$.
- To this end, choose q^{t+1} such that

$$E[\Delta r^{t+1}] \perp [\bar{r}^t]_+$$

A strategy q such that $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$: 3rd attempt

- Recall: our objective is $[\bar{r}^t]_+ \to 0$.
- To this end, choose q^{t+1} such that

$$E[\Delta r^{t+1}] \perp [\bar{r}^t]_+$$

So:

$$\begin{aligned} & E[\Delta r^{t+1}] \cdot [\bar{r}^t]_+ = 0 \\ \Leftrightarrow\ & (-\alpha^{t+1} q_2^{t+1},\ \alpha^{t+1} q_1^{t+1}) \cdot [\bar{r}^t]_+ = 0 \\ \Leftrightarrow\ & \alpha^{t+1}(q_1^{t+1}[\bar{r}_2^t]_+ - q_2^{t+1}[\bar{r}_1^t]_+) = 0 \\ \Leftrightarrow\ & q_1^{t+1} : q_2^{t+1} = [\bar{r}_1^t]_+ : [\bar{r}_2^t]_+. \end{aligned}$$

A strategy q such that $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$: 3rd attempt

- Recall: our objective is $[\bar{r}^t]_+ \to 0$.
- To this end, choose q^{t+1} such that

$$E[\Delta r^{t+1}] \perp [\bar{r}^t]_+$$

So:

$$\begin{aligned} & E[\Delta r^{t+1}] \cdot [\bar{r}^t]_+ = 0 \\ \Leftrightarrow\ & (-\alpha^{t+1} q_2^{t+1},\ \alpha^{t+1} q_1^{t+1}) \cdot [\bar{r}^t]_+ = 0 \\ \Leftrightarrow\ & \alpha^{t+1}(q_1^{t+1}[\bar{r}_2^t]_+ - q_2^{t+1}[\bar{r}_1^t]_+) = 0 \\ \Leftrightarrow\ & q_1^{t+1} : q_2^{t+1} = [\bar{r}_1^t]_+ : [\bar{r}_2^t]_+. \end{aligned}$$

The last equation amounts to proportional regret matching.

A strategy q such that $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$: 3rd attempt

- Recall: our objective is $[\bar{r}^t]_+ \to 0$.
- To this end, choose q^{t+1} such that

$$E[\Delta r^{t+1}] \perp [\bar{r}^t]_+$$

So:

$$
\begin{aligned}
& E[\Delta r^{t+1}] \cdot [\bar{r}^t]_+ = 0 \\
\Leftrightarrow\ & (-\alpha^{t+1} q_2^{t+1},\ \alpha^{t+1} q_1^{t+1}) \cdot [\bar{r}^t]_+ = 0 \\
\Leftrightarrow\ & \alpha^{t+1}(q_1^{t+1}[\bar{r}_2^t]_+ - q_2^{t+1}[\bar{r}_1^t]_+) = 0 \\
\Leftrightarrow\ & q_1^{t+1} : q_2^{t+1} = [\bar{r}_1^t]_+ : [\bar{r}_2^t]_+ .
\end{aligned}
$$

The last equation amounts to proportional regret matching.

(Notice that α^{t+1} has left the stage.)

A strategy q such that $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$: 3rd attempt

- Recall: our objective is $[\bar{r}^t]_+ \to 0$.
- To this end, choose q^{t+1} such that

$$E[\Delta r^{t+1}] \perp [\bar{r}^t]_+$$

So:

$$\begin{aligned} & E[\Delta r^{t+1}] \cdot [\bar{r}^t]_+ = 0 \\ \Leftrightarrow \; & (-\alpha^{t+1} q_2^{t+1},\, \alpha^{t+1} q_1^{t+1}) \cdot [\bar{r}^t]_+ = 0 \\ \Leftrightarrow \; & \alpha^{t+1}(q_1^{t+1}[\bar{r}_2^t]_+ - q_2^{t+1}[\bar{r}_1^t]_+) = 0 \\ \Leftrightarrow \; & q_1^{t+1} : q_2^{t+1} = [\bar{r}_1^t]_+ : [\bar{r}_2^t]_+. \end{aligned}$$

The last equation amounts to proportional regret matching.

(Notice that α^{t+1} has left the stage.)

- Boundary cases are obvious and can be treated as follows:

A strategy q such that $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$: 3rd attempt

- Recall: our objective is $[\bar{r}^t]_+ \to 0$.
- To this end, choose q^{t+1} such that

$$E[\Delta r^{t+1}] \perp [\bar{r}^t]_+$$

So:

$$\begin{aligned} & E[\Delta r^{t+1}] \cdot [\bar{r}^t]_+ = 0 \\ \Leftrightarrow\ & (-\alpha^{t+1} q_2^{t+1},\ \alpha^{t+1} q_1^{t+1}) \cdot [\bar{r}^t]_+ = 0 \\ \Leftrightarrow\ & \alpha^{t+1}(q_1^{t+1}[\bar{r}_2^t]_+ - q_2^{t+1}[\bar{r}_1^t]_+) = 0 \\ \Leftrightarrow\ & q_1^{t+1} : q_2^{t+1} = [\bar{r}_1^t]_+ : [\bar{r}_2^t]_+. \end{aligned}$$

The last equation amounts to proportional regret matching.

(Notice that α^{t+1} has left the stage.)

- Boundary cases are obvious and can be treated as follows:
  - If $\bar{r}_1^t \leq 0$ and $\bar{r}_2^t > 0$, then let $q^{t+1} =_{Def} (0, 1)$.

A strategy q such that $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$: 3rd attempt

- Recall: our objective is $[\bar{r}^t]_+ \to 0$.
- To this end, choose q^{t+1} such that

$$E[\Delta r^{t+1}] \perp [\bar{r}^t]_+$$

So:

$$\begin{aligned} & E[\Delta r^{t+1}] \cdot [\bar{r}^t]_+ = 0 \\ \Leftrightarrow\ & (-\alpha^{t+1} q_2^{t+1},\ \alpha^{t+1} q_1^{t+1}) \cdot [\bar{r}^t]_+ = 0 \\ \Leftrightarrow\ & \alpha^{t+1}(q_1^{t+1}[\bar{r}_2^t]_+ - q_2^{t+1}[\bar{r}_1^t]_+) = 0 \\ \Leftrightarrow\ & q_1^{t+1} : q_2^{t+1} = [\bar{r}_1^t]_+ : [\bar{r}_2^t]_+. \end{aligned}$$

The last equation amounts to proportional regret matching.

(Notice that α^{t+1} has left the stage.)

- Boundary cases are obvious and can be treated as follows:
  - If $\bar{r}_1^t \leq 0$ and $\bar{r}_2^t > 0$, then let $q^{t+1} =_{Def} (0, 1)$.
  - If $\bar{r}_1^t > 0$ and $\bar{r}_2^t \leq 0$, then let $q^{t+1} =_{Def} (1, 0)$.

A strategy q such that $\lim_{t\to\infty}[\bar{r}^t]_+ = 0$: 3rd attempt

- Recall: our objective is $[\bar{r}^t]_+ \to 0$.
- To this end, choose q^{t+1} such that

$$E[\Delta r^{t+1}] \perp [\bar{r}^t]_+$$

So:

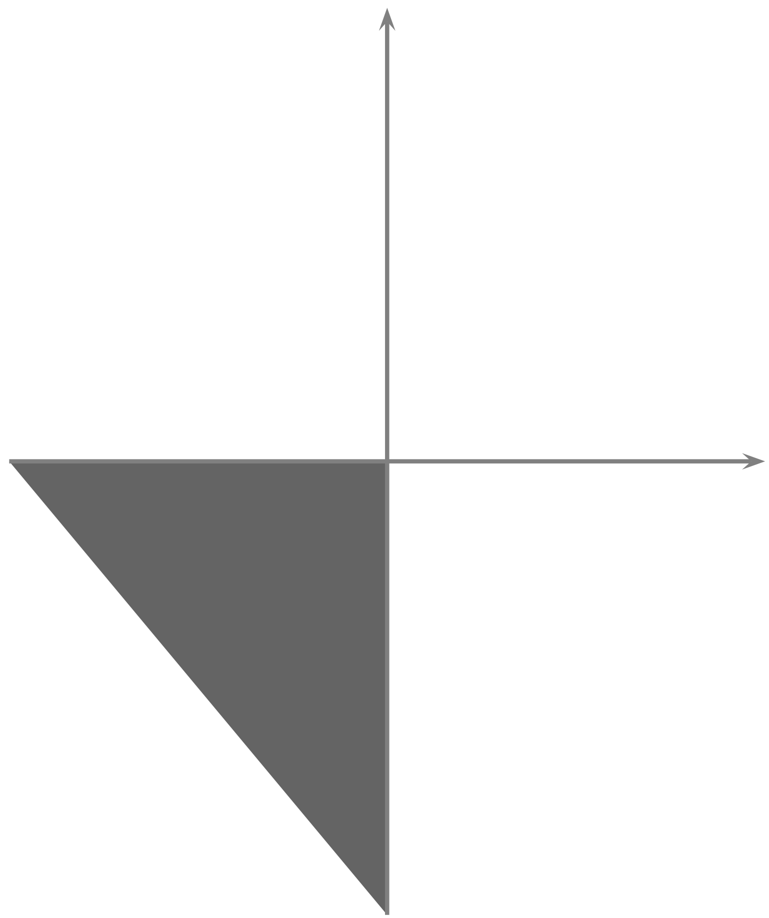
$$E[\Delta r^{t+1}] \cdot [\bar{r}^t]_+ = 0$$
$$\Leftrightarrow \ (-\alpha^{t+1} q_2^{t+1},\ \alpha^{t+1} q_1^{t+1}) \cdot [\bar{r}^t]_+ = 0$$
$$\Leftrightarrow \ \alpha^{t+1}(q_1^{t+1}[\bar{r}_2^t]_+ - q_2^{t+1}[\bar{r}_1^t]_+) = 0$$
$$\Leftrightarrow \ q_1^{t+1} : q_2^{t+1} = [\bar{r}_1^t]_+ : [\bar{r}_2^t]_+.$$

The last equation amounts to proportional regret matching.

(Notice that α^{t+1} has left the stage.)

- Boundary cases are obvious and can be treated as follows:
  - If $\bar{r}_1^t \leq 0$ and $\bar{r}_2^t > 0$, then let $q^{t+1} =_{Def} (0, 1)$.
  - If $\bar{r}_1^t > 0$ and $\bar{r}_2^t \leq 0$, then let $q^{t+1} =_{Def} (1, 0)$.
  - If all regret is non-positive, then play an action at random.

Stochastic dynamics of regret matching



- Expected incremental regret, $E[\Delta r^{t+1}]$ is made orthogonal to the current regret, independently of the unknown α^{t+1}.

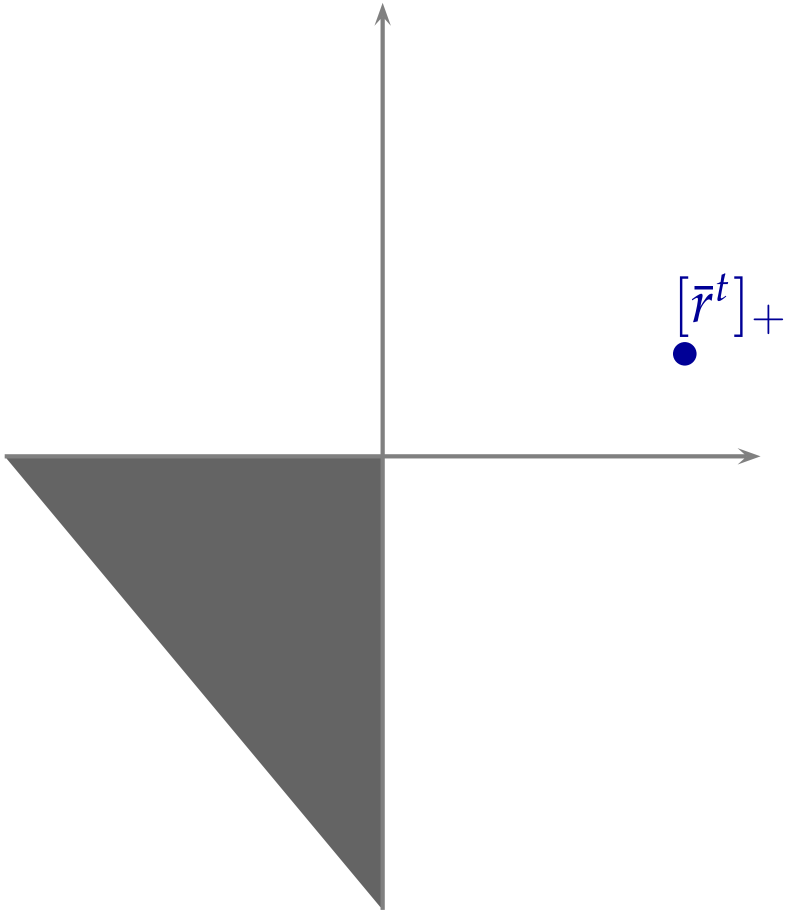
  Because at A does not know what B will play next, this is crucial.

- $E[\bar{r}^{t+1}]$ is a convex combination of $\bar{r}^t_+$ and $E[\Delta r^{t+1}]$.

- Since $E[\Delta r^{t+1}] \perp \bar{r}^t_+$, $E[\bar{r}^{t+1}]$ lies closer to the non-positive orthant than $\bar{r}^t_+$ does, provided t is large.

- Ultimately, the result follows from Blackwell's approachability theorem (Strategic Learning and its Limits, 2004, Ch. 4).

Stochastic dynamics of regret matching

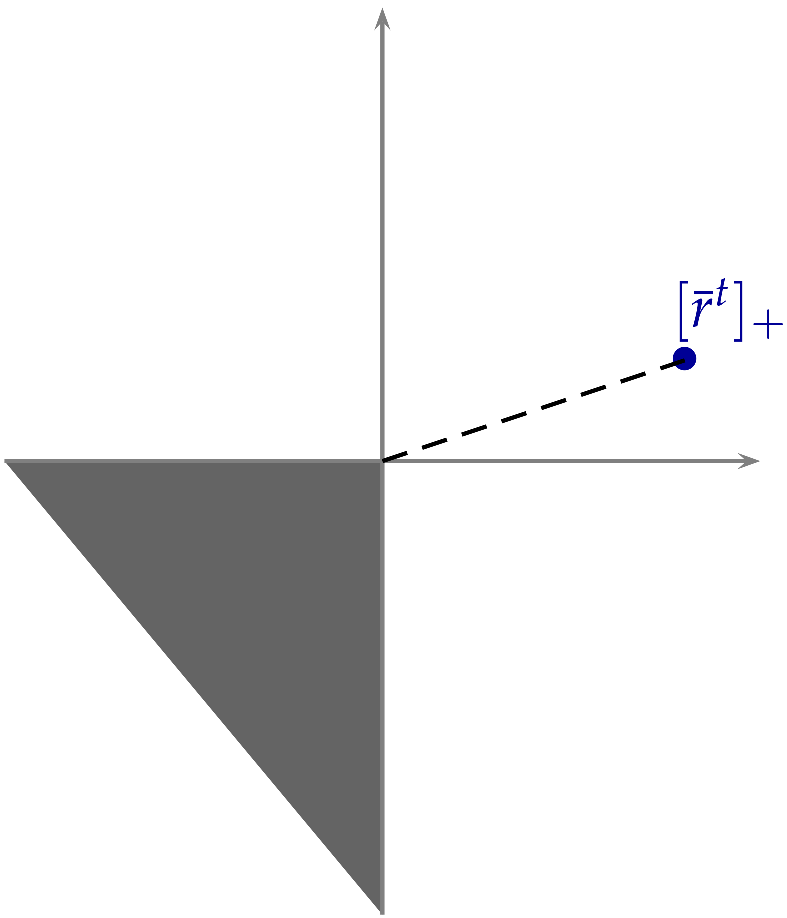


- Expected incremental regret, $E[\Delta r^{t+1}]$ is made orthogonal to the current regret, independently of the unknown α^{t+1}.

  Because at A does not know what B will play next, this is crucial.

- $E[\bar{r}^{t+1}]$ is a convex combination of $\bar{r}^t_+$ and $E[\Delta r^{t+1}]$.
- Since $E[\Delta r^{t+1}] \perp \bar{r}^t_+$, $E[\bar{r}^{t+1}]$ lies closer to the non-positive orthant than $\bar{r}^t_+$ does, provided t is large.
- Ultimately, the result follows from Blackwell's approachability theorem (Strategic Learning and its Limits, 2004, Ch. 4).

Stochastic dynamics of regret matching



- Expected incremental regret, $E[\Delta r^{t+1}]$ is made orthogonal to the current regret, independently of the unknown α^{t+1}.

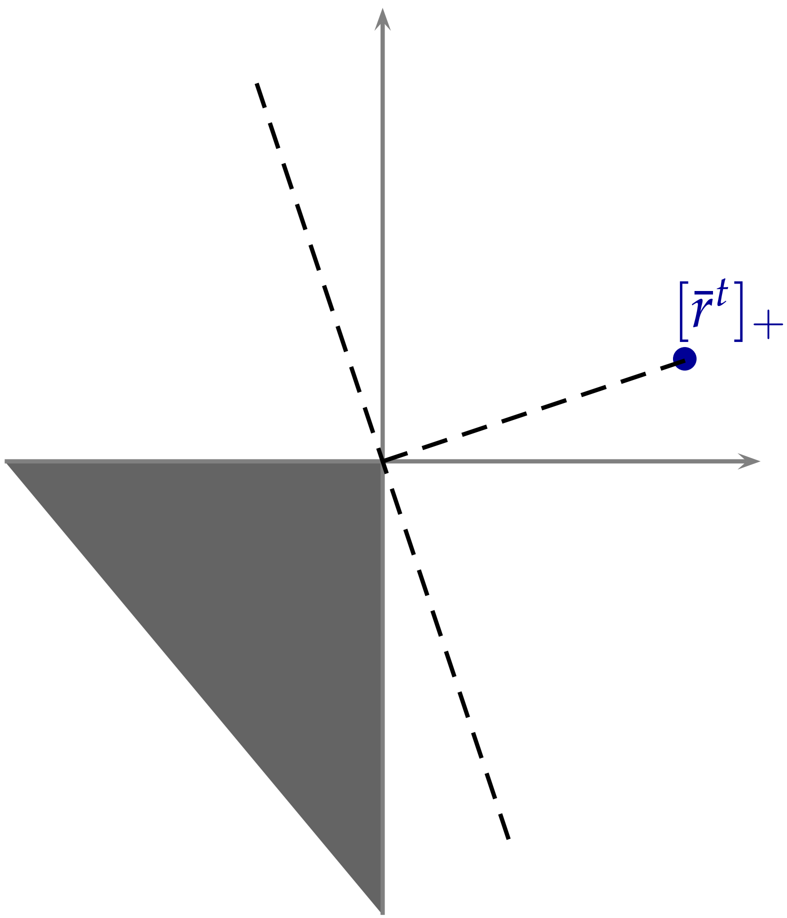
  Because at A does not know what B will play next, this is crucial.

- $E[\bar{r}^{t+1}]$ is a convex combination of $\bar{r}^t_+$ and $E[\Delta r^{t+1}]$.

- Since $E[\Delta r^{t+1}] \perp \bar{r}^t_+$, $E[\bar{r}^{t+1}]$ lies closer to the non-positive orthant than $\bar{r}^t_+$ does, provided t is large.

- Ultimately, the result follows from Blackwell's approachability theorem (Strategic Learning and its Limits, 2004, Ch. 4).

Stochastic dynamics of regret matching

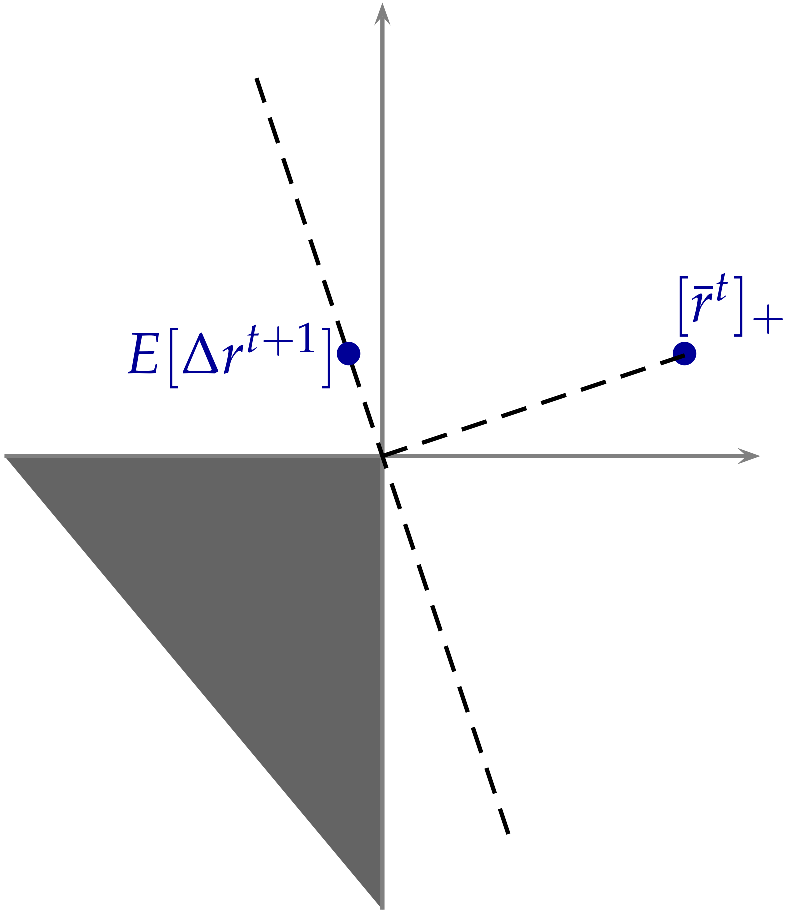


- Expected incremental regret, $E[\Delta r^{t+1}]$ is made orthogonal to the current regret, independently of the unknown α^{t+1}.

  Because at A does not know what B will play next, this is crucial.

- $E[\bar{r}^{t+1}]$ is a convex combination of $\bar{r}^t_+$ and $E[\Delta r^{t+1}]$.
- Since $E[\Delta r^{t+1}] \perp \bar{r}^t_+$, $E[\bar{r}^{t+1}]$ lies closer to the non-positive orthant than $\bar{r}^t_+$ does, provided t is large.
- Ultimately, the result follows from Blackwell's approachability theorem (Strategic Learning and its Limits, 2004, Ch. 4).

Stochastic dynamics of regret matching



- Expected incremental regret, $E[\Delta r^{t+1}]$ is made orthogonal to the current regret, independently of the unknown α^{t+1}.

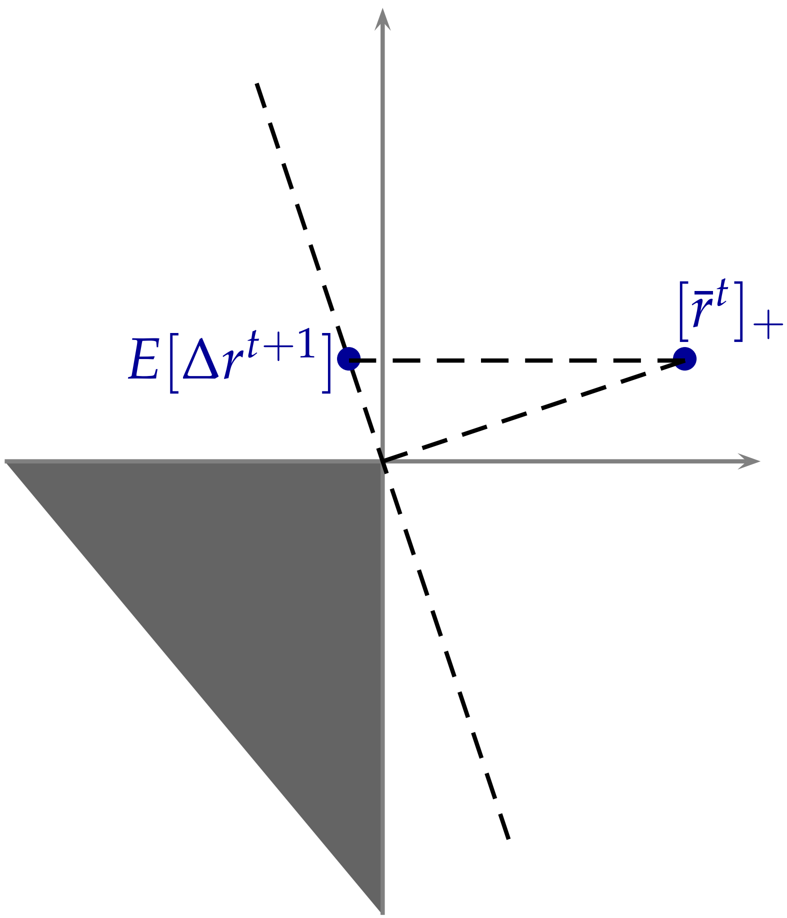
  Because at A does not know what B will play next, this is crucial.

- $E[\bar{r}^{t+1}]$ is a convex combination of $\bar{r}^t_+$ and $E[\Delta r^{t+1}]$.

- Since $E[\Delta r^{t+1}] \perp \bar{r}^t_+$, $E[\bar{r}^{t+1}]$ lies closer to the non-positive orthant than $\bar{r}^t_+$ does, provided t is large.

- Ultimately, the result follows from Blackwell's approachability theorem (Strategic Learning and its Limits, 2004, Ch. 4).

Stochastic dynamics of regret matching



- Expected incremental regret, $E[\Delta r^{t+1}]$ is made orthogonal to the current regret, independently of the unknown α^{t+1}.

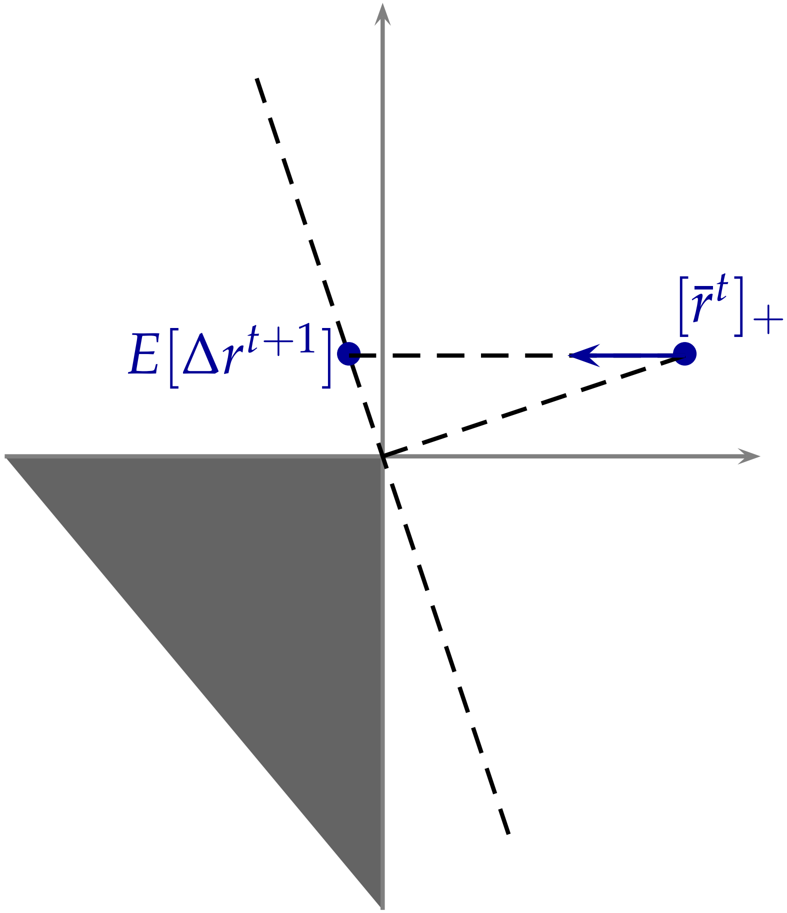
  Because at A does not know what B will play next, this is crucial.

- $E[\bar{r}^{t+1}]$ is a convex combination of $\bar{r}^t_+$ and $E[\Delta r^{t+1}]$.

- Since $E[\Delta r^{t+1}] \perp \bar{r}^t_+$, $E[\bar{r}^{t+1}]$ lies closer to the non-positive orthant than $\bar{r}^t_+$ does, provided t is large.

- Ultimately, the result follows from Blackwell's approachability theorem (Strategic Learning and its Limits, 2004, Ch. 4).

Stochastic dynamics of regret matching



- Expected incremental regret, $E[\Delta r^{t+1}]$ is made orthogonal to the current regret, independently of the unknown α^{t+1}.

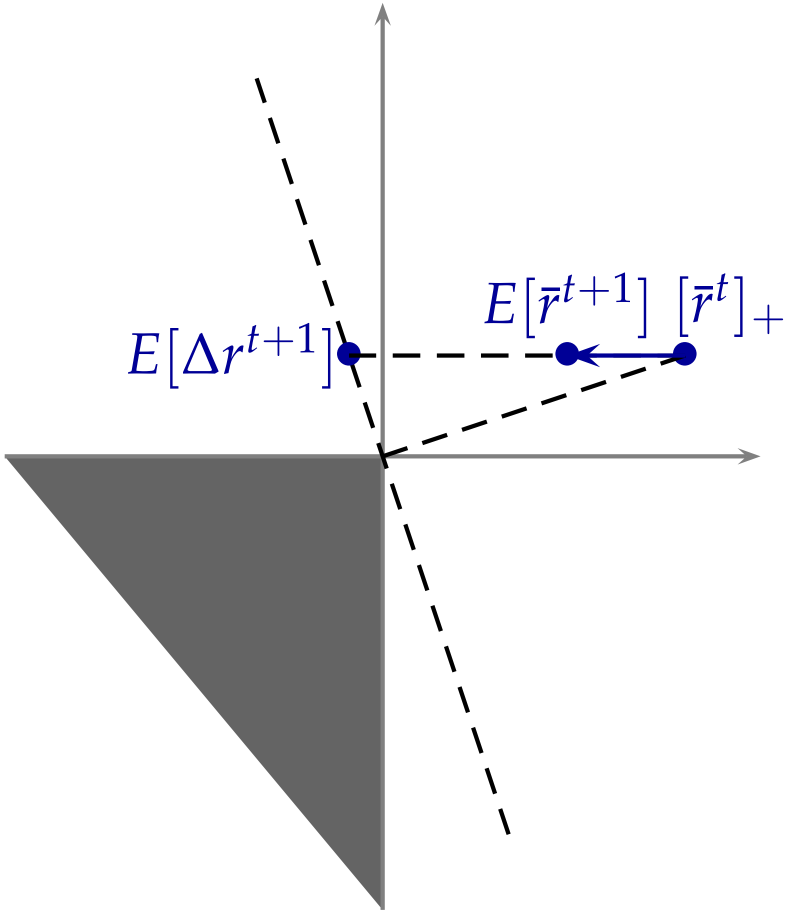
  Because at A does not know what B will play next, this is crucial.

- $E[\bar{r}^{t+1}]$ is a convex combination of $\bar{r}^t_+$ and $E[\Delta r^{t+1}]$.

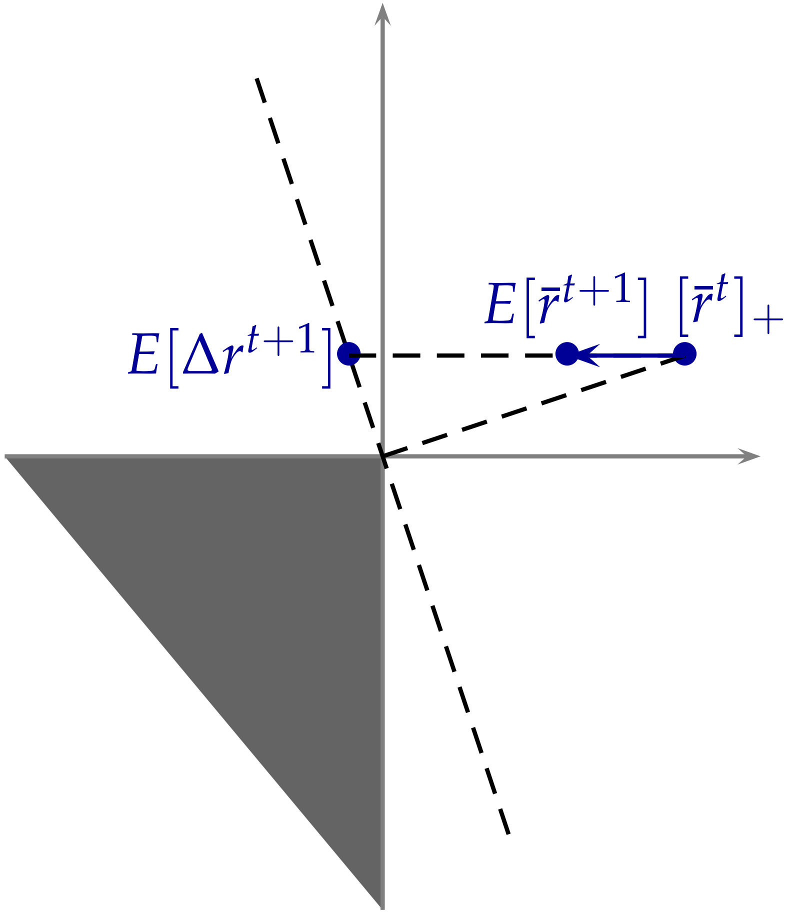
- Since $E[\Delta r^{t+1}] \perp \bar{r}^t_+$, $E[\bar{r}^{t+1}]$ lies closer to the non-positive orthant than $\bar{r}^t_+$ does, provided t is large.

- Ultimately, the result follows from Blackwell's approachability theorem (Strategic Learning and its Limits, 2004, Ch. 4).

Stochastic dynamics of regret matching

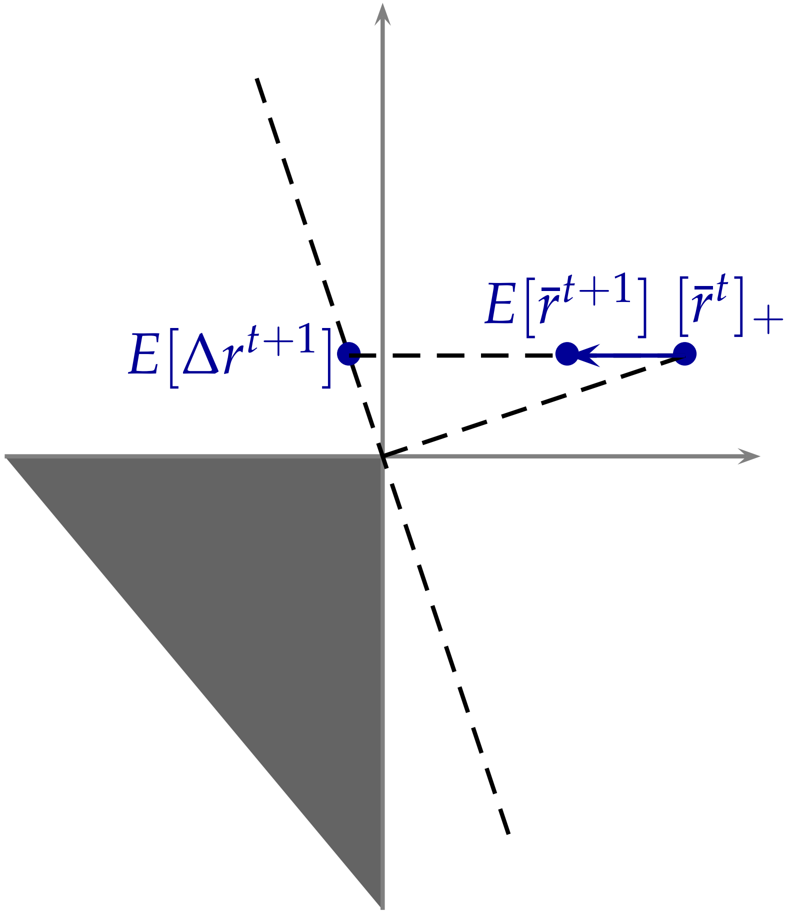


Stochastic dynamics of regret matching



- Expected incremental regret, $E[\Delta r^{t+1}]$ is made orthogonal to the current regret, independently of the unknown α^{t+1}.

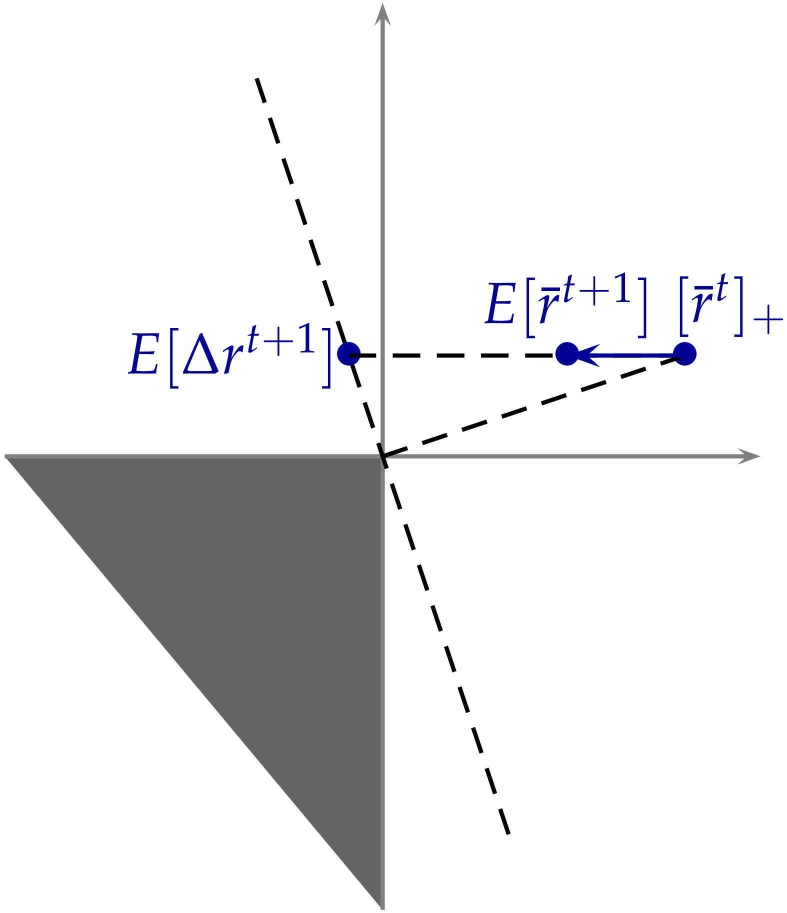
Stochastic dynamics of regret matching



- Expected incremental regret, $E[\Delta r^{t+1}]$ is made orthogonal to the current regret, independently of the unknown α^{t+1}.

  Because at A does not know what B will play next, this is crucial.

Stochastic dynamics of regret matching

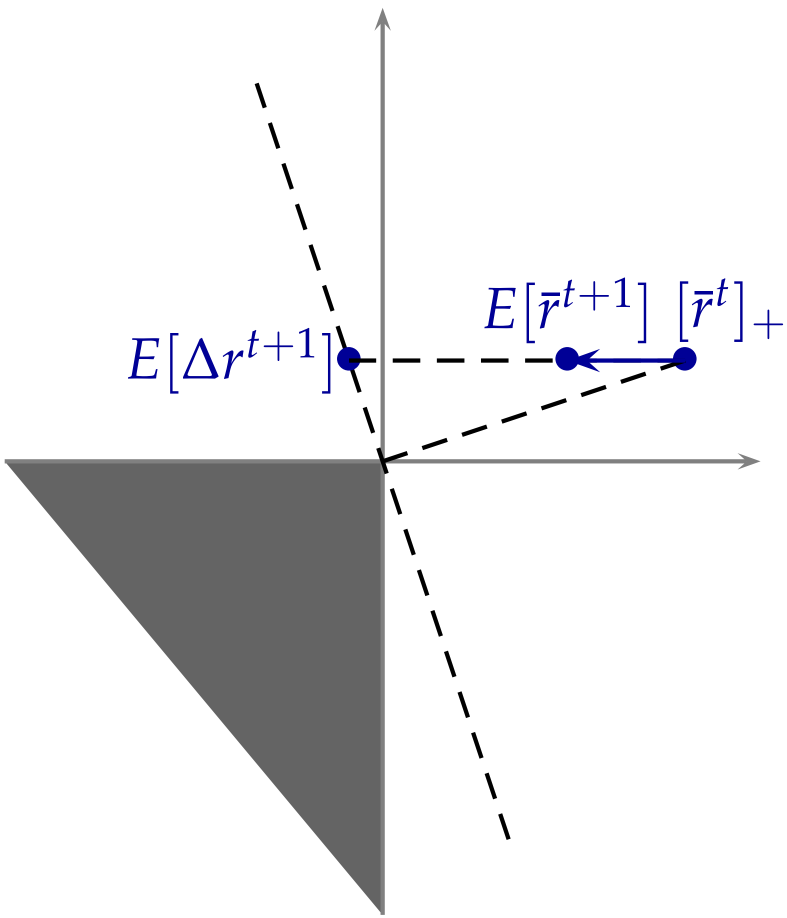


- Expected incremental regret, $E[\Delta r^{t+1}]$ is made orthogonal to the current regret, independently of the unknown α^{t+1}.

  Because at A does not know what B will play next, this is crucial.

- $E[\bar{r}^{t+1}]$ is a convex combination of $\bar{r}^t_+$ and $E[\Delta r^{t+1}]$.

Stochastic dynamics of regret matching



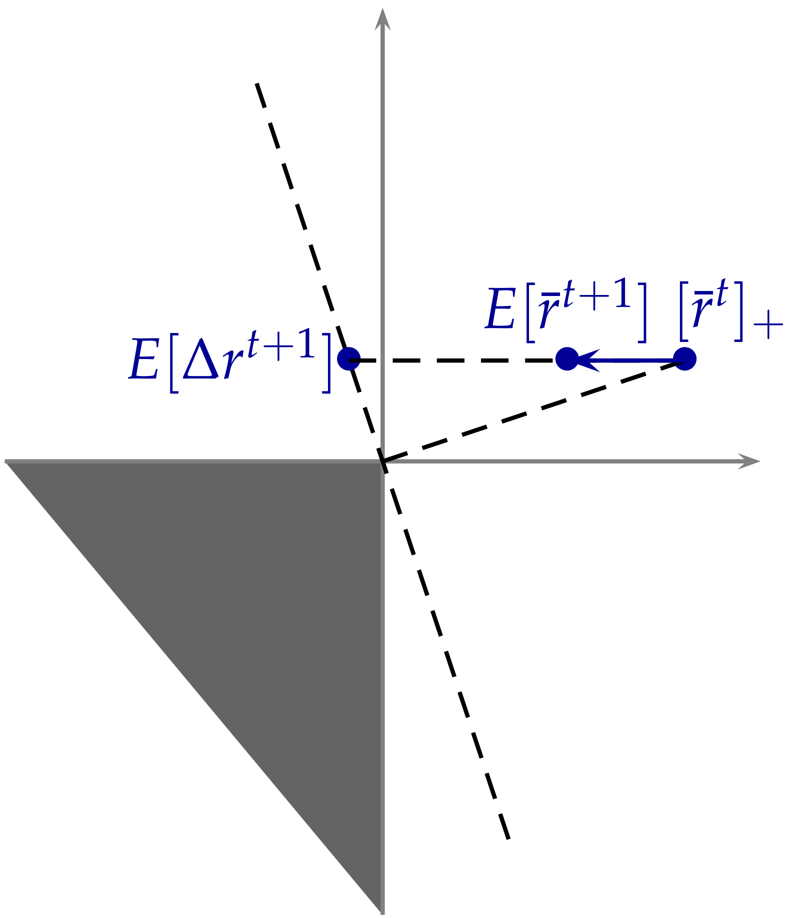
- Expected incremental regret, $E[\Delta r^{t+1}]$ is made orthogonal to the current regret, independently of the unknown α^{t+1}.

  Because at A does not know what B will play next, this is crucial.

- $E[\bar{r}^{t+1}]$ is a convex combination of $\bar{r}^t_+$ and $E[\Delta r^{t+1}]$.

- Since $E[\Delta r^{t+1}] \perp \bar{r}^t_+$, $E[\bar{r}^{t+1}]$ lies closer to the non-positive orthant than $\bar{r}^t_+$ does, provided t is large.

Stochastic dynamics of regret matching



- Expected incremental regret, $E[\Delta r^{t+1}]$ is made orthogonal to the current regret, independently of the unknown α^{t+1}.

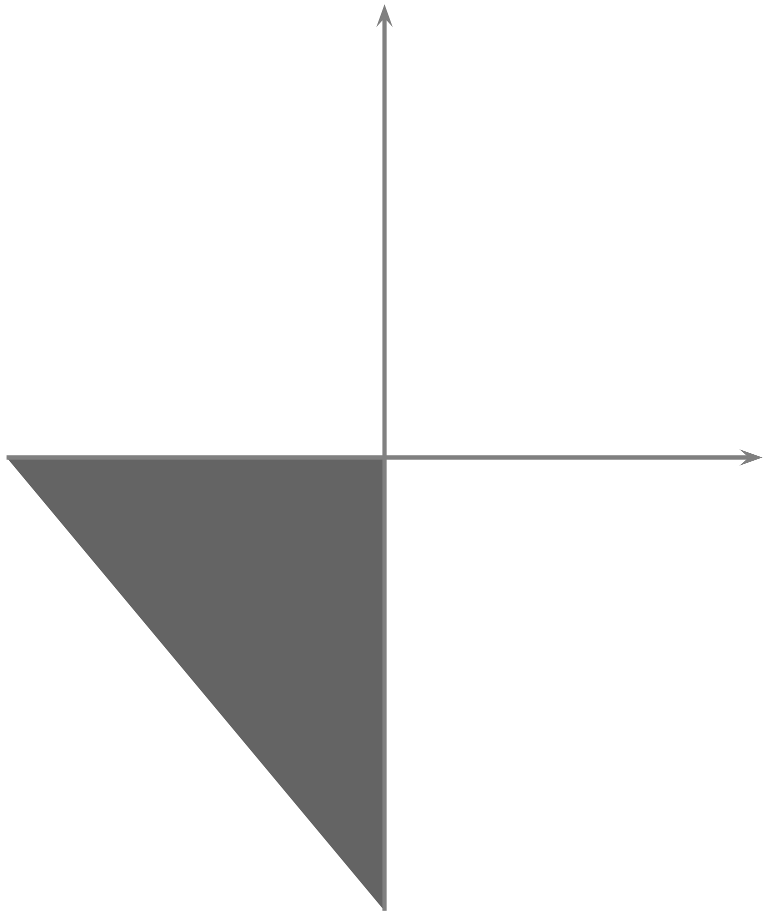
  Because at A does not know what B will play next, this is crucial.

- $E[\bar{r}^{t+1}]$ is a convex combination of $\bar{r}^t_+$ and $E[\Delta r^{t+1}]$.

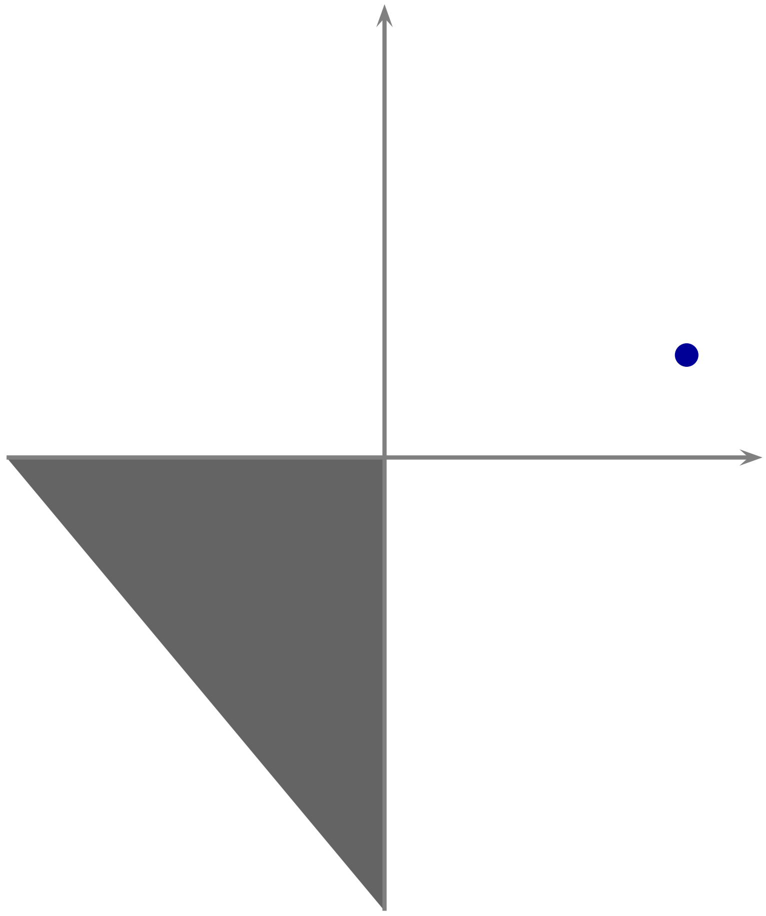
- Since $E[\Delta r^{t+1}] \perp \bar{r}^t_+$, $E[\bar{r}^{t+1}]$ lies closer to the non-positive orthant than $\bar{r}^t_+$ does, provided t is large.

- Ultimately, the result follows from Blackwell's approachability theorem (Strategic Learning and its Limits, 2004, Ch. 4).

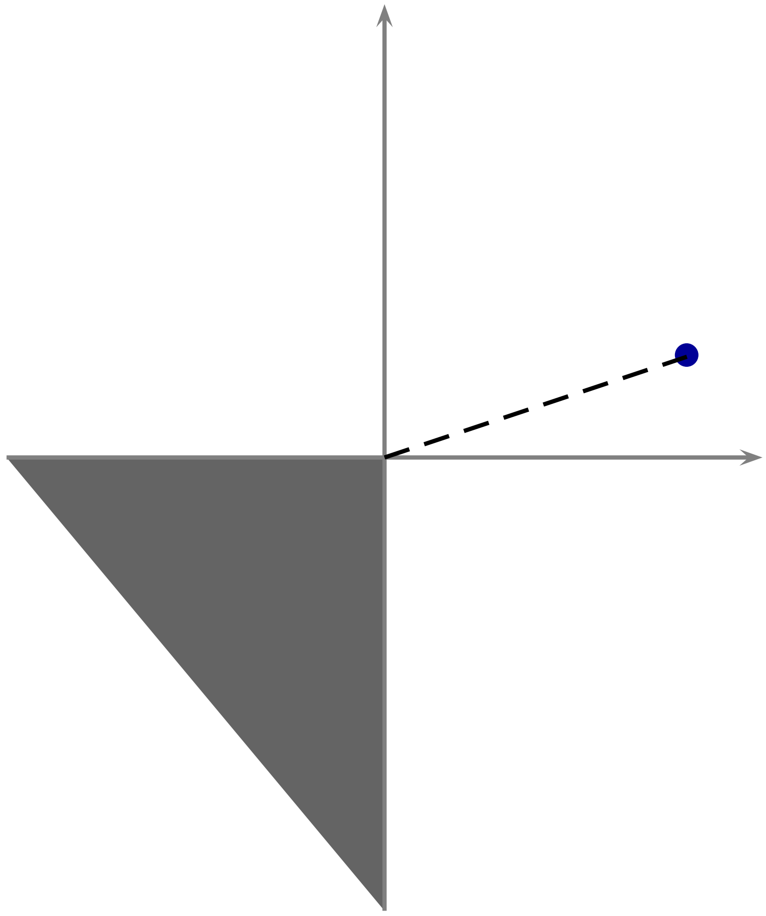
Stochastic dynamics of regret matching



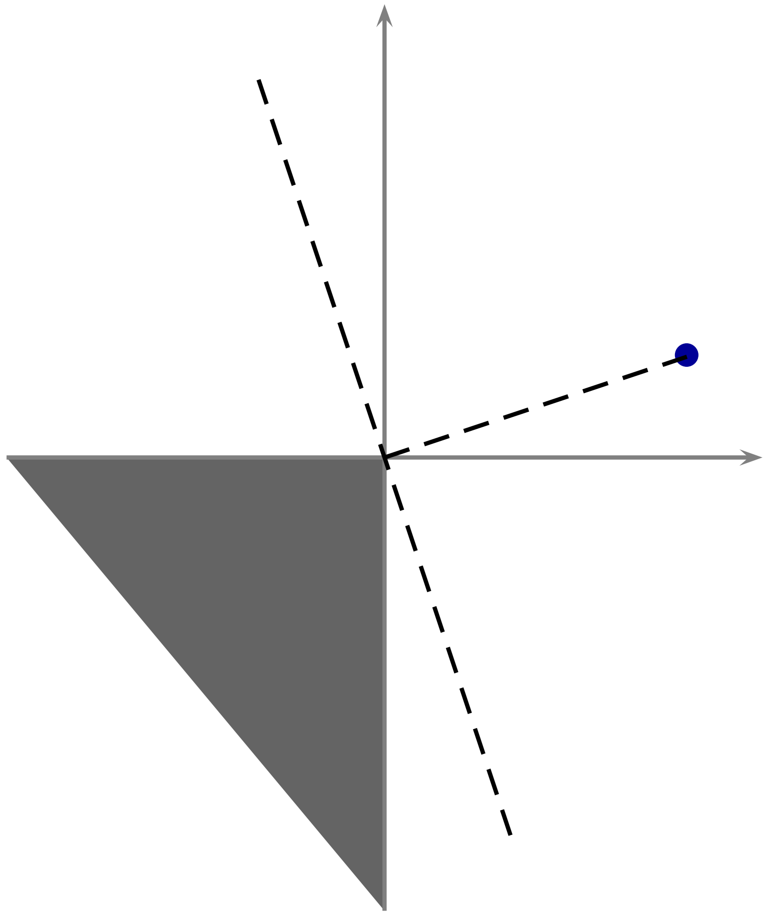
Stochastic dynamics of regret matching



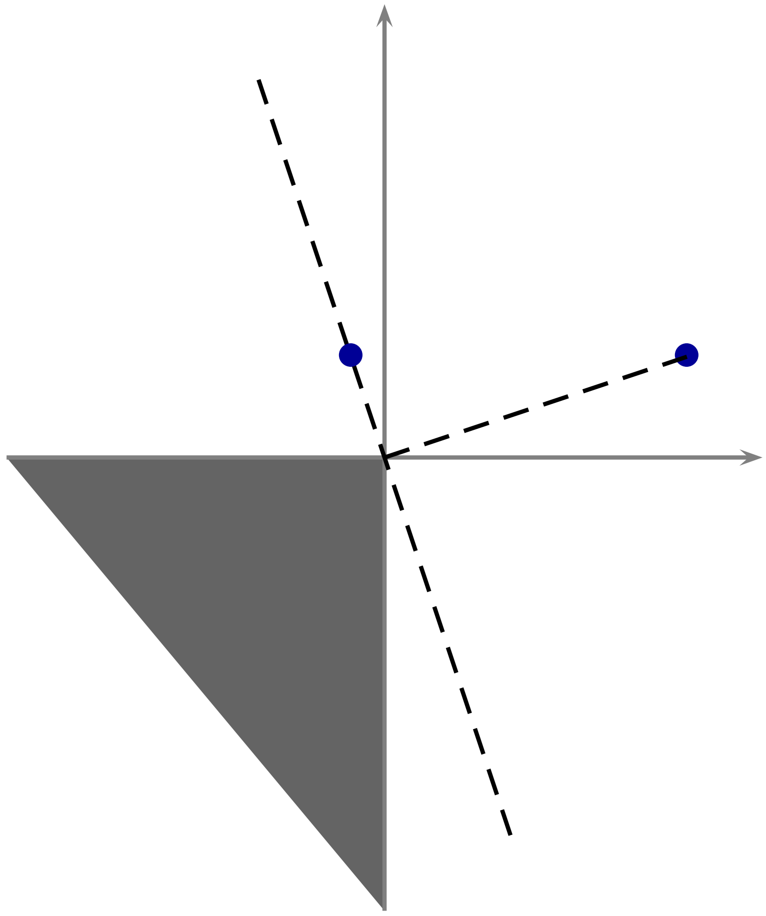
Stochastic dynamics of regret matching



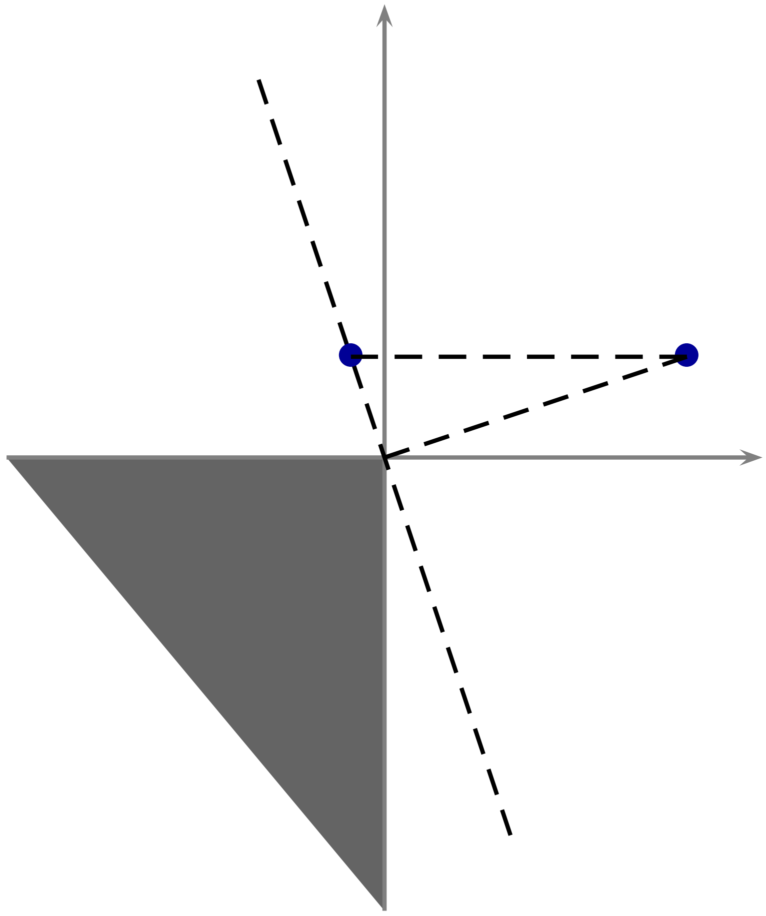
Stochastic dynamics of regret matching



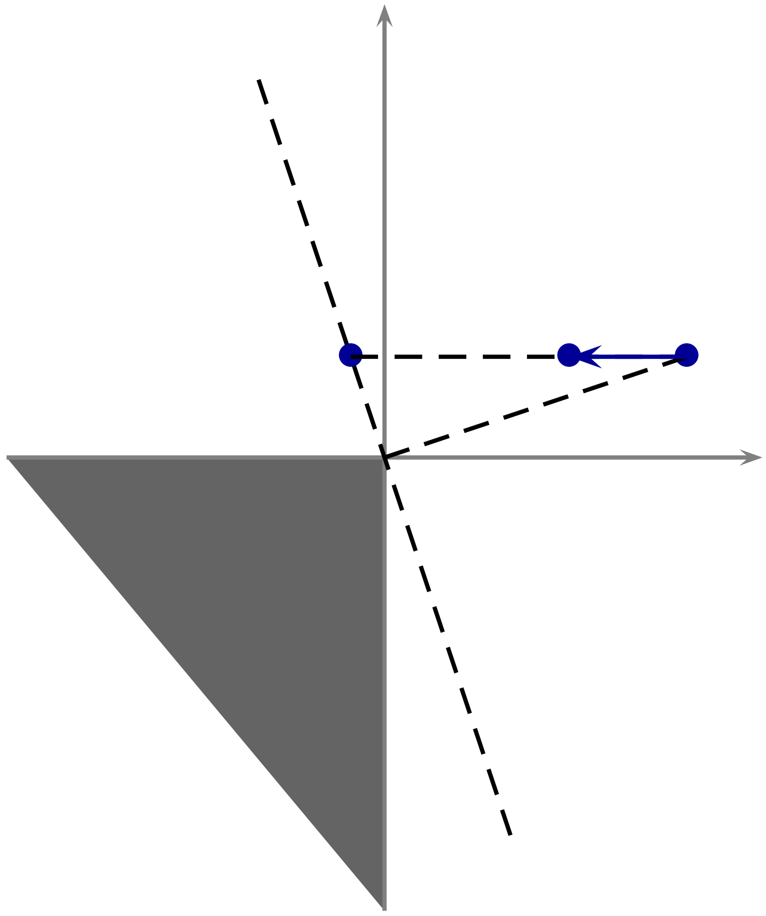
Stochastic dynamics of regret matching



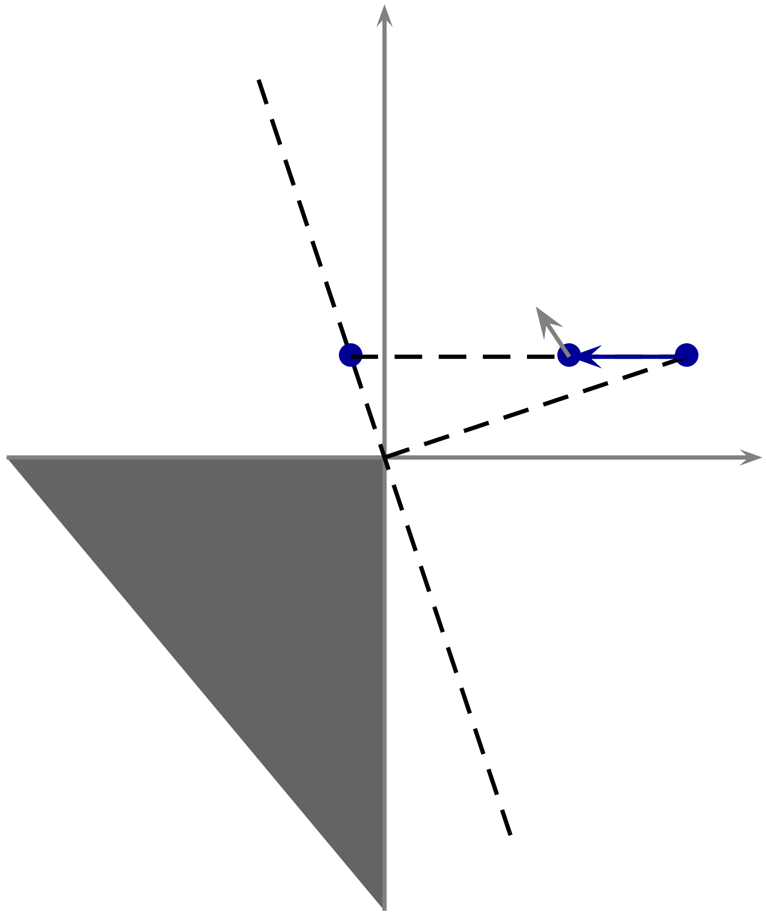
Stochastic dynamics of regret matching



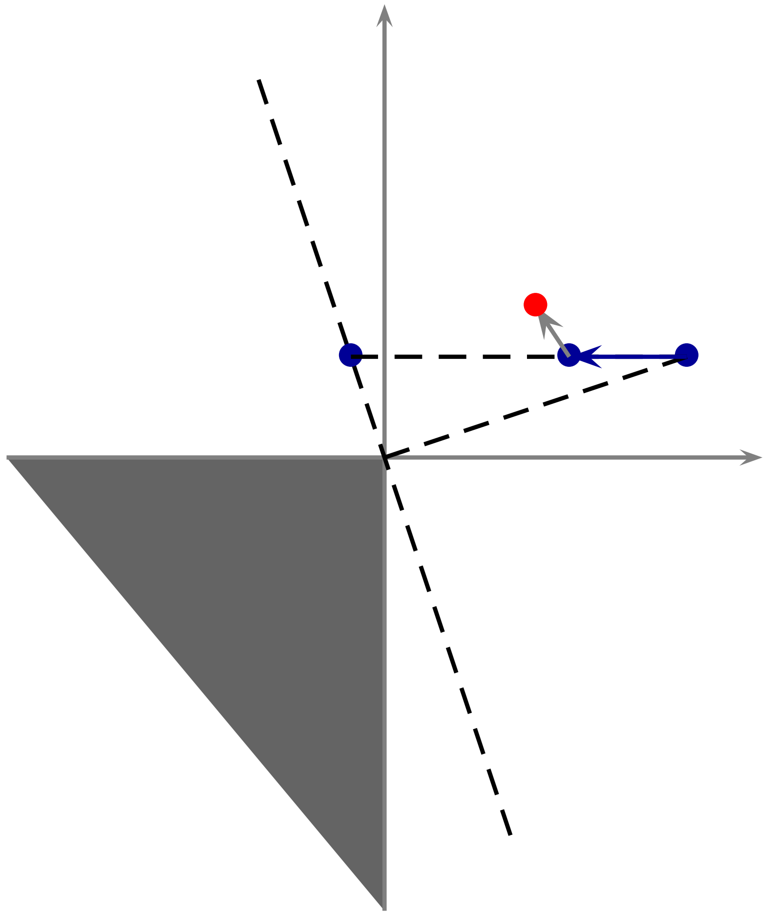
Stochastic dynamics of regret matching



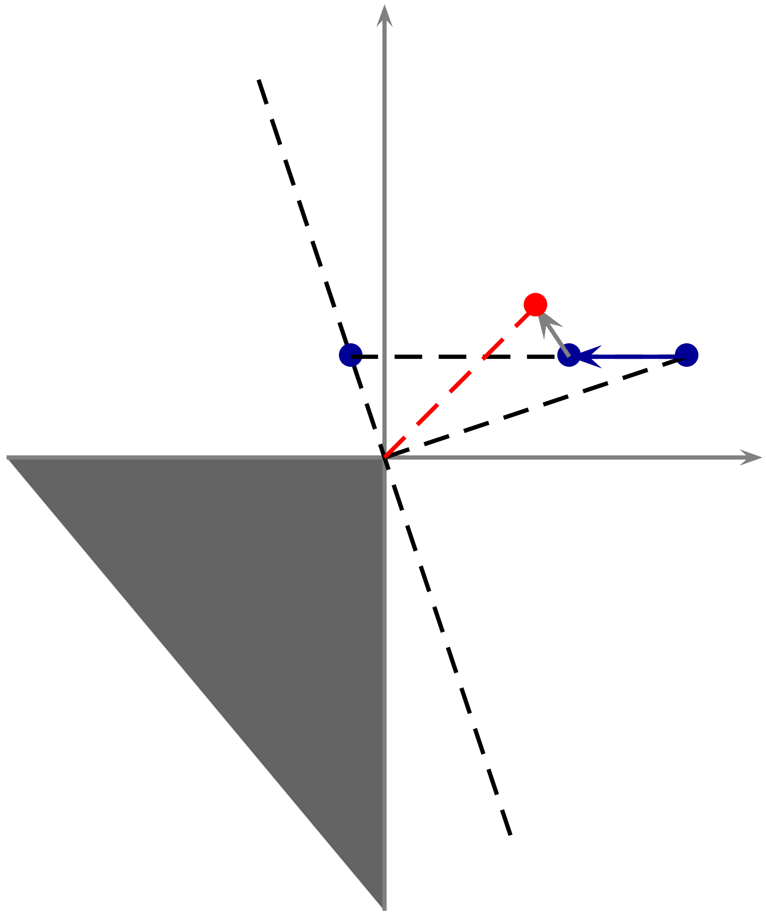
Stochastic dynamics of regret matching



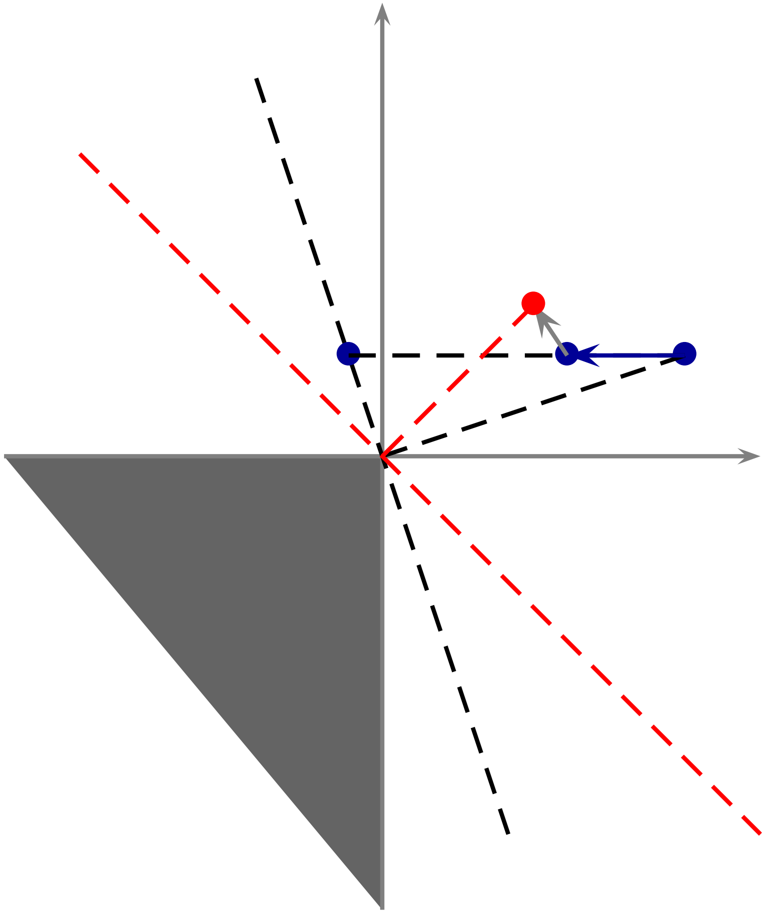
Stochastic dynamics of regret matching



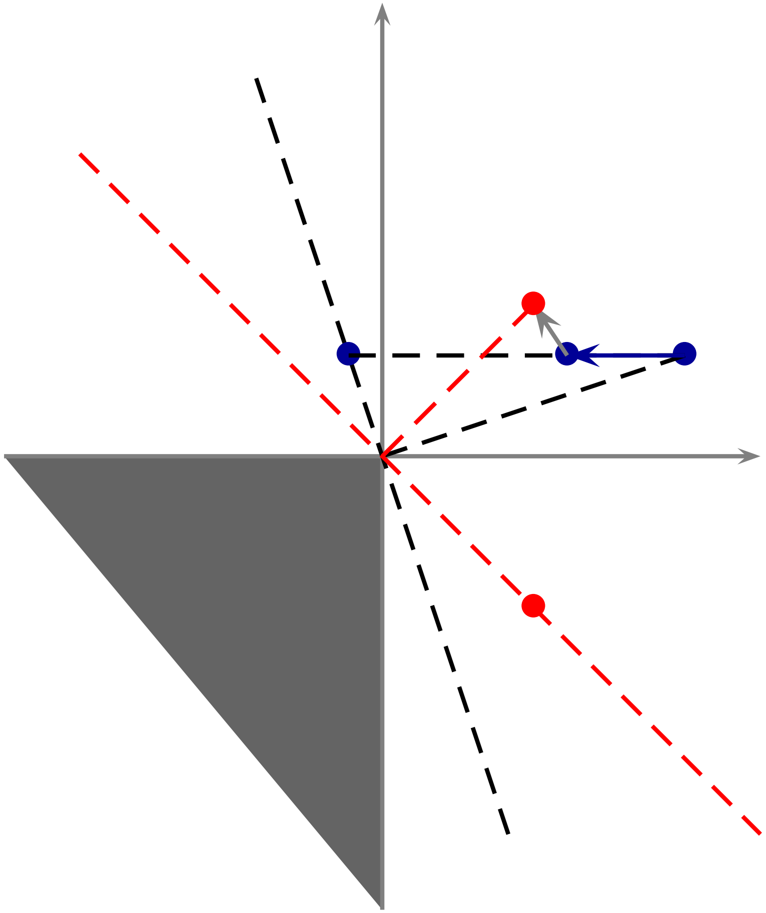
Stochastic dynamics of regret matching



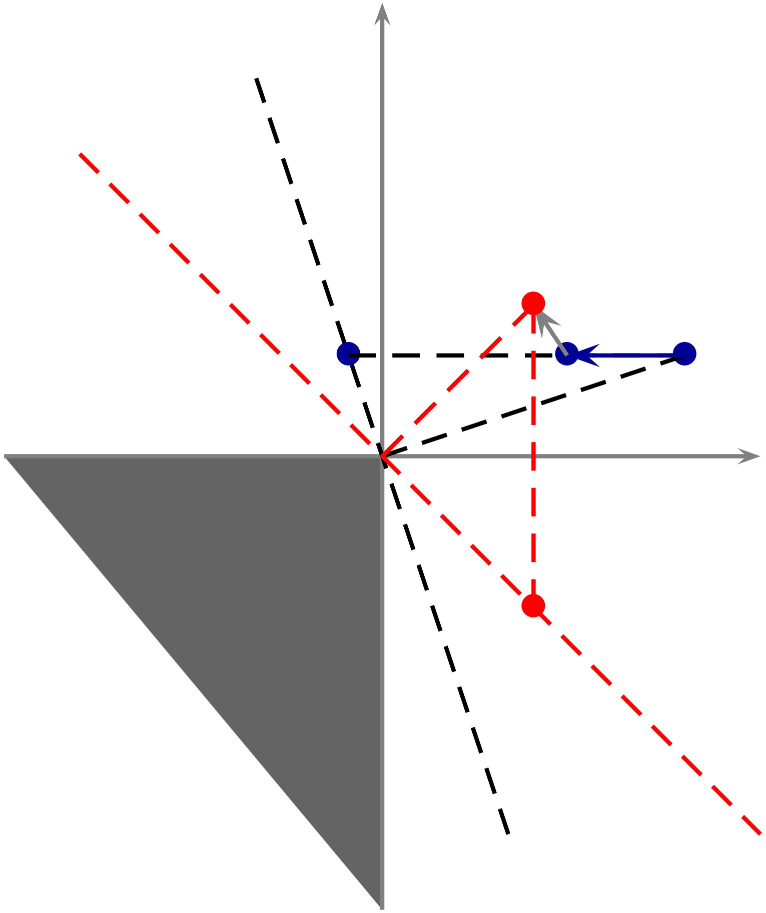
Stochastic dynamics of regret matching



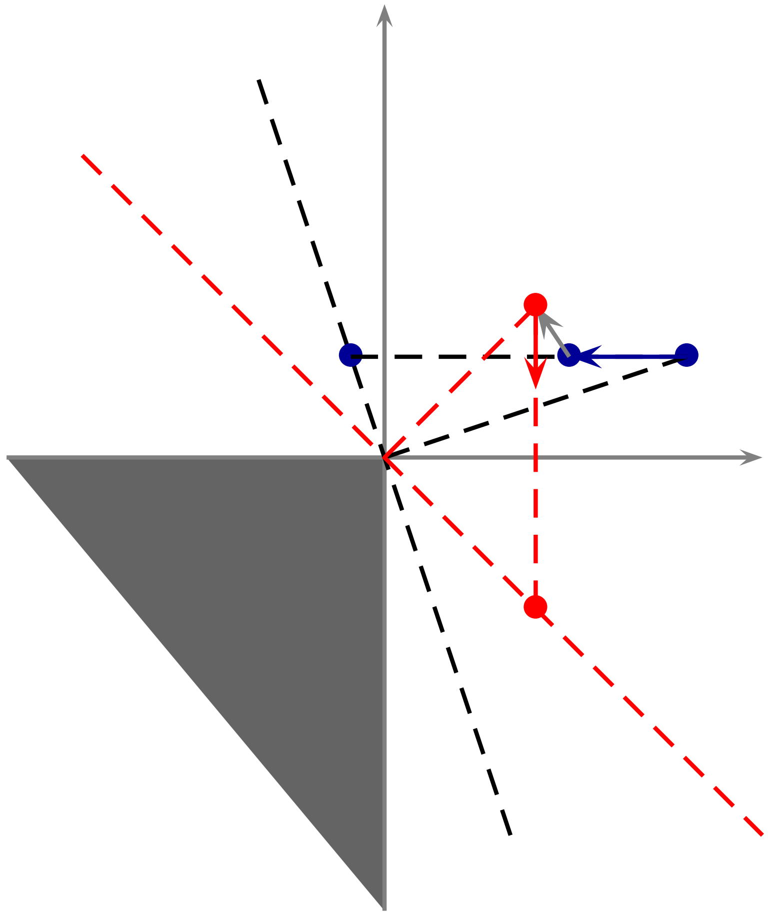
Stochastic dynamics of regret matching



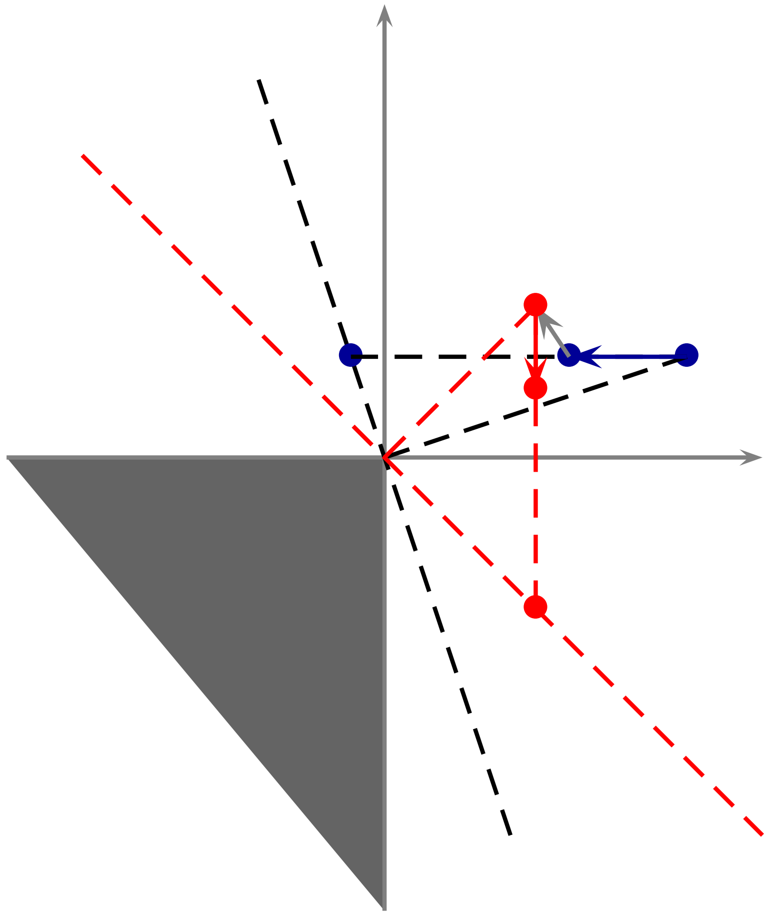
Stochastic dynamics of regret matching



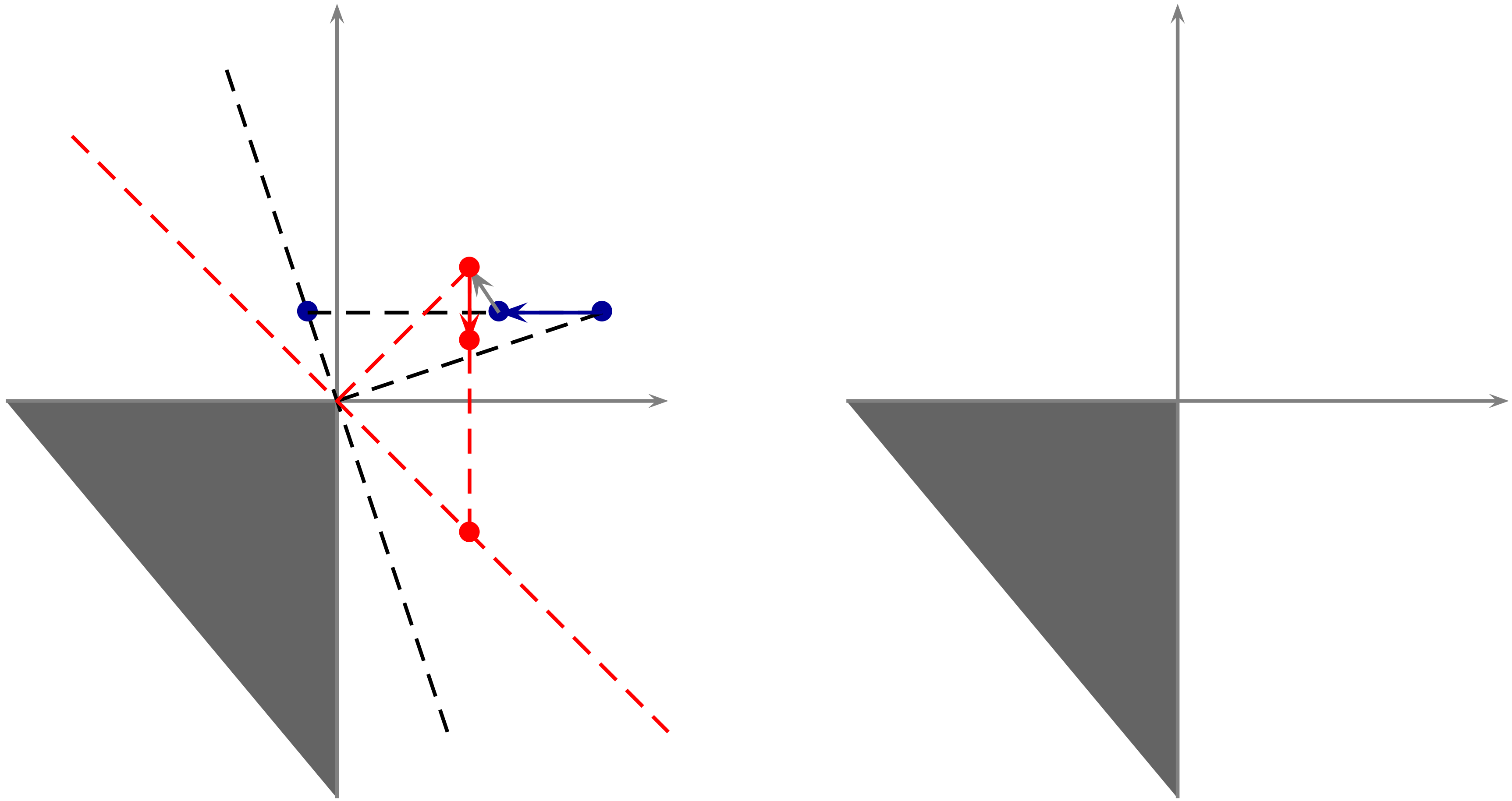
Stochastic dynamics of regret matching



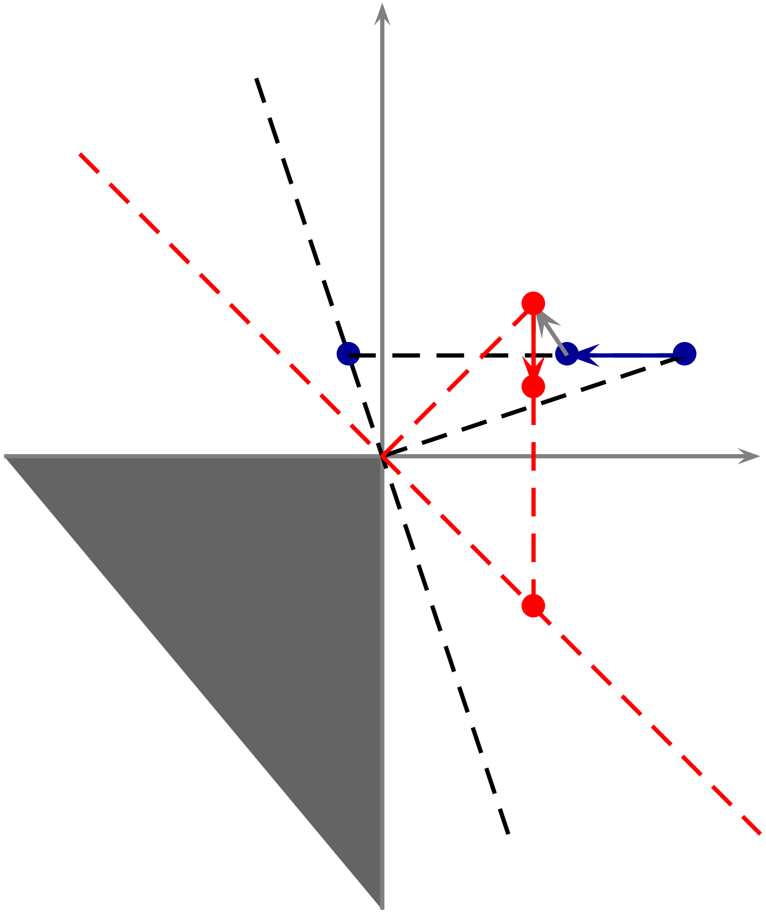
Stochastic dynamics of regret matching

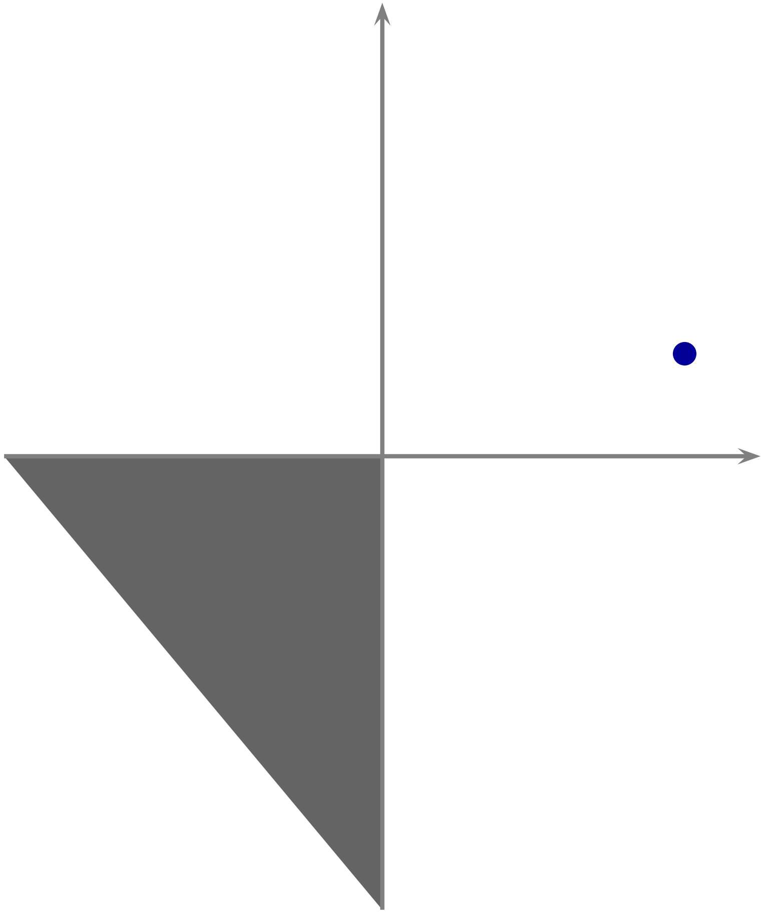


Stochastic dynamics of regret matching

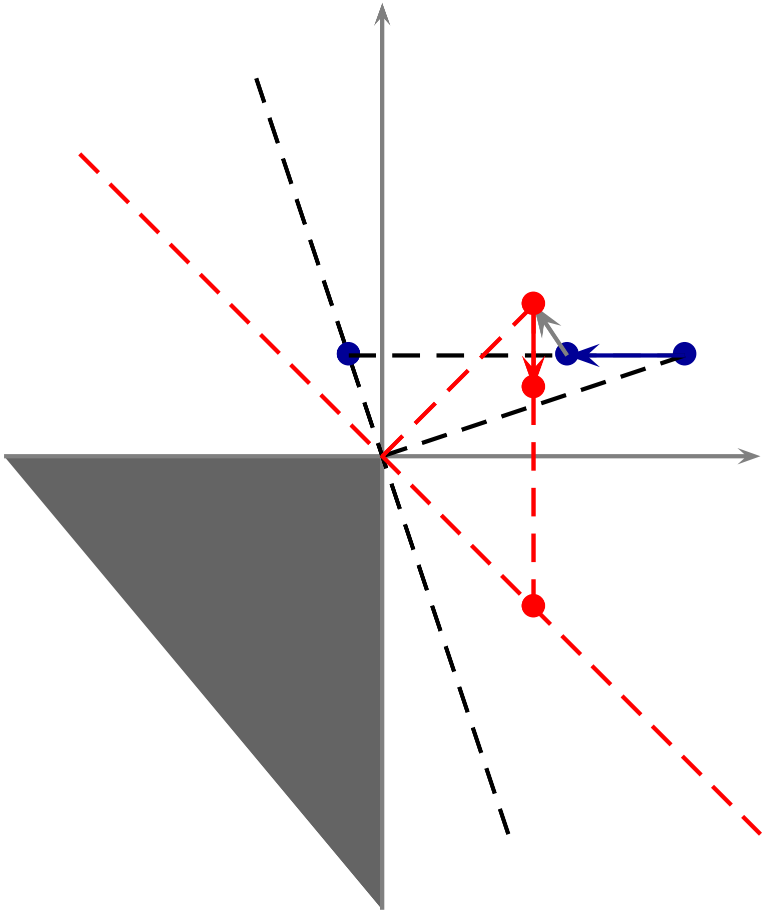


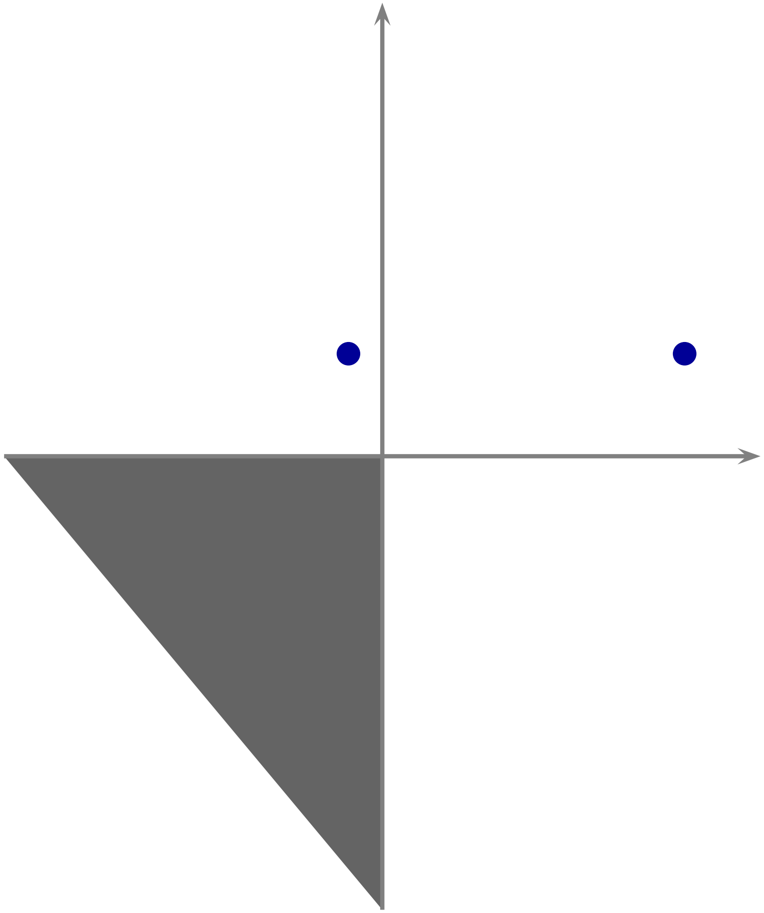
Stochastic dynamics of regret matching



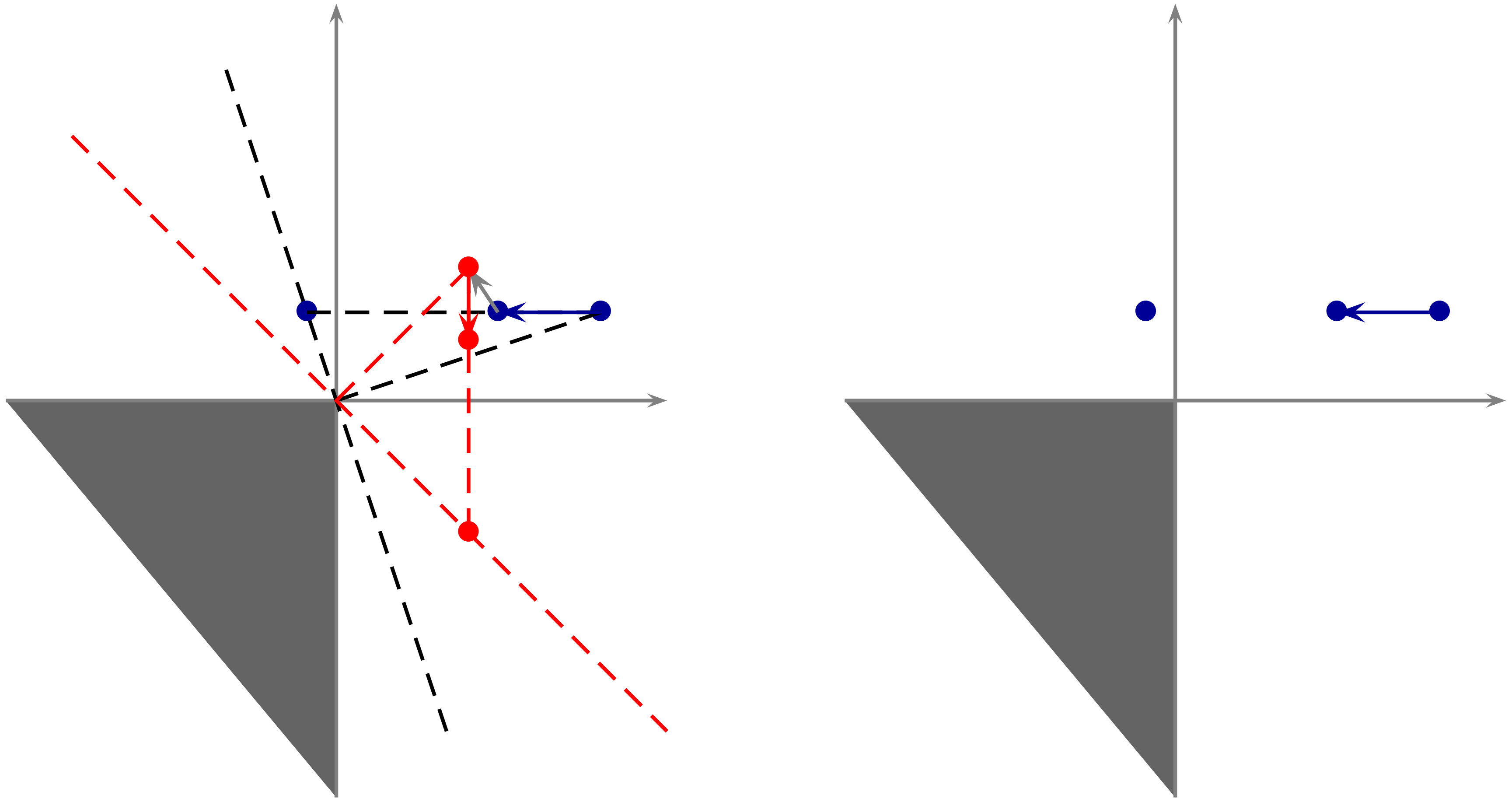


Stochastic dynamics of regret matching

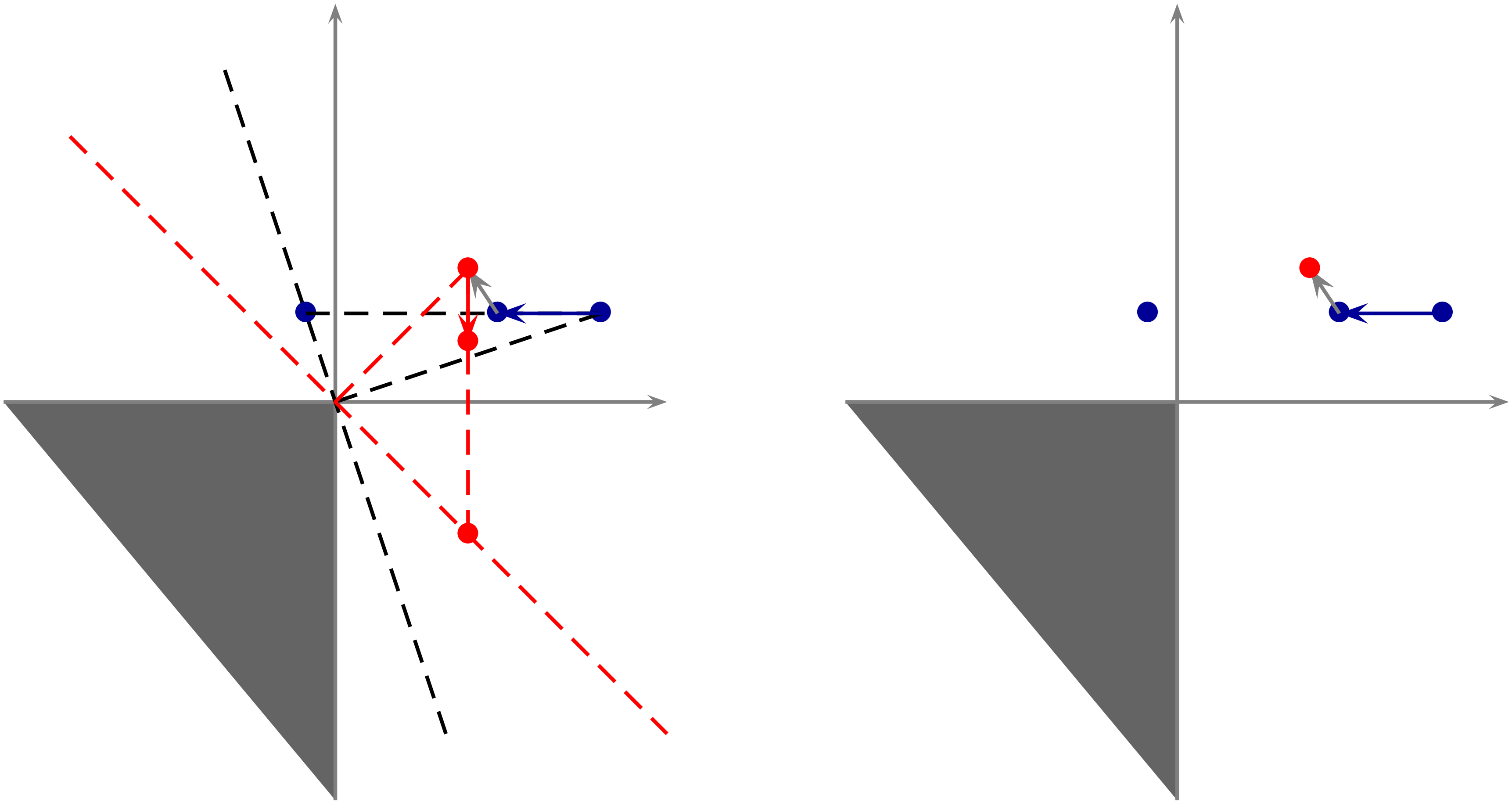




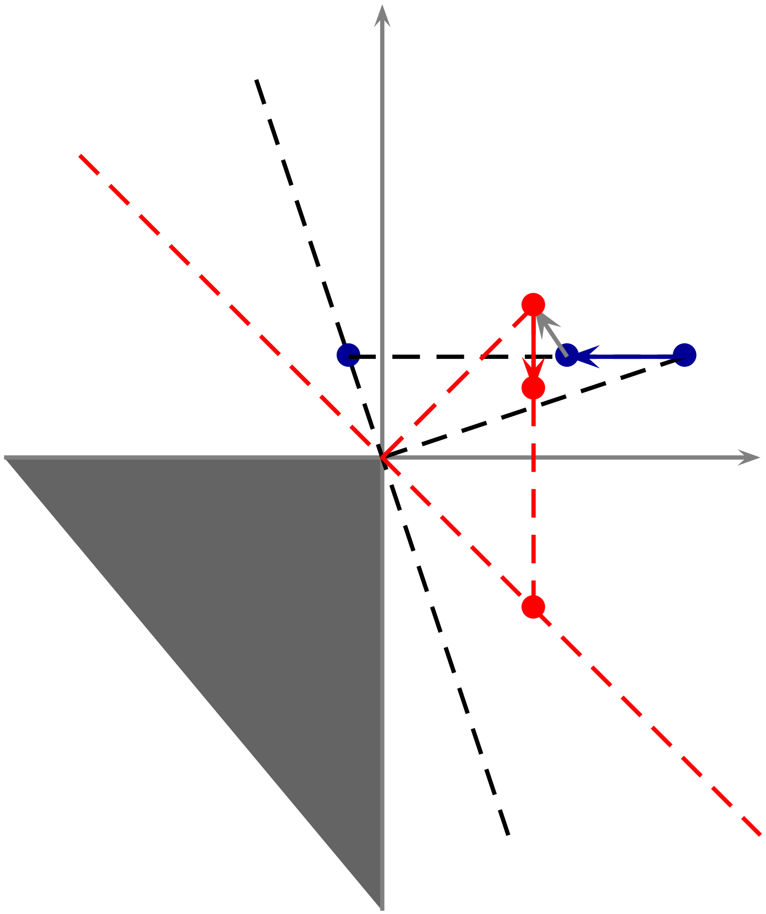
Stochastic dynamics of regret matching

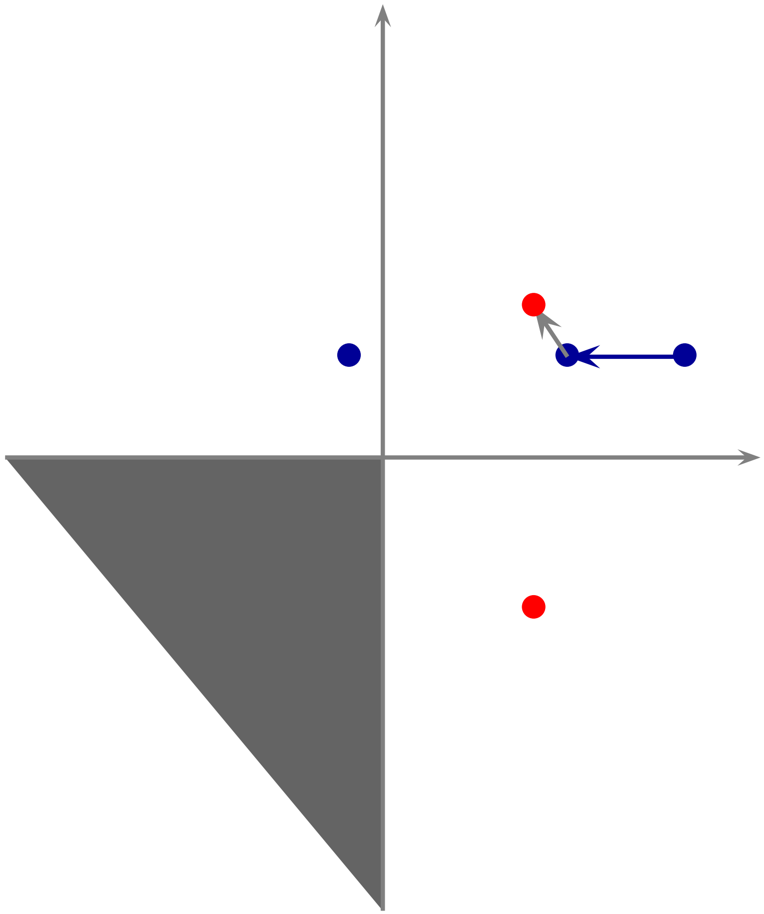


Stochastic dynamics of regret matching

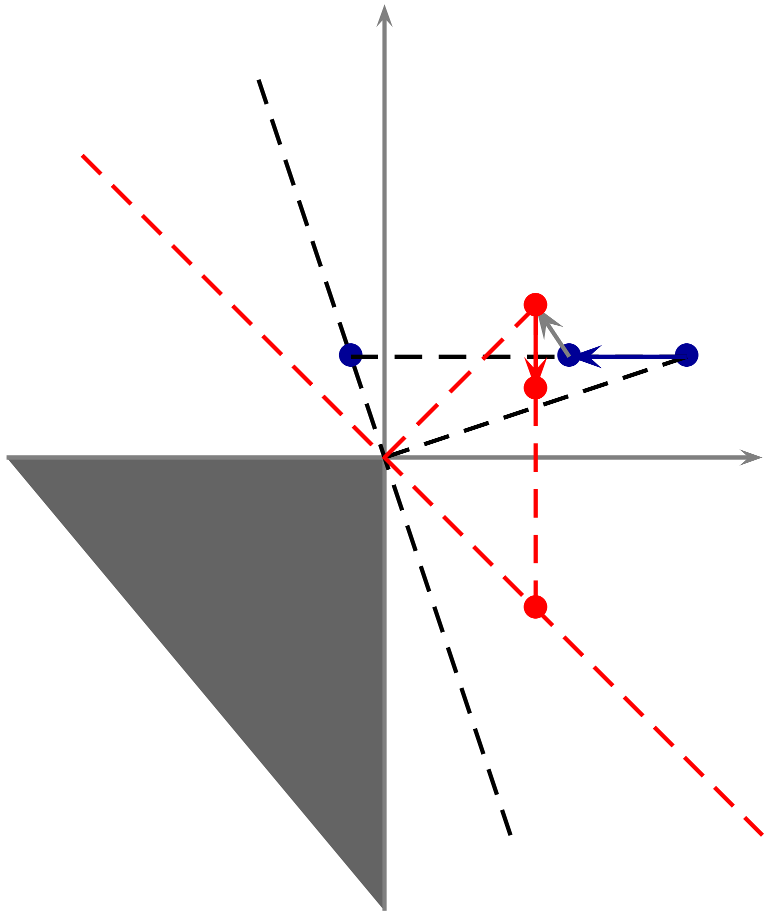


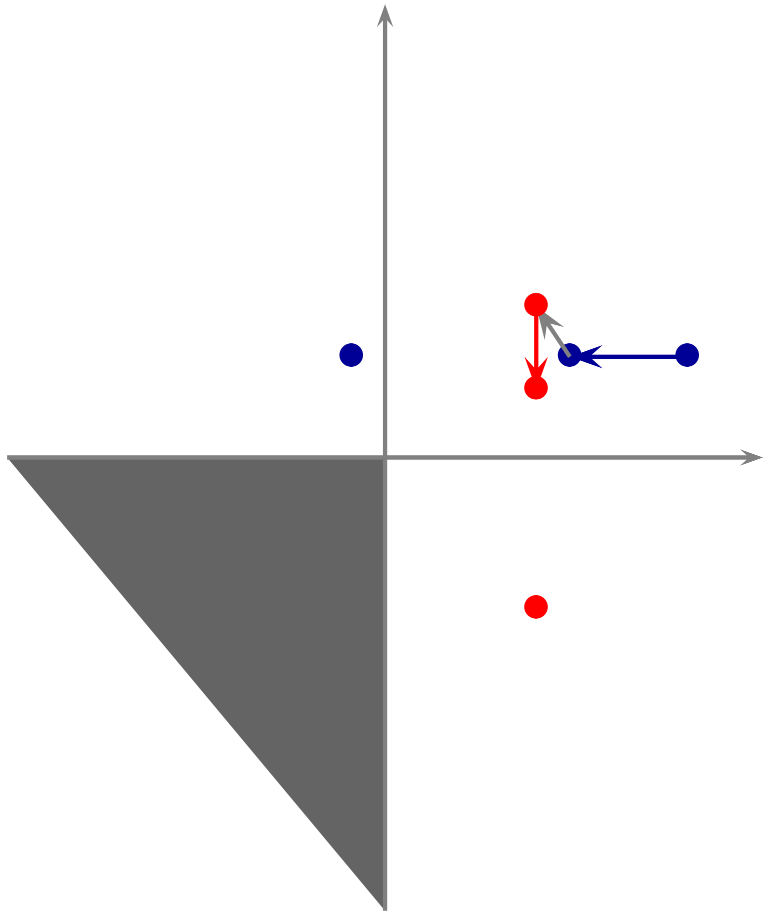
Stochastic dynamics of regret matching



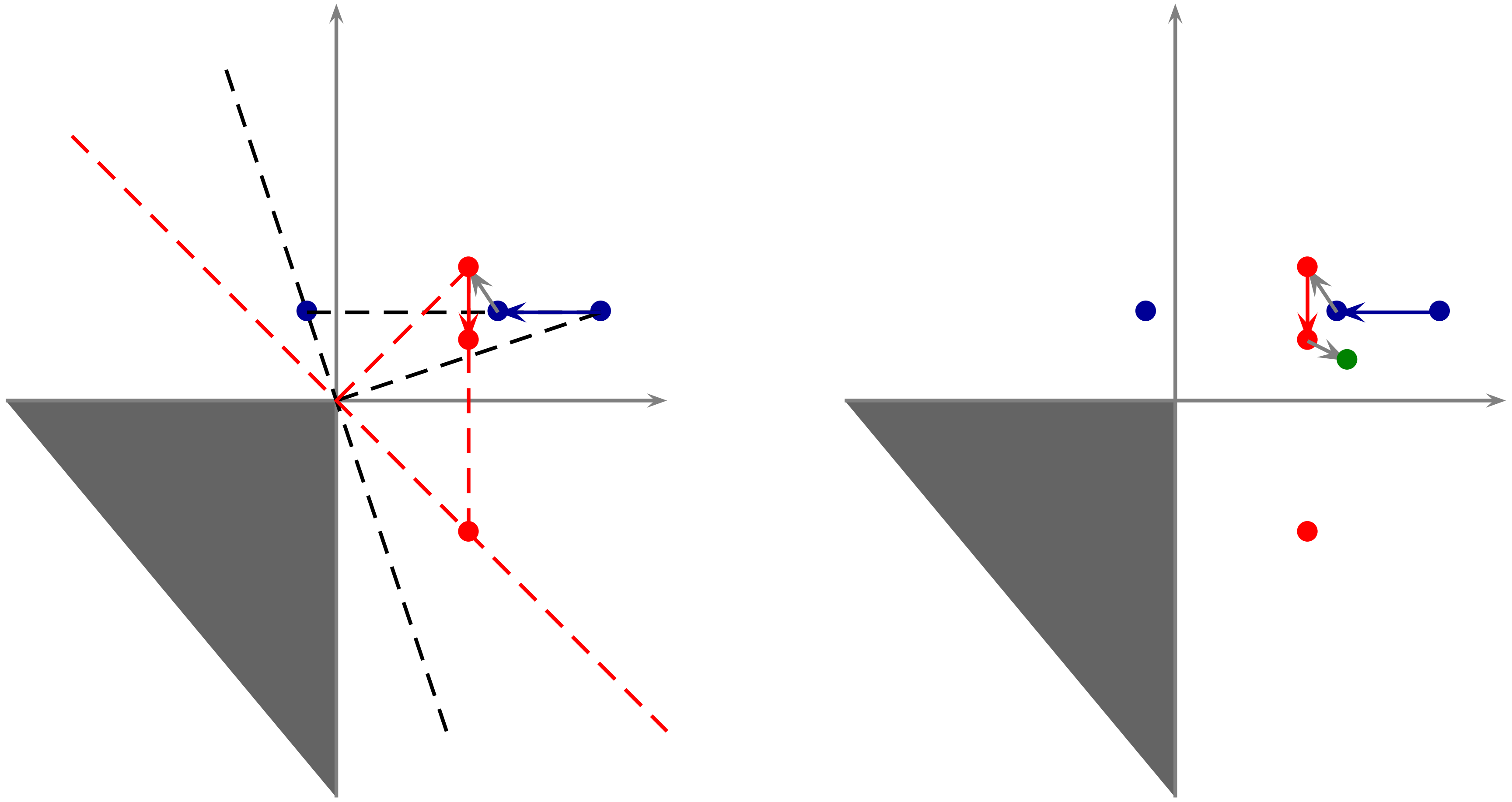


Stochastic dynamics of regret matching

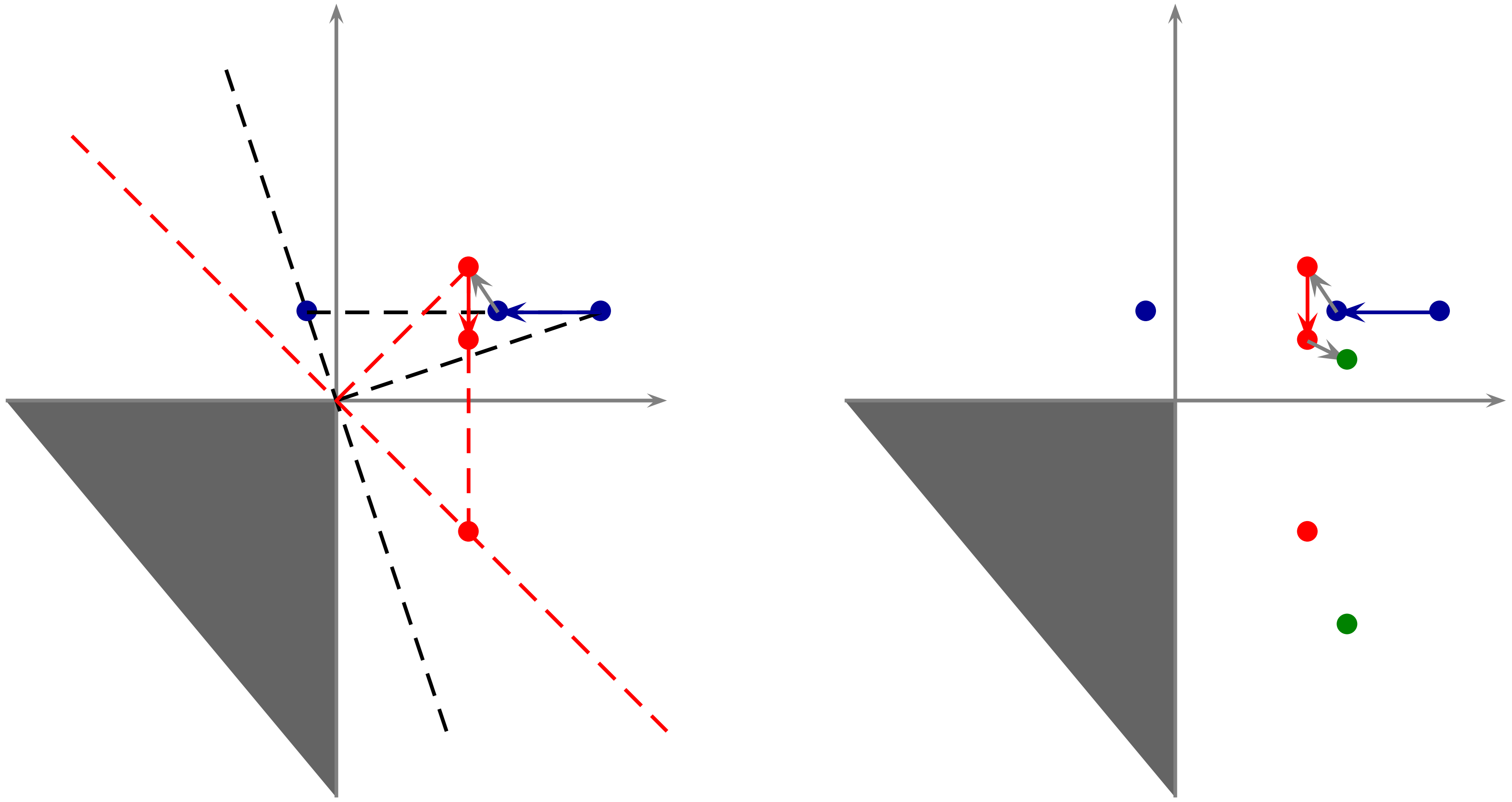




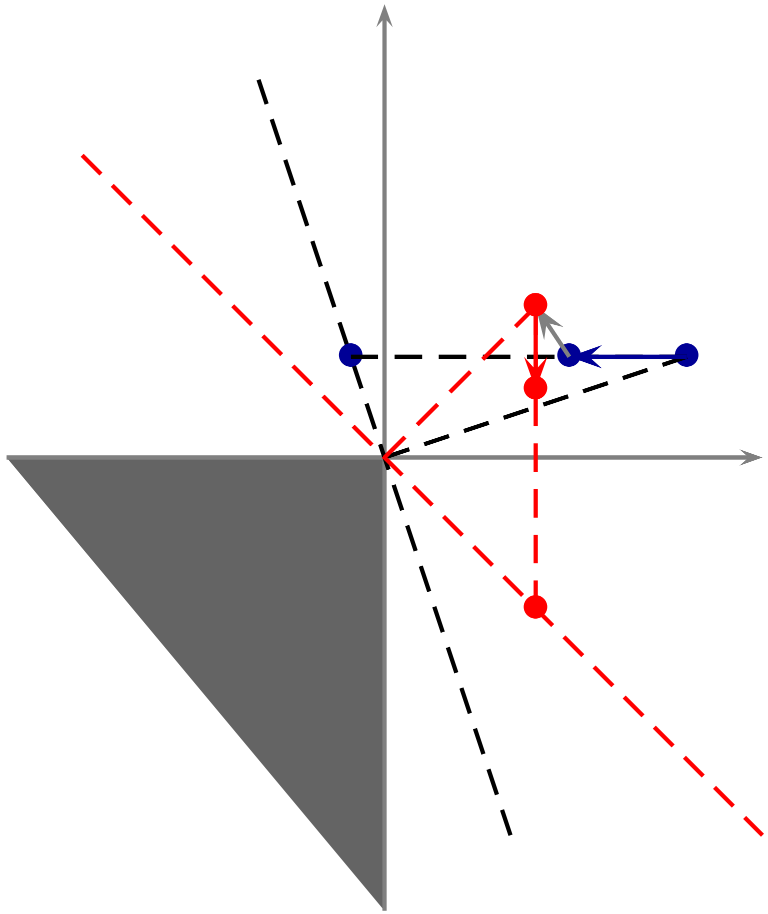
Stochastic dynamics of regret matching

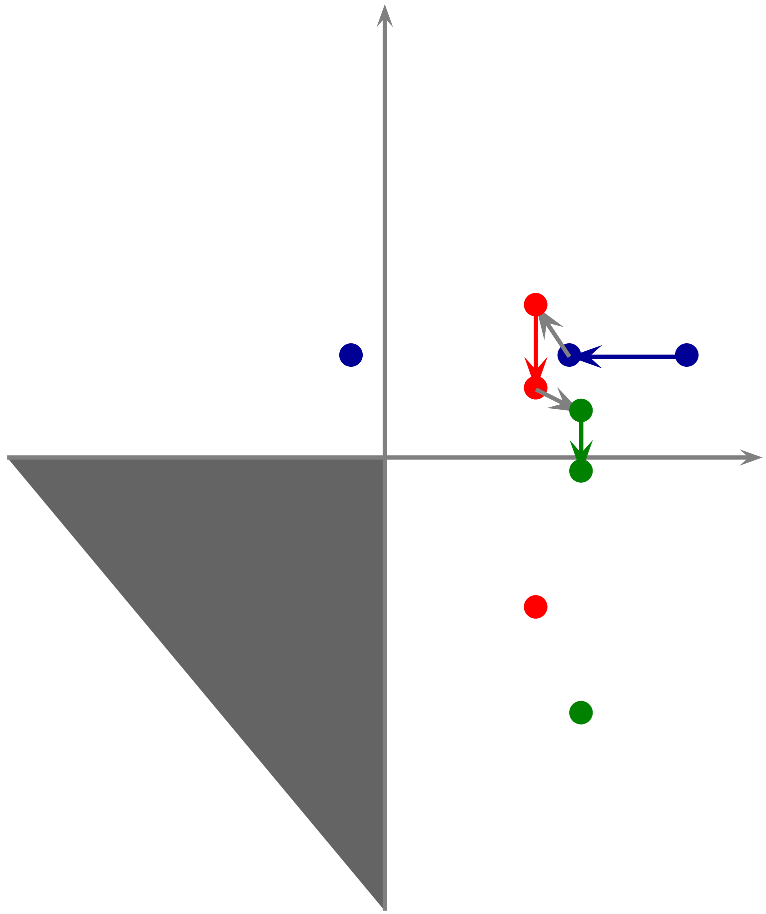


Stochastic dynamics of regret matching

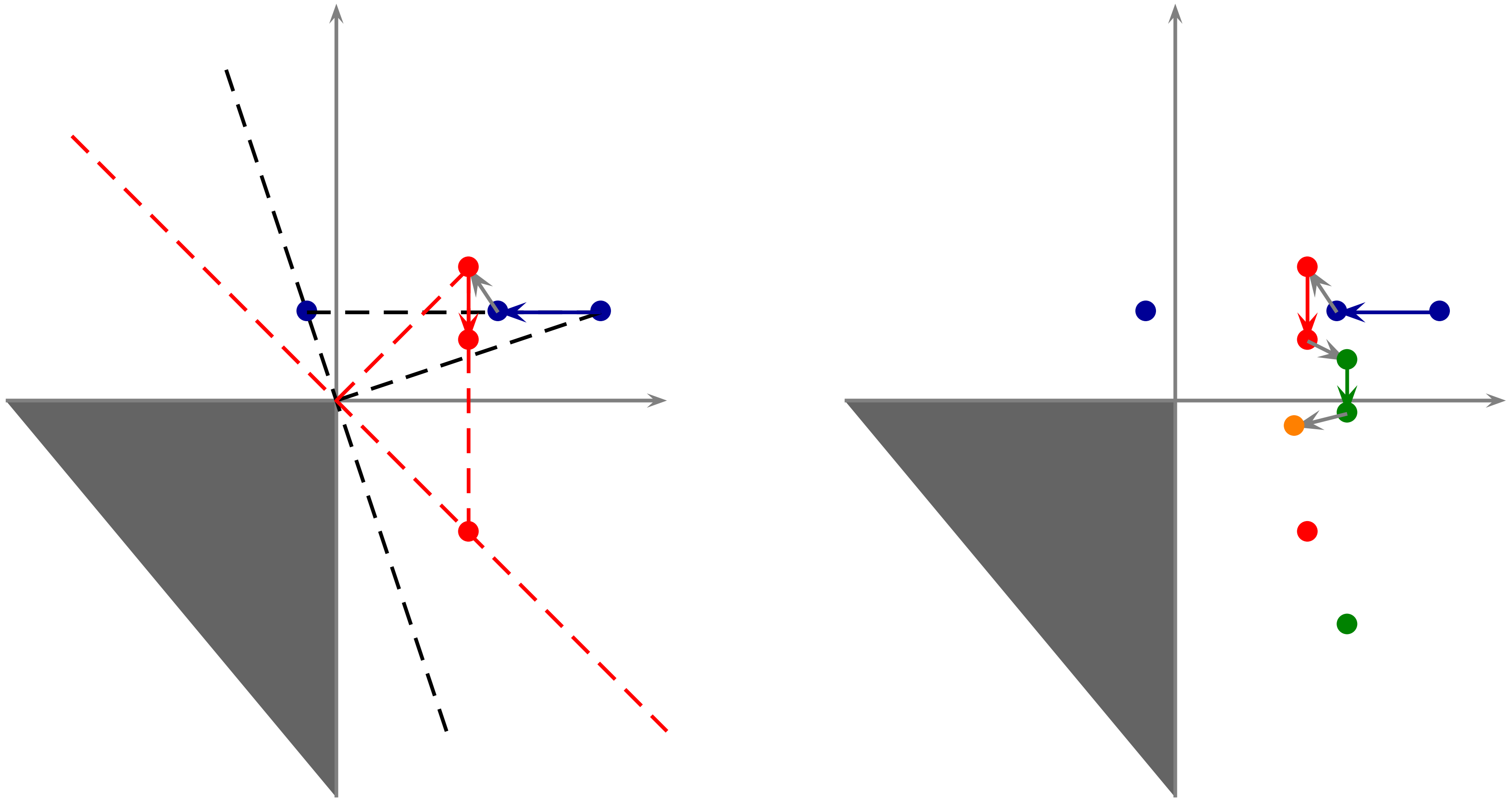


Stochastic dynamics of regret matching

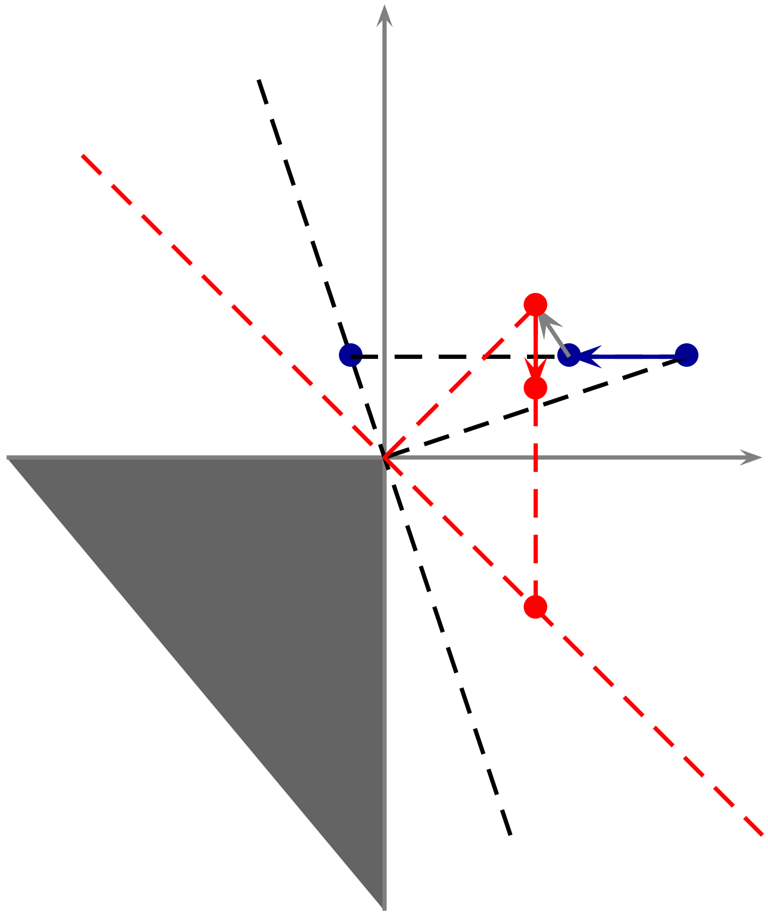


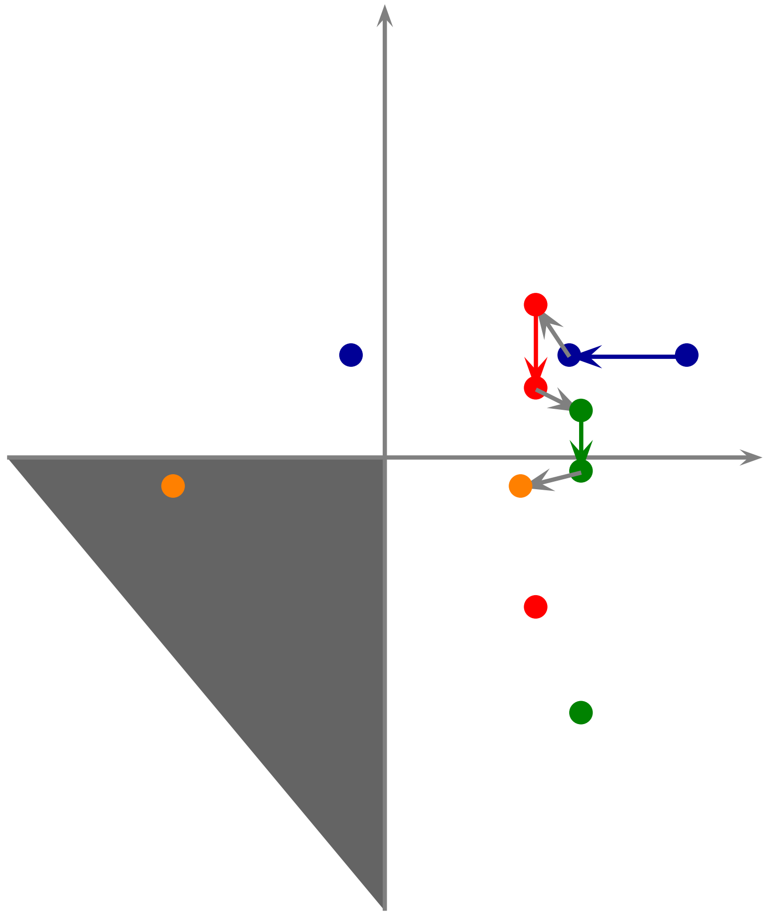


Stochastic dynamics of regret matching

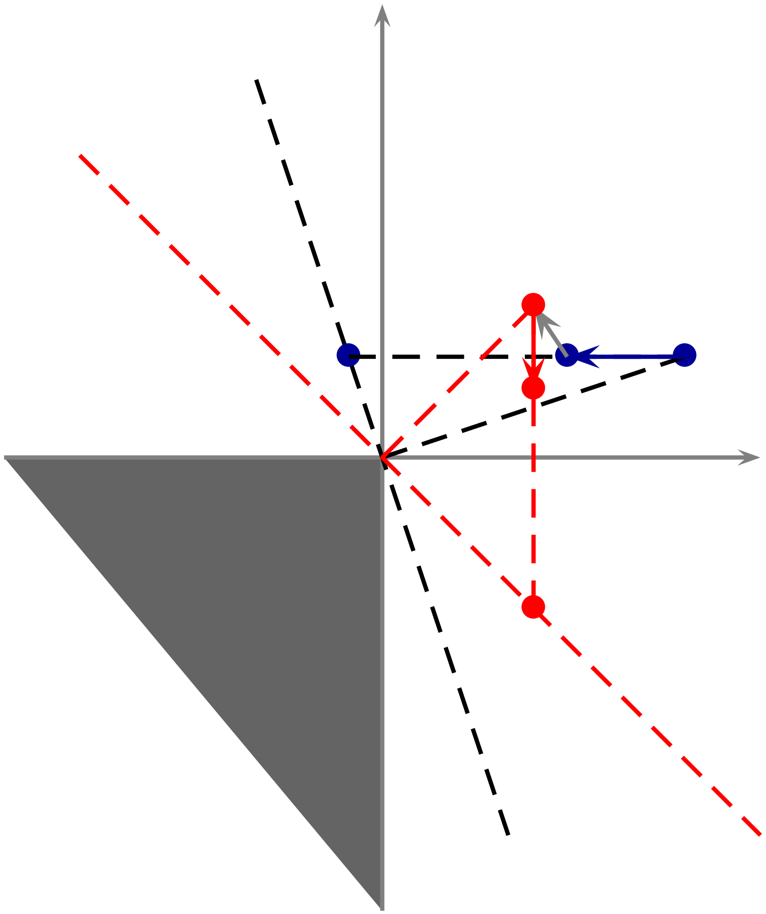


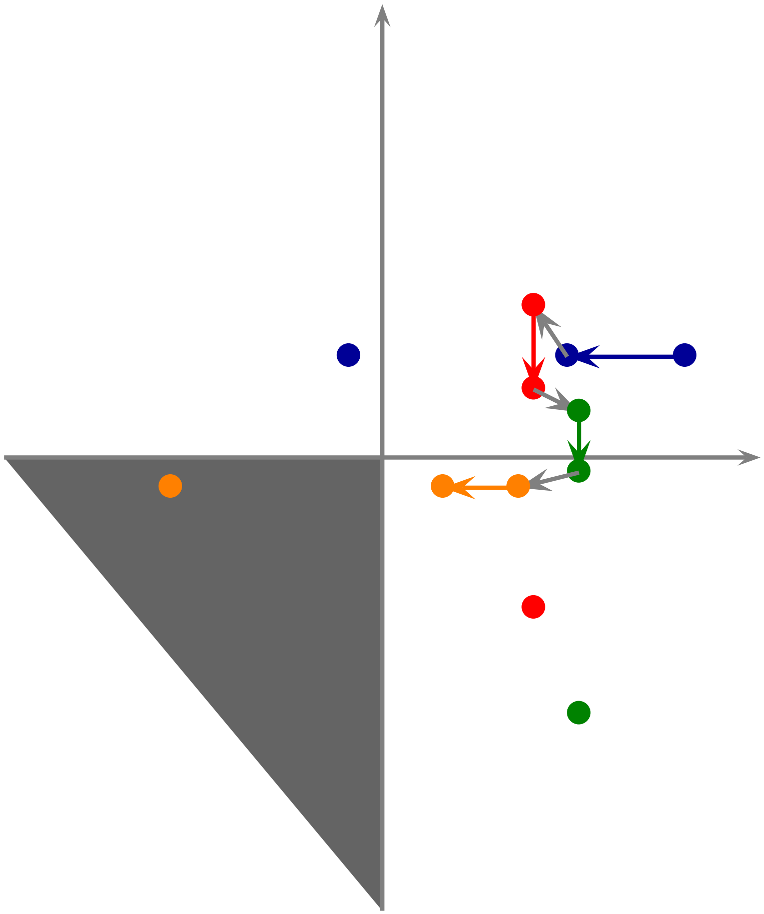
Stochastic dynamics of regret matching





Stochastic dynamics of regret matching





Part III: ϵ-Greedy Off-policy Regret Matching

ϵ-Greedy regret matching (Foster & Vohra, 1998)

ϵ-greedy regret matching. Let $\epsilon > 0$ small.

1. **Explore**. Play randomly $\epsilon\%$ of the time.
2. **Exploit**. Else, play off-policy regret matching.

ϵ-Greedy regret matching (Foster & Vohra, 1998)

ϵ-greedy regret matching. Let $\epsilon > 0$ small.

1. **Explore**. Play randomly ϵ% of the time.
2. **Exploit**. Else, play off-policy regret matching.

Define off-policy regret for x in round t as

$$\bar{r}_x^t =_{Def} \bar{u}_x^t(E) - \bar{u}^t, \qquad \text{where} \quad \bar{u}_x^t(E) = \left[\frac{1}{|E_x|} \sum_{t \in E_x} u(x^t, y^t) \right]$$

and $E_x = \{ t \mid \text{player } A \text{ experimented in round } t \text{ and played } x \}$.

ϵ-Greedy regret matching (Foster & Vohra, 1998)

ϵ-greedy regret matching. Let $\epsilon > 0$ small.

1. **Explore.** Play randomly $\epsilon\%$ of the time.
2. **Exploit.** Else, play off-policy regret matching.

Define off-policy regret for x in round t as

$$\bar{r}_x^t =_{Def} \bar{u}_x^t(E) - \bar{u}^t, \qquad \text{where} \quad \bar{u}_x^t(E) = \left[\frac{1}{|E_x|} \sum_{t \in E_x} u(x^t, y^t) \right]$$

and $E_x = \{ t \mid \text{player } A \text{ experimented in round } t \text{ and played } x \}$.

- Proposed as a forecasting heuristic by Foster and Vohra (1993).

ϵ-Greedy regret matching (Foster & Vohra, 1998)

ϵ-greedy regret matching. Let $\epsilon > 0$ small.

1. **Explore**. Play randomly ϵ% of the time.
2. **Exploit**. Else, play off-policy regret matching.

Define off-policy regret for x in round t as

$$\bar{r}^t_x =_{Def} \bar{u}^t_x(E) - \bar{u}^t, \qquad \text{where} \quad \bar{u}^t_x(E) = \left[\frac{1}{|E_x|} \sum_{t \in E_x} u(x^t, y^t) \right]$$

and $E_x = \{ t \mid \text{player } A \text{ experimented in round } t \text{ and played } x \}$.

- Proposed as a forecasting heuristic by Foster and Vohra (1993).
- Does not need to know the actions of its opponents.

ϵ-Greedy regret matching (Foster & Vohra, 1998)

ϵ-greedy regret matching. Let $\epsilon > 0$ small.

1. **Explore**. Play randomly ϵ% of the time.
2. **Exploit**. Else, play off-policy regret matching.

Define off-policy regret for x in round t as

$$\bar{r}_x^t =_{Def} \bar{u}_x^t(E) - \bar{u}^t, \qquad \text{where} \quad \bar{u}_x^t(E) = \left[\frac{1}{|E_x|} \sum_{t \in E_x} u(x^t, y^t) \right]$$

and $E_x = \{ \, t \mid \text{player } A \text{ experimented in round } t \text{ and played } x \, \}$.

- Proposed as a forecasting heuristic by Foster and Vohra (1993).
- Does not need to know the actions of its opponents.
- Turns out to estimate regrets.

ϵ-Greedy regret matching (outline of proof)

Theorem (Foster *et al.*, 1998). *For all* $\delta > 0$ *there exists an* $\epsilon > 0$ *such that* ϵ*-greedy regret matching has at most* δ *regret.*

ϵ-Greedy regret matching (outline of proof)

Theorem (Foster *et al.*, 1998). *For all* $\delta > 0$ *there exists an* $\epsilon > 0$ *such that* ϵ*-greedy regret matching has at most* δ *regret.*

If $\epsilon_t \to 0$ *at a rate* $\mathcal{O}(t^{-1/3})$*, there is no regret in the long run.*

ϵ-Greedy regret matching (outline of proof)

Theorem (Foster *et al.*, 1998). *For all $\delta > 0$ there exists an $\epsilon > 0$ such that ϵ-greedy regret matching has at most δ regret.*

If $\epsilon_t \to 0$ at a rate $\mathcal{O}(t^{-1/3})$, there is no regret in the long run.

Proof. Suppose there are k different actions.

ϵ-Greedy regret matching (outline of proof)

Theorem (Foster *et al.*, 1998). *For all $\delta > 0$ there exists an $\epsilon > 0$ such that ϵ-greedy regret matching has at most δ regret.*

If $\epsilon_t \to 0$ at a rate $\mathcal{O}(t^{-1/3})$, there is no regret in the long run.

Proof. Suppose there are k different actions. Let $e^t \in R^k$ such that

$$e^t_x = (\text{ player } A \text{ explores at } t \text{ and chooses } x\ ) \ ?\ 1 : 0.$$

ϵ-Greedy regret matching (outline of proof)

Theorem (Foster *et al.*, 1998). *For all $\delta > 0$ there exists an $\epsilon > 0$ such that ϵ-greedy regret matching has at most δ regret.*

If $\epsilon_t \to 0$ at a rate $\mathcal{O}(t^{-1/3})$, there is no regret in the long run.

Proof. Suppose there are k different actions. Let $e^t \in R^k$ such that

$$e^t_x = (\,\text{player } A \text{ explores at } t \text{ and chooses } x\,)\ ?\ 1 : 0.$$

For each action i

$$\Pr(x^t = i \mid A \text{ explores at round } t) = \frac{1}{k}.$$

ϵ-Greedy regret matching (outline of proof)

Theorem (Foster *et al.*, 1998). *For all $\delta > 0$ there exists an $\epsilon > 0$ such that ϵ-greedy regret matching has at most δ regret.*

If $\epsilon_t \to 0$ at a rate $\mathcal{O}(t^{-1/3})$, there is no regret in the long run.

Proof. Suppose there are k different actions. Let $e^t \in R^k$ such that

$$e^t_x = (\text{ player } A \text{ explores at } t \text{ and chooses } x\ ) \ ?\ 1 : 0.$$

For each action i

$$\Pr(x^t = i \mid A \text{ explores at round } t) = \frac{1}{k}.$$

It follows that

$$E[e^t] = \left(\frac{\epsilon}{k}, \ldots, \frac{\epsilon}{k}\right).$$

ϵ-Greedy regret matching (outline of proof)

Define

$$z_x^t =_{Def} \left(\frac{k}{\epsilon} \cdot e_x^t \cdot u(x, y^t) \right) - u(x, y^t).$$

ϵ-Greedy regret matching (outline of proof)

Define

$$z_x^t =_{Def} \left(\frac{k}{\epsilon} \cdot e_x^t \cdot u(x, y^t) \right) \; - \; u(x, y^t).$$

In words, z_x^t is the difference between the properly magnified empirical payoff for x and the (correct but) virtual payoff for x.

ϵ-Greedy regret matching (outline of proof)

Define

$$z_x^t =_{Def} \left(\frac{k}{\epsilon} \cdot e_x^t \cdot u(x, y^t) \right) - u(x, y^t).$$

In words, z_x^t is the difference between the properly magnified empirical payoff for x and the (correct but) virtual payoff for x. It follows that

ϵ-Greedy regret matching (outline of proof)

Define

$$z^t_x =_{Def} \left(\frac{k}{\epsilon} \cdot e^t_x \cdot u(x, y^t) \right) - u(x, y^t).$$

In words, z^t_x is the difference between the properly magnified empirical payoff for x and the (correct but) virtual payoff for x. It follows that

$$E[z^t_x]$$

ϵ-Greedy regret matching (outline of proof)

Define

$$z_x^t =_{Def} \left(\frac{k}{\epsilon} \cdot e_x^t \cdot u(x, y^t) \right) - u(x, y^t).$$

In words, z_x^t is the difference between the properly magnified empirical payoff for x and the (correct but) virtual payoff for x. It follows that

$$E[z_x^t] = E\left[\left(\frac{k}{\epsilon} \cdot e_x^t \cdot u(x, y^t) \right) - u(x, y^t) \right]$$

ϵ-Greedy regret matching (outline of proof)

Define

$$z_x^t =_{Def} \left(\frac{k}{\epsilon} \cdot e_x^t \cdot u(x, y^t) \right) - u(x, y^t).$$

In words, z_x^t is the difference between the properly magnified empirical payoff for x and the (correct but) virtual payoff for x. It follows that

$$\begin{aligned} E[z_x^t] &= E\left[\left(\frac{k}{\epsilon} \cdot e_x^t \cdot u(x, y^t) \right) - u(x, y^t) \right] \\ &= \frac{k}{\epsilon} \cdot E\left[e_x^t \cdot u(x, y^t) \right] - E[u(x, y^t)] \end{aligned}$$

ϵ-Greedy regret matching (outline of proof)

Define

$$z_x^t =_{Def} \left(\frac{k}{\epsilon} \cdot e_x^t \cdot u(x, y^t) \right) - u(x, y^t).$$

In words, z_x^t is the difference between the properly magnified empirical payoff for x and the (correct but) virtual payoff for x. It follows that

$$\begin{aligned} E[z_x^t] &= E\left[\left(\frac{k}{\epsilon} \cdot e_x^t \cdot u(x, y^t) \right) - u(x, y^t) \right] \\ &= \frac{k}{\epsilon} \cdot E\left[e_x^t \cdot u(x, y^t) \right] - E[u(x, y^t)] \\ &= \frac{k}{\epsilon} \cdot E[e_x^t] \cdot E[u(x, y^t)] - E[u(x, y^t)] \quad (e^t \text{ and } u^t \text{ are independent}) \end{aligned}$$

ϵ-Greedy regret matching (outline of proof)

Define

$$z_x^t =_{Def} \left(\frac{k}{\epsilon} \cdot e_x^t \cdot u(x, y^t) \right) - u(x, y^t).$$

In words, z_x^t is the difference between the properly magnified empirical payoff for x and the (correct but) virtual payoff for x. It follows that

$$\begin{aligned}
E[z_x^t] &= E\left[\left(\frac{k}{\epsilon} \cdot e_x^t \cdot u(x, y^t) \right) - u(x, y^t) \right] \\
&= \frac{k}{\epsilon} \cdot E\left[e_x^t \cdot u(x, y^t) \right] - E[u(x, y^t)] \\
&= \frac{k}{\epsilon} \cdot E[e_x^t] \cdot E[u(x, y^t)] - E[u(x, y^t)] \quad (e^t \text{ and } u^t \text{ are independent}) \\
&= \frac{k}{\epsilon} \cdot \frac{\epsilon}{k} \cdot E[u(x, y^t)] - E[u(x, y^t)]
\end{aligned}$$

ϵ-Greedy regret matching (outline of proof)

Define

$$z_x^t =_{Def} \left(\frac{k}{\epsilon} \cdot e_x^t \cdot u(x, y^t) \right) - u(x, y^t).$$

In words, z_x^t is the difference between the properly magnified empirical payoff for x and the (correct but) virtual payoff for x. It follows that

$$\begin{aligned}
E[z_x^t] &= E\left[\left(\frac{k}{\epsilon} \cdot e_x^t \cdot u(x, y^t) \right) - u(x, y^t) \right] \\
&= \frac{k}{\epsilon} \cdot E\left[e_x^t \cdot u(x, y^t) \right] - E[u(x, y^t)] \\
&= \frac{k}{\epsilon} \cdot E[e_x^t] \cdot E[u(x, y^t)] - E[u(x, y^t)] \quad (e^t \text{ and } u^t \text{ are independent}) \\
&= \frac{k}{\epsilon} \cdot \frac{\epsilon}{k} \cdot E[u(x, y^t)] - E[u(x, y^t)] \\
&= 0.
\end{aligned}$$

ϵ-Greedy regret matching (outline of proof)

ϵ-Greedy regret matching (outline of proof)

Strong law of large numbers for dependent random variables. Let $\{w^t\}^t$ be a bounded sequence of possibly dependent random variables in R^k. Let $z^t = E[\, w^t \mid w^{t-1}, w^{t-2}, \dots, w^1\,] - w^t$, and $\bar{z}^t$ the average of the z^t's. Then $\lim_{t\to\infty} \bar{z}^t = 0$ with probability one.[a]

[a] PY refers to Loève, 1978, Book II, Th. 32.E.1.

ϵ-Greedy regret matching (outline of proof)

Strong law of large numbers for dependent random variables. Let $\{w^t\}^t$ be a bounded sequence of possibly dependent random variables in R^k. Let $z^t = E[\, w^t \mid w^{t-1}, w^{t-2}, \dots, w^1\,] - w^t$, and $\bar{z}^t$ the average of the z^t's. Then $\lim_{t\to\infty} \bar{z}^t = 0$ with probability one.[a]

[a] PY refers to Loève, 1978, Book II, Th. 32.E.1.

We have

$$z_x^t =_{Def} \left(\frac{k}{\epsilon} \cdot e_x^t \cdot u(x, y^t) \right) - u(x, y^t).$$

ϵ-Greedy regret matching (outline of proof)

Strong law of large numbers for dependent random variables. Let $\{w^t\}^t$ be a bounded sequence of possibly dependent random variables in R^k. Let $z^t = E[\, w^t \mid w^{t-1}, w^{t-2}, \dots, w^1 \,] - w^t$, and $\bar{z}^t$ the average of the z^t's. Then $\lim_{t\to\infty} \bar{z}^t = 0$ with probability one.[a]

[a] PY refers to Loève, 1978, Book II, Th. 32.E.1.

We have

$$z^t_x =_{Def} \left(\frac{k}{\epsilon} \cdot e^t_x \cdot u(x, y^t) \right) - u(x, y^t).$$

and

$$E[z^t_x] = 0.$$

ϵ-Greedy regret matching (outline of proof)

Strong law of large numbers for dependent random variables. Let $\{w^t\}^t$ be a bounded sequence of possibly dependent random variables in R^k. Let $z^t = E[\, w^t \mid w^{t-1}, w^{t-2}, \dots, w^1 \,] - w^t$, and $\bar{z}^t$ the average of the z^t's. Then $\lim_{t\to\infty} \bar{z}^t = 0$ with probability one.[a]

[a] PY refers to Loève, 1978, Book II, Th. 32.E.1.

We have

$$z_x^t =_{Def} \left(\frac{k}{\epsilon} \cdot e_x^t \cdot u(x, y^t) \right) - u(x, y^t).$$

and

$$E[z_x^t] = 0.$$

If

$$\bar{z}^t =_{Def} \text{average of } z^s,\ s \leq t$$

ϵ-Greedy regret matching (outline of proof)

Strong law of large numbers for dependent random variables. Let $\{w^t\}^t$ be a bounded sequence of possibly dependent random variables in R^k. Let $z^t = E[\, w^t \mid w^{t-1}, w^{t-2}, \dots, w^1 \,] - w^t$, and $\bar{z}^t$ the average of the z^t's. Then $\lim_{t\to\infty} \bar{z}^t = 0$ with probability one.[a]

[a] PY refers to Loève, 1978, Book II, Th. 32.E.1.

We have

$$z^t_x =_{Def} \left(\frac{k}{\epsilon} \cdot e^t_x \cdot u(x, y^t) \right) - u(x, y^t).$$

and

$$E[z^t_x] = 0.$$

If

$$\bar{z}^t =_{Def} \text{average of } z^s,\ s \leq t$$

then from the strong law of large numbers for dependent random variables it follows that $\lim_{t\to\infty} \bar{z}^t = 0$ a.s.

Estimated vs. true regret

Now write $\bar{z}_x^t$ as follows (!):

$$\bar{z}_x^t = \underbrace{\frac{1}{t}\sum_{s=1}^{t}\frac{k}{\epsilon}\cdot e_x^s\cdot u(x, y^s) - \bar{u}^t}_{\substack{\text{scaled}\\ \text{empirical regret}}} \;-\; \underbrace{\frac{1}{t}\sum_{s=1}^{t} u(x, y^s) - \bar{u}^t}_{\text{true regret}}$$

Estimated vs. true regret

Now write $\bar{z}_x^t$ as follows (!):

$$\bar{z}_x^t = \underbrace{\frac{1}{t}\sum_{s=1}^{t}\frac{k}{\epsilon}\cdot e_x^s\cdot u(x, y^s) - \bar{u}^t}_{\substack{\text{scaled}\\ \text{empirical regret}}} \; - \; \underbrace{\frac{1}{t}\sum_{s=1}^{t} u(x, y^s) - \bar{u}^t}_{\text{true regret}}$$

1. Since $\lim_{t\to\infty} \bar{z}^t = 0$, scaled empirical regret converges to true regret a.s.

Estimated vs. true regret

Now write $\bar{z}_x^t$ as follows (!):

$$\bar{z}_x^t = \underbrace{\frac{1}{t}\sum_{s=1}^{t}\frac{k}{\epsilon}\cdot e_x^s\cdot u(x,y^s) - \bar{u}^t}_{\substack{\text{scaled}\\ \text{empirical regret}}} - \underbrace{\frac{1}{t}\sum_{s=1}^{t} u(x,y^s) - \bar{u}^t}_{\text{true regret}}$$

1. Since $\lim_{t\to\infty} \bar{z}^t = 0$, scaled empirical regret converges to true regret a.s.

2. ϵ% of the time A explores.

Estimated vs. true regret

Now write $\bar{z}_x^t$ as follows (!):

$$\bar{z}_x^t = \underbrace{\frac{1}{t}\sum_{s=1}^{t}\frac{k}{\epsilon}\cdot e_x^s \cdot u(x,y^s) - \bar{u}^t}_{\substack{\text{scaled}\\ \text{empirical regret}}} - \underbrace{\frac{1}{t}\sum_{s=1}^{t} u(x,y^s) - \bar{u}^t}_{\text{true regret}}$$

1. Since $\lim_{t\to\infty} \bar{z}^t = 0$, scaled empirical regret converges to true regret a.s.
2. $\epsilon\%$ of the time A explores.
3. $(1-\epsilon)\%$ of the time A plays empirical regret

Estimated vs. true regret

Now write $\bar{z}_x^t$ as follows (!):

$$\bar{z}_x^t = \underbrace{\frac{1}{t}\sum_{s=1}^{t} \frac{k}{\epsilon} \cdot e_x^s \cdot u(x, y^s) - \bar{u}^t}_{\substack{\text{scaled} \\ \text{empirical regret}}} - \underbrace{\frac{1}{t}\sum_{s=1}^{t} u(x, y^s) - \bar{u}^t}_{\text{true regret}}$$

1. Since $\lim_{t\to\infty} \bar{z}^t = 0$, scaled empirical regret converges to true regret a.s.
2. ϵ% of the time A explores.
3. $(1-\epsilon)$% of the time A plays empirical regret $\rightsquigarrow$ true regret.

Estimated vs. true regret

Now write $\bar{z}^t_x$ as follows (!):

$$\bar{z}^t_x = \underbrace{\frac{1}{t}\sum_{s=1}^{t}\frac{k}{\epsilon}\cdot e^s_x \cdot u(x, y^s) - \bar{u}^t}_{\text{scaled empirical regret}} - \underbrace{\frac{1}{t}\sum_{s=1}^{t} u(x, y^s) - \bar{u}^t}_{\text{true regret}}$$

1. Since $\lim_{t\to\infty} \bar{z}^t = 0$, scaled empirical regret converges to true regret a.s.
2. ϵ% of the time A explores.
3. $(1-\epsilon)$% of the time A plays empirical regret $\rightsquigarrow$ true regret.
4. In the long run, empirical regret is within 2ϵ from true regret.

Estimated vs. true regret

Now write $\bar{z}_x^t$ as follows (!):

$$\bar{z}_x^t = \underbrace{\frac{1}{t}\sum_{s=1}^{t}\frac{k}{\epsilon}\cdot e_x^s \cdot u(x, y^s) - \bar{u}^t}_{\text{scaled empirical regret}} - \underbrace{\frac{1}{t}\sum_{s=1}^{t} u(x, y^s) - \bar{u}^t}_{\text{true regret}}$$

1. Since $\lim_{t\to\infty} \bar{z}^t = 0$, scaled empirical regret converges to true regret a.s.
2. ϵ% of the time A explores.
3. $(1-\epsilon)$% of the time A plays empirical regret $\rightsquigarrow$ true regret.
4. In the long run, empirical regret is within 2ϵ from true regret.
5. If ϵ is set to $\delta/2$, then empirical regret remains within $2 \cdot \delta/2$ from zero. $\square$

Literature

Literature

- Regret matching can be traced to Blackwell's approachability theorem and Hannan's notion of universal consistency.

Literature

- Regret matching can be traced to Blackwell's approachability theorem and Hannan's notion of universal consistency.
- The diagram on regret matching is taken from Peyton Young, and Foster and Vohra (who formulate the problem from a decision theoretic point of view).

Literature

- Regret matching can be traced to Blackwell's approachability theorem and Hannan's notion of universal consistency.
- The diagram on regret matching is taken from Peyton Young, and Foster and Vohra (who formulate the problem from a decision theoretic point of view).
- The regret-matching algorithm and the analysis of its convergence to coarse correlated equilibria (a generalisation of Nash equilibria) is given by Hart and Mas-Colell.

Literature

- Regret matching can be traced to Blackwell's approachability theorem and Hannan's notion of universal consistency.
- The diagram on regret matching is taken from Peyton Young, and Foster and Vohra (who formulate the problem from a decision theoretic point of view).
- The regret-matching algorithm and the analysis of its convergence to coarse correlated equilibria (a generalisation of Nash equilibria) is given by Hart and Mas-Colell.

Blackwell, D. (1956). "Controlled random walks". *Proc. of the Int. Congress of Mathematicians*, North-Holland Publishing Comp., pp. 336-338.

Literature

- Regret matching can be traced to Blackwell's approachability theorem and Hannan's notion of universal consistency.
- The diagram on regret matching is taken from Peyton Young, and Foster and Vohra (who formulate the problem from a decision theoretic point of view).
- The regret-matching algorithm and the analysis of its convergence to coarse correlated equilibria (a generalisation of Nash equilibria) is given by Hart and Mas-Colell.

Blackwell, D. (1956). "Controlled random walks". *Proc. of the Int. Congress of Mathematicians*, North-Holland Publishing Comp., pp. 336-338.

Hannan, J. F. (1957). "Approximation to Bayes risk in repeated plays". *Contributions to the Theory of Games*, **3**, pp. 97-139.

Literature

- Regret matching can be traced to Blackwell's approachability theorem and Hannan's notion of universal consistency.
- The diagram on regret matching is taken from Peyton Young, and Foster and Vohra (who formulate the problem from a decision theoretic point of view).
- The regret-matching algorithm and the analysis of its convergence to coarse correlated equilibria (a generalisation of Nash equilibria) is given by Hart and Mas-Colell.

Blackwell, D. (1956). "Controlled random walks". *Proc. of the Int. Congress of Mathematicians*, North-Holland Publishing Comp., pp. 336-338.

Hannan, J. F. (1957). "Approximation to Bayes risk in repeated plays". *Contributions to the Theory of Games*, **3**, pp. 97-139.

Hart, S., and Mas-Colell, A. (2000). "A simple adaptive procedure leading to correlated equilibrium". *Econometrica*, **68**, pp. 1127-1150.

Literature

- Regret matching can be traced to Blackwell's approachability theorem and Hannan's notion of universal consistency.
- The diagram on regret matching is taken from Peyton Young, and Foster and Vohra (who formulate the problem from a decision theoretic point of view).
- The regret-matching algorithm and the analysis of its convergence to coarse correlated equilibria (a generalisation of Nash equilibria) is given by Hart and Mas-Colell.

Blackwell, D. (1956). "Controlled random walks". *Proc. of the Int. Congress of Mathematicians*, North-Holland Publishing Comp., pp. 336-338.

Hannan, J. F. (1957). "Approximation to Bayes risk in repeated plays". *Contributions to the Theory of Games*, **3**, pp. 97-139.

Hart, S., and Mas-Colell, A. (2000). "A simple adaptive procedure leading to correlated equilibrium". *Econometrica*, **68**, pp. 1127-1150.

Foster, D., and Vohra, R. (1999). "Regret in the on-line decision problem". GEB: *Games and Economic Behavior*, **29**, pp. 7-36.

What next?

What next?

- **Fictitious Play**. Monitor actions of opponent(s) and play a best response to most frequent actions. As opposed to no-regret, fictitious play is interested in the opponent's behaviour to predict future play.

What next?

- **Fictitious Play**. Monitor actions of opponent(s) and play a best response to most frequent actions. As opposed to no-regret, fictitious play is interested in the opponent's behaviour to predict future play.
- **Smoothed fictitious play**. With fictitious play, the probability to play sub-optimal responses is zero. Smoothed fictitious play plays sub-optimal responses proportional to their expected payoff, given opponents' play.

What next?

- **Fictitious Play**. Monitor actions of opponent(s) and play a best response to most frequent actions. As opposed to no-regret, fictitious play is interested in the opponent's behaviour to predict future play.
- **Smoothed fictitious play**. With fictitious play, the probability to play sub-optimal responses is zero. Smoothed fictitious play plays sub-optimal responses proportional to their expected payoff, given opponents' play.
- **Conditional no-regret**. Conditions on particular actions. There is regret if there is a pair of actions (x, x') such that, with hindsight, playing x' was better than playing x.

What next?

- **Fictitious Play**. Monitor actions of opponent(s) and play a best response to most frequent actions. As opposed to no-regret, fictitious play is interested in the opponent's behaviour to predict future play.
- **Smoothed fictitious play**. With fictitious play, the probability to play sub-optimal responses is zero. Smoothed fictitious play plays sub-optimal responses proportional to their expected payoff, given opponents' play.
- **Conditional no-regret**. Conditions on particular actions. There is regret if there is a pair of actions (x, x') such that, with hindsight, playing x' was better than playing x.
- **Satisficing Play**. While payoffs equal or supersede the average of past payoffs, keep playing the same action.

Exam problem

Regret matching

Regret matching

Problem.

Regret matching

Problem. The following game is played with regret matching with initial action profile (T, L).

	L	R
T	$1, 4$	$3, 1$
B	$2, 0$	$0, 5$

.

Regret matching

Problem. The following game is played with regret matching with initial action profile (T, L).

	L	R
T	$1, 4$	$3, 1$
B	$2, 0$	$0, 5$

.

Give the action and the cumulative regrets of the column player at the 5th round.

Regret matching

Problem. The following game is played with regret matching with initial action profile (T, L).

	L	R
T	$1, 4$	$3, 1$
B	$2, 0$	$0, 5$

.

Give the action and the cumulative regrets of the column player at the 5th round.

Answer.

Regret matching

Problem. The following game is played with regret matching with initial action profile (T, L).

	L	R
T	$1,4$	$3,1$
B	$2,0$	$0,5$

.

Give the action and the cumulative regrets of the column player at the 5th round.

Answer. Emulate regret matching for 5 rounds.

Regret matching

Problem. The following game is played with regret matching with initial action profile (T, L).

	L	R
T	1, 4	3, 1
B	2, 0	0, 5

Give the action and the cumulative regrets of the column player at the 5th round.

Answer. Emulate regret matching for 5 rounds.

t :	1	2	3	4	5
a_{row} :	T	B	B	T	T
a_{col} :	L	L	R	R	**R**
x_{row} :	1	2	0	3	3
x_{col} :	4	0	5	1	1
$\Delta r(T)$:	0	-1	3	0	0
$r(T)$:	0	-1	2	2	2
$\Delta r(B)$:	1	0	0	-3	-3
$r(B)$:	1	1	1	-2	-5
$\Delta r(L)$:	0	0	-5	3	3
$r(L)$:	0	0	-5	-2	**1**
$\Delta r(R)$:	-3	5	0	0	0
$r(R)$:	-3	2	2	2	**2**

Regret matching

Problem. The following game is played with regret matching with initial action profile (T, L).

	L	R
T	1,4	3,1
B	2,0	0,5

.

Give the action and the cumulative regrets of the column player at the 5th round.

Answer. Emulate regret matching for 5 rounds.

t :	1	2	3	4	5
a_{row} :	T	B	B	T	T
a_{col} :	L	L	R	R	$\mathbf{R}$
x_{row} :	1	2	0	3	3
x_{col} :	4	0	5	1	1
$\Delta r(T)$:	0	-1	3	0	0
$r(T)$:	0	-1	2	2	2
$\Delta r(B)$:	1	0	0	-3	-3
$r(B)$:	1	1	1	-2	-5
$\Delta r(L)$:	0	0	-5	3	3
$r(L)$:	0	0	-5	-2	$\mathbf{1}$
$\Delta r(R)$:	-3	5	0	0	0
$r(R)$:	-3	2	2	2	$\mathbf{2}$

The answer is R, $(2, 1)$. $\square$

Exam problem

ϵ-Greedy regret matching

ϵ-Greedy regret matching

Problem.

ϵ-Greedy regret matching

Problem. What is the benefit of ϵ-Greedy regret matching

ϵ-Greedy regret matching

Problem. What is the benefit of ϵ-Greedy regret matching, compared to ordinary regret matching?

ϵ-Greedy regret matching

Problem. What is the benefit of ϵ-Greedy regret matching, compared to ordinary regret matching?

a) Opponent behaviour is irrelevant.

ϵ-Greedy regret matching

Problem. What is the benefit of ϵ-Greedy regret matching, compared to ordinary regret matching?

a) Opponent behaviour is irrelevant.

b) Regret matching is done off-policy.

ϵ-Greedy regret matching

Problem. What is the benefit of ϵ-Greedy regret matching, compared to ordinary regret matching?

a) Opponent behaviour is irrelevant.

b) Regret matching is done off-policy.

c) Regret matching is done only ϵ% of the time, not all of the time.

ϵ-Greedy regret matching

Problem. What is the benefit of ϵ-Greedy regret matching, compared to ordinary regret matching?

a) Opponent behaviour is irrelevant.

b) Regret matching is done off-policy.

c) Regret matching is done only ϵ% of the time, not all of the time.

d) For all $\delta > 0$ there exists an $\epsilon > 0$ such that ϵ-Greedy regret matching yields at most δ regret.

ϵ-Greedy regret matching

Problem. What is the benefit of ϵ-Greedy regret matching, compared to ordinary regret matching?

a) Opponent behaviour is irrelevant.

b) Regret matching is done off-policy.

c) Regret matching is done only ϵ% of the time, not all of the time.

d) For all $\delta > 0$ there exists an $\epsilon > 0$ such that ϵ-Greedy regret matching yields at most δ regret.

Answer.

ϵ-Greedy regret matching

Problem. What is the benefit of ϵ-Greedy regret matching, compared to ordinary regret matching?

a) Opponent behaviour is irrelevant.

b) Regret matching is done off-policy.

c) Regret matching is done only ϵ% of the time, not all of the time.

d) For all $\delta > 0$ there exists an $\epsilon > 0$ such that ϵ-Greedy regret matching yields at most δ regret.

Answer. Option a) is correct

ϵ-Greedy regret matching

Problem. What is the benefit of ϵ-Greedy regret matching, compared to ordinary regret matching?

a) Opponent behaviour is irrelevant.

b) Regret matching is done off-policy.

c) Regret matching is done only $\epsilon\%$ of the time, not all of the time.

d) For all $\delta > 0$ there exists an $\epsilon > 0$ such that ϵ-Greedy regret matching yields at most δ regret.

Answer. Option a) is correct: knowledge of opponent's actions is not needed, because regrets are computed when experimenting.

ϵ-Greedy regret matching

Problem. What is the benefit of ϵ-Greedy regret matching, compared to ordinary regret matching?

a) Opponent behaviour is irrelevant.

b) Regret matching is done off-policy.

c) Regret matching is done only ϵ% of the time, not all of the time.

d) For all $\delta > 0$ there exists an $\epsilon > 0$ such that ϵ-Greedy regret matching yields at most δ regret.

Answer. Option a) is correct: knowledge of opponent's actions is not needed, because regrets are computed when experimenting. Option b) is wrong: regret matching *is* the policy

ϵ-Greedy regret matching

Problem. What is the benefit of ϵ-Greedy regret matching, compared to ordinary regret matching?

a) Opponent behaviour is irrelevant.

b) Regret matching is done off-policy.

c) Regret matching is done only ϵ% of the time, not all of the time.

d) For all $\delta > 0$ there exists an $\epsilon > 0$ such that ϵ-Greedy regret matching yields at most δ regret.

Answer. Option a) is correct: knowledge of opponent's actions is not needed, because regrets are computed when experimenting. Option b) is wrong: regret matching *is* the policy; experimenting is done off-policy.

ϵ-Greedy regret matching

Problem. What is the benefit of ϵ-Greedy regret matching, compared to ordinary regret matching?

a) Opponent behaviour is irrelevant.

b) Regret matching is done off-policy.

c) Regret matching is done only ϵ% of the time, not all of the time.

d) For all $\delta > 0$ there exists an $\epsilon > 0$ such that ϵ-Greedy regret matching yields at most δ regret.

Answer. Option a) is correct: knowledge of opponent's actions is not needed, because regrets are computed when experimenting. Option b) is wrong: regret matching *is* the policy; experimenting is done off-policy. Option c) is wrong because with ϵ-Greedy regret matching, regret matching is done most of the time, based on occasional experiments.

ϵ-Greedy regret matching

Problem. What is the benefit of ϵ-Greedy regret matching, compared to ordinary regret matching?

a) Opponent behaviour is irrelevant.

b) Regret matching is done off-policy.

c) Regret matching is done only $\epsilon\%$ of the time, not all of the time.

d) For all $\delta > 0$ there exists an $\epsilon > 0$ such that ϵ-Greedy regret matching yields at most δ regret.

Answer. Option a) is correct: knowledge of opponent's actions is not needed, because regrets are computed when experimenting. Option b) is wrong: regret matching *is* the policy; experimenting is done off-policy. Option c) is wrong because with ϵ-Greedy regret matching, regret matching is done most of the time, based on occasional experiments. Option d) is not wrong, it merely quotes a property of ϵ-Greedy regret matching.

ϵ-Greedy regret matching

Problem. What is the benefit of ϵ-Greedy regret matching, compared to ordinary regret matching?

a) Opponent behaviour is irrelevant.

b) Regret matching is done off-policy.

c) Regret matching is done only ϵ% of the time, not all of the time.

d) For all $\delta > 0$ there exists an $\epsilon > 0$ such that ϵ-Greedy regret matching yields at most δ regret.

Answer. Option a) is correct: knowledge of opponent's actions is not needed, because regrets are computed when experimenting. Option b) is wrong: regret matching *is* the policy; experimenting is done off-policy. Option c) is wrong because with ϵ-Greedy regret matching, regret matching is done most of the time, based on occasional experiments. Option d) is not wrong, it merely quotes a property of ϵ-Greedy regret matching. See also Peyton Young, CH2, p. 23.

ϵ-Greedy regret matching

Problem. What is the benefit of ϵ-Greedy regret matching, compared to ordinary regret matching?

a) Opponent behaviour is irrelevant.

b) Regret matching is done off-policy.

c) Regret matching is done only $\epsilon\%$ of the time, not all of the time.

d) For all $\delta > 0$ there exists an $\epsilon > 0$ such that ϵ-Greedy regret matching yields at most δ regret.

Answer. Option a) is correct: knowledge of opponent's actions is not needed, because regrets are computed when experimenting. Option b) is wrong: regret matching *is* the policy; experimenting is done off-policy. Option c) is wrong because with ϵ-Greedy regret matching, regret matching is done most of the time, based on occasional experiments. Option d) is not wrong, it merely quotes a property of ϵ-Greedy regret matching. See also Peyton Young, CH2, p. 23. $\square$