Low-Regret for Online Decision-Making

Siddhartha Banerjee and Alberto Vera

Cornell University

November 6, 2018

Motivation

Motivation

OFFLINE is a host who knew all the arrivals ahead of time

Example: Online Packing (NRM)

- Resources with initial budgets B_i, $i = 1, \ldots, d$.

Example: Online Packing (NRM)

- Resources with initial budgets B_i, $i = 1, \ldots, d$.
- Type j wants a_{ij} units of resource i.

Example: Online Packing (NRM)

- Resources with initial budgets B_i, $i = 1, \ldots, d$.
- Type j wants a_{ij} units of resource i.
- Type j arrives with probability p_j and has reward r_j, $j = 1, \ldots, n$.

Example: Online Packing (NRM)

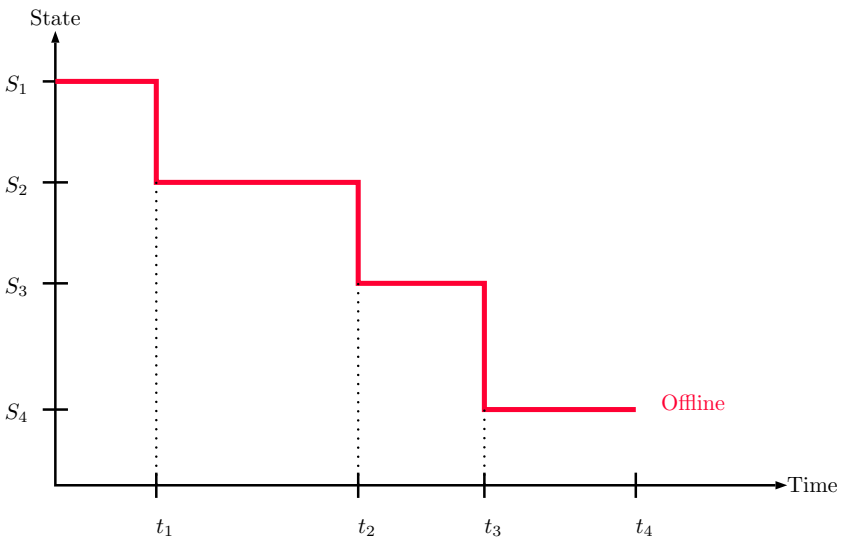
- Resources with initial budgets B_i, $i = 1, \ldots, d$.
- Type j wants a_{ij} units of resource i.
- Type j arrives with probability p_j and has reward r_j, $j = 1, \ldots, n$.

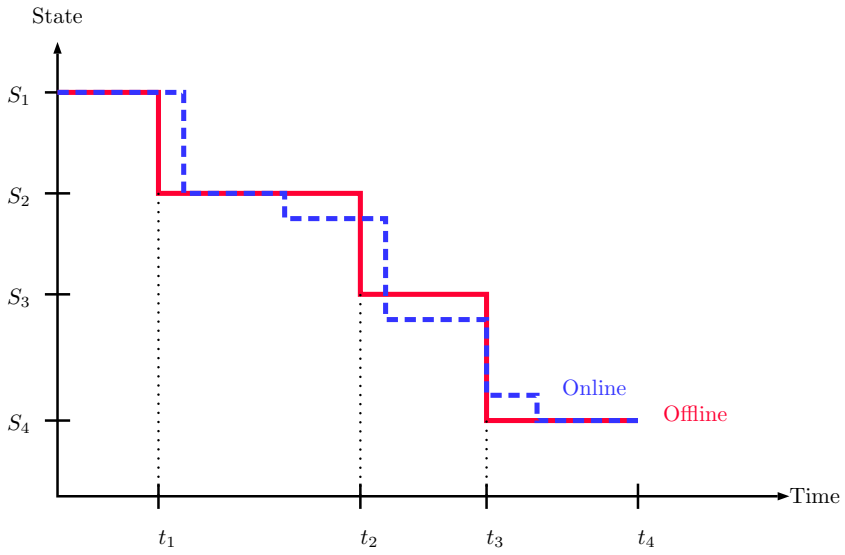
If we had full information, solve

$$\begin{array}{lrcll} \max & r'x & & & \\ \text{s.t.} & Ax & \leq & B & \text{(Budget Consumption)} \\ & x & \leq & Z(T) & \text{(Respect Arrivals)} \\ & x & \in & \mathbb{N}^n. & \end{array}$$

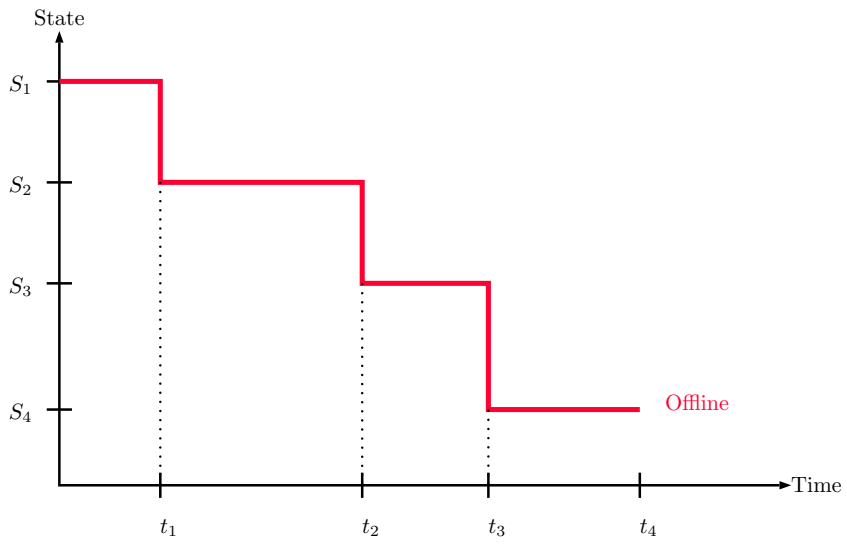
Paradigm: two agents solving this problem, ONLINE and OFFLINE

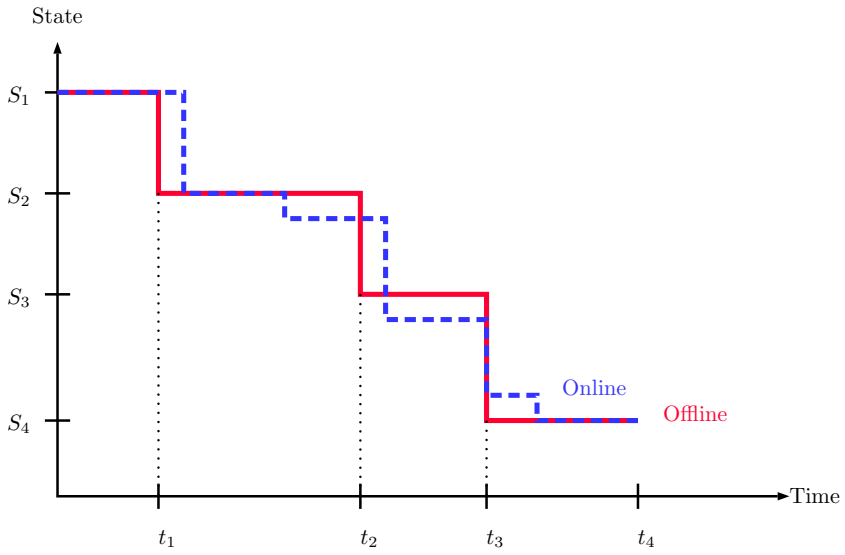
Traditional Approach and Our Contributions





Traditional Approach and Our Contributions





Our Contributions

- Compensated Coupling: for analyzing algorithms
- Bayes Selector: meta-algorithm based on estimators
- Applications: constant regret

Application: Constant Regret for Online Packing

V^{on} and V^{off} collected values, Regret $:= V^{\text{off}} - V^{\text{on}}$.

Application: Constant Regret for Online Packing

V^{on} and V^{off} collected values, Regret $:= V^{\text{off}} - V^{\text{on}}$.
$V^{\text{off}} = \max\{r'x : Ax \leq B, 0 \leq x \leq Z(T)\}$.

Application: Constant Regret for Online Packing

V^{on} and V^{off} collected values, Regret $:= V^{\text{off}} - V^{\text{on}}$.
$V^{\text{off}} = \max\{r'x : Ax \leq B, 0 \leq x \leq Z(T)\}$.

Theorem (Informal)

We provide a natural policy with constant expected regret for online packing problems.

Application: Constant Regret for Online Packing

V^{on} and V^{off} collected values, Regret $:= V^{\text{off}} - V^{\text{on}}$.
$V^{\text{off}} = \max\{r'x : Ax \leq B, 0 \leq x \leq Z(T)\}$.

Theorem (Informal)

We provide a natural policy with constant expected regret for online packing problems.

- Regret depends on $d, n, \kappa(A)$, and the distribution of $Z(t)$
- Independent of T and initial budget levels B
- Correlated and time-varying distributions
- $1 - o(1)$ approximation whenever T and B are large

Related Work

- Constant Regret for Multi-Secretary Arlotto and Gurvich (2017)
- Network Revenue Management / Online Packing:
  - Asymptotic analysis $O(\sqrt{T})$ regret Talluri and Van Ryzin (2004)
  - Later $o(\sqrt{T})$, $O(1)$ conditional, $O(1)$ Poisson Reiman and Wang (2008),Jasin and Kumar (2012),Bumpensanti and Wang (2018)
- Prophet: worst case distribution (competitive ratio)
  - Ratio $1 - e^{-1}$ for maximum of i.i.d. Hill and Kertz (1982)
  - Best ratio ≈ 0.745, used in posted pricing Correa et al. (2017)
- Online Matching Manshadi et al (2012)
- Approximate Dynamic Programming Powell (2011)

Framework for Online Decision-Making

- Arrival type $J^t \in \mathcal{J}$ presented at $t = T, T-1, \ldots, 1$

Framework for Online Decision-Making

- Arrival type $J^t \in \mathcal{J}$ presented at $t = T, T-1, \ldots, 1$
- States $s \in \mathcal{S}$ and actions $a \in \mathcal{A}$

Framework for Online Decision-Making

- Arrival type $J^t \in \mathcal{J}$ presented at $t = T, T-1, \ldots, 1$
- States $s \in \mathcal{S}$ and actions $a \in \mathcal{A}$
- Rewards $\mathcal{R}(a, s, j)$

Framework for Online Decision-Making

- Arrival type $J^t \in \mathcal{J}$ presented at $t = T, T-1, \ldots, 1$
- States $s \in \mathcal{S}$ and actions $a \in \mathcal{A}$
- Rewards $\mathcal{R}(a, s, j)$
- Transitions $\mathcal{T}(a, s, j)$

Framework for Online Decision-Making

- Arrival type $J^t \in \mathcal{J}$ presented at $t = T, T-1, \ldots, 1$
- States $s \in \mathcal{S}$ and actions $a \in \mathcal{A}$
- Rewards $\mathcal{R}(a, s, j)$
- Transitions $\mathcal{T}(a, s, j)$

Examples: Online Packing (NRM), Multi-Secretary, Online Matching, Ski-Rental, etc.

Example: Multi-Secretary

- One resource ($d = 1$) with capacity B.
- Goal: choose B arrivals to maximize their sum

6	10	10	8	7	8
$t = 6$	5	4	3	2	1

t is time to go

Value of Offline

$V^{\text{off}}(t, s)[\omega]$ is the optimal offline value starting on s with t periods to go on sample path ω.

Value of Offline

$V^{\text{off}}(t, s)[\omega]$ is the optimal offline value starting on s with t periods to go on sample path ω.

6 10 10 8 7 8

t	6	5	4	3	2	1
$B = 1$	10	10	10	8	8	8
$B = 2$	10+10	10+10	10+8	8+8	8+7	8

Value of Offline

$V^{\text{off}}(t, s)[\omega]$ is the optimal offline value starting on s with t periods to go on sample path ω.

6 10 10 8 7 8

t	6	5	4	3	2	1
$B = 1$	10	10	10	8	8	8
$B = 2$	10+10	10+10	10+8	8+8	8+7	8

It satisfies the Bellman Equation

$$V^{\text{off}}(t, s)[\omega] = \max\{\mathcal{R}(a, s, J^t) + V^{\text{off}}(t-1, \mathcal{T}(a, s, J^t))[\omega] : a \in \mathcal{A}\}.$$

Offline Follows Online

ONLINE uses a policy π^{on}. We want OFFLINE to follow π^{on}

Offline Follows Online

ONLINE uses a policy π^{on}. We want OFFLINE to follow π^{on}
What if OFFLINE is not satisfied with $a = \pi^{\text{on}}(t, s, j)$?

$$V^{\text{off}}(t, s)[\omega] > \mathcal{R}(a, s, J^t) + V^{\text{off}}(t-1, \mathcal{T}(a, s, J^t))[\omega].$$

Offline Follows Online

ONLINE uses a policy π^{on}. We want OFFLINE to follow π^{on}
What if OFFLINE is not satisfied with $a = \pi^{\text{on}}(t, s, j)$?

$$V^{\text{off}}(t, s)[\omega] > \mathcal{R}(a, s, J^t) + V^{\text{off}}(t-1, \mathcal{T}(a, s, J^t))[\omega].$$

We need to compensate him by paying

$$V^{\text{off}}(t, s)[\omega] - \mathcal{R}(a, s, J^t) - V^{\text{off}}(t-1, \mathcal{T}(a, s, J^t))[\omega].$$

Offline Follows Online

ONLINE uses a policy π^{on}. We want OFFLINE to follow π^{on}
What if OFFLINE is not satisfied with $a = \pi^{\text{on}}(t, s, j)$?

$$V^{\text{off}}(t, s)[\omega] > \mathcal{R}(a, s, J^t) + V^{\text{off}}(t-1, \mathcal{T}(a, s, J^t))[\omega].$$

We need to compensate him by paying

$$V^{\text{off}}(t, s)[\omega] - \mathcal{R}(a, s, J^t) - V^{\text{off}}(t-1, \mathcal{T}(a, s, J^t))[\omega].$$

6	10	10	8	7	8
$t = 6$	5	4	3	2	1

Compensations

How big are the compensations?

Compensations

How big are the compensations?

$r_{\max}$ for multi-secretary and $|A_j| r_{\max} \leq d r_{\max}$ for packing.

Compensations

How big are the compensations?
$r_{\max}$ for multi-secretary and $|A_j| r_{\max} \leq d r_{\max}$ for packing.

$Q(t, s, a)$ is the **disagreement set**: $\omega \in \Omega$ such that OFFLINE loses optimality by taking a.

Compensations

How big are the compensations?
$r_{\max}$ for multi-secretary and $|A_j| r_{\max} \leq d r_{\max}$ for packing.

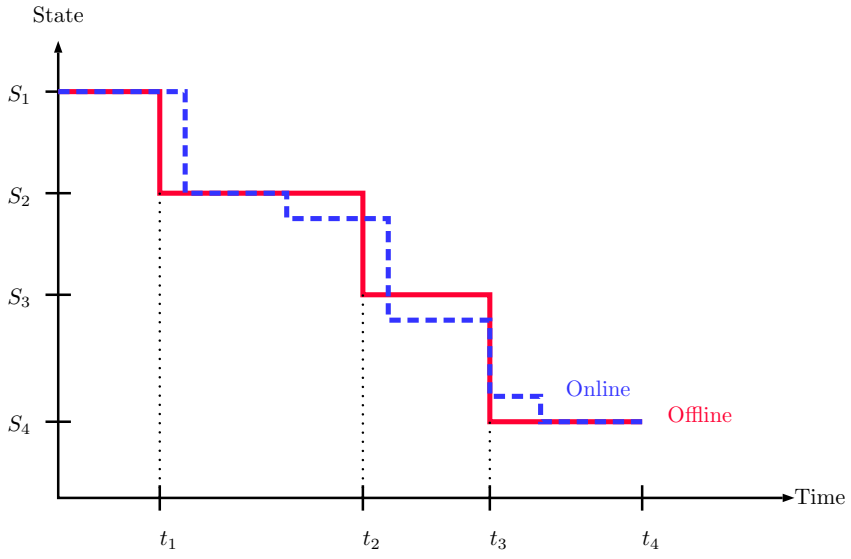
$Q(t, s, a)$ is the **disagreement set**: $\omega \in \Omega$ such that OFFLINE loses optimality by taking a.

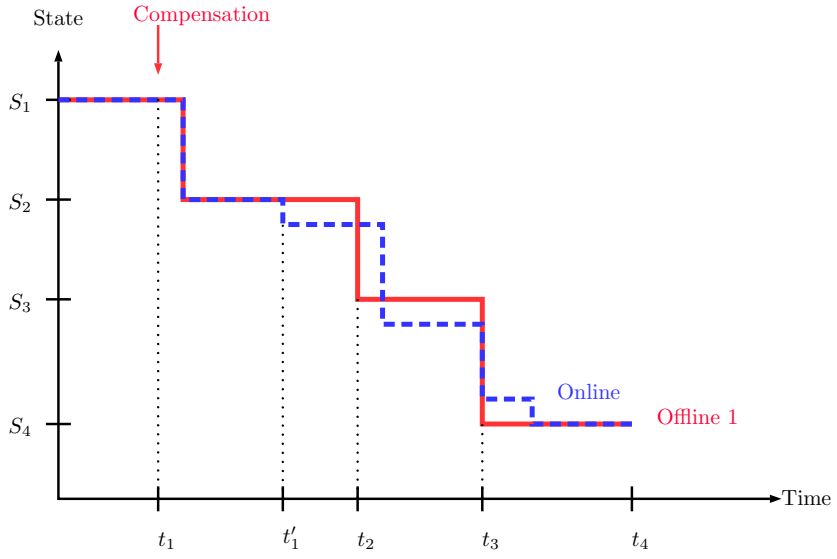
Lemma (Compensated Coupling)

Fix any policy for ONLINE. *Let* S^t *be* ONLINE*'s state over time and* A^t ONLINE*'s action over time. If* $\bar{r}$ *is a bound on the compensations, then*

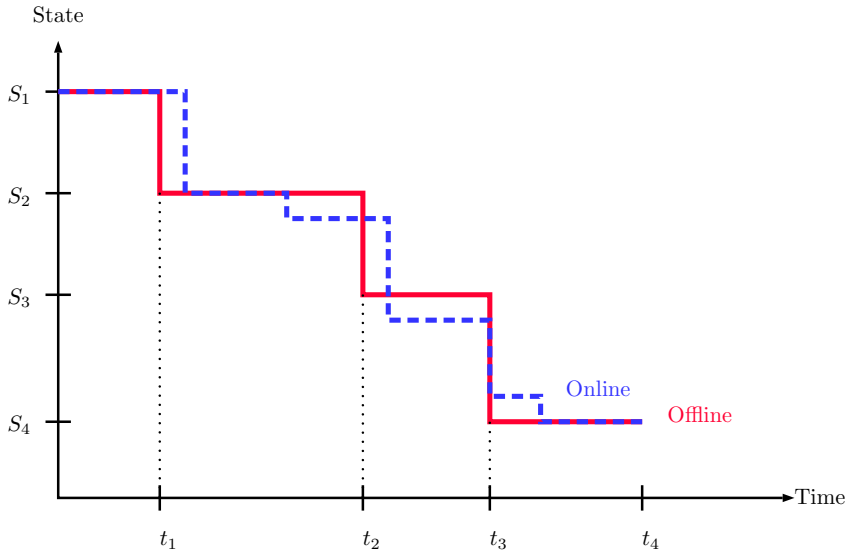
$$\mathbb{E}[Regret] \leq \bar{r}\mathbb{E}\left[\sum_t \mathbb{P}[Q(t, S^t, A^t)]\right]$$

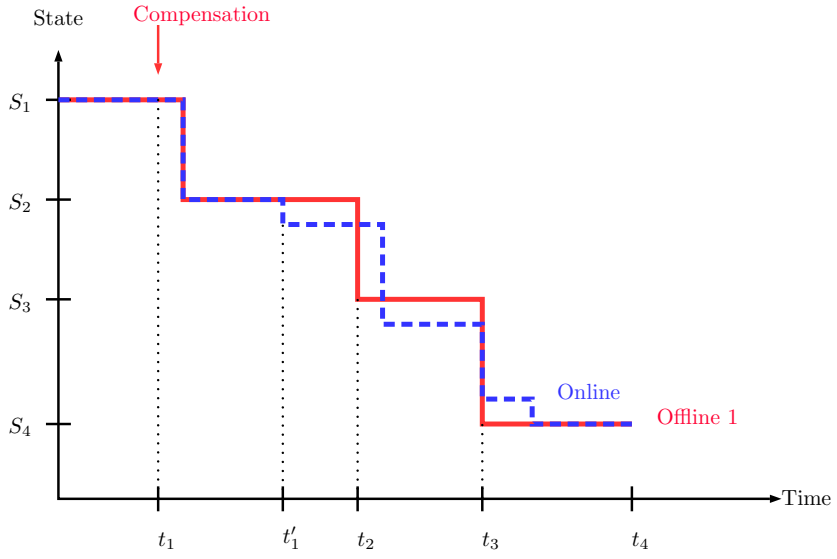
Compensated Coupling View

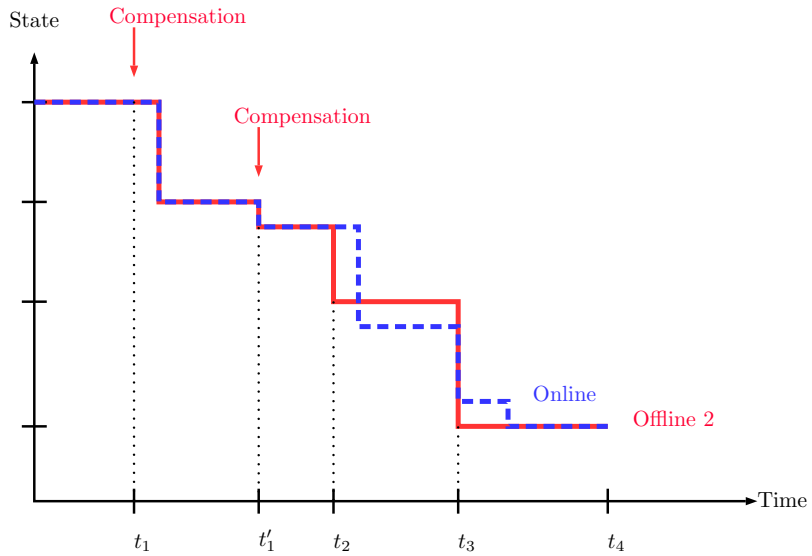




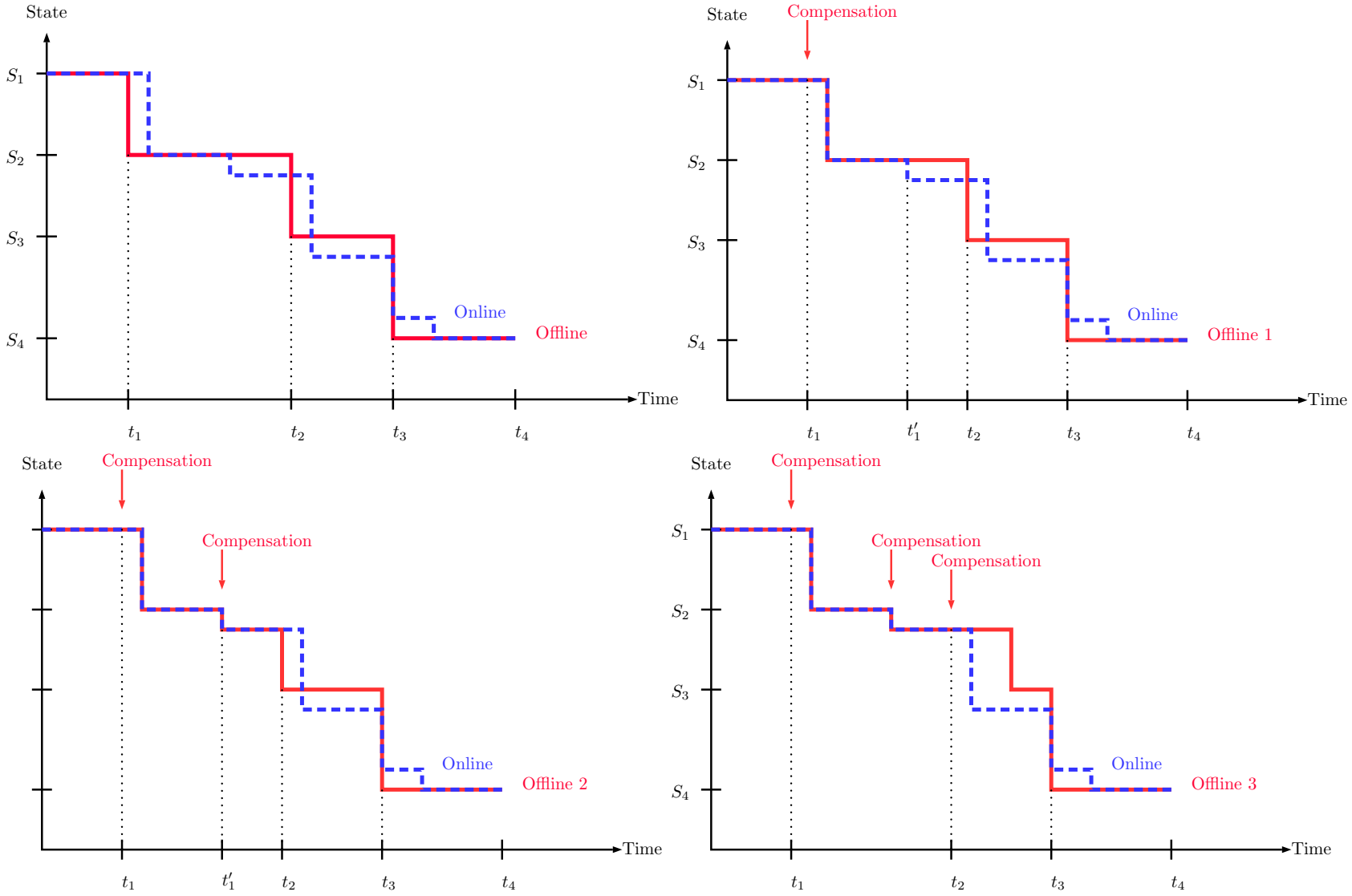
Compensated Coupling View







Compensated Coupling View



The Bayes Selector Policy

Compensated coupling works for any policy. Can we turn it around and find a policy?

The Bayes Selector Policy

Compensated coupling works for any policy. Can we turn it around and find a policy?

Alice is making decisions and we want to copy her.
An oracle tells us she picked red/blue with probabilities 0.7/0.3.
What to do?

The Bayes Selector Policy

Compensated coupling works for any policy. Can we turn it around and find a policy?

Alice is making decisions and we want to copy her.
An oracle tells us she picked red/blue with probabilities 0.7/0.3.
What to do?

$Q(t, s, a)$ is the **disagreement set**: $\omega \in \Omega$ such that OFFLINE loses optimality by taking a.

The Bayes Selector Policy

Compensated coupling works for any policy. Can we turn it around and find a policy?

Alice is making decisions and we want to copy her.
An oracle tells us she picked red/blue with probabilities 0.7/0.3.
What to do?

$Q(t, s, a)$ is the **disagreement set**: $\omega \in \Omega$ such that OFFLINE loses optimality by taking a.

Policy: at each time t, take action

$$A^t_{\mathrm{BS}} \in \operatorname{argmin}\{\mathbb{P}[Q(t, S^t_{\mathrm{BS}}, a)] : a \in \mathcal{A}\}.$$

Regret Guarantee of Bayes Selector

Ideal Policy: at time t, $A^t_{\mathrm{BS}} \in \operatorname{argmin}\{\mathbb{P}[Q(t, S^t_{\mathrm{BS}}, a)] : a \in \mathcal{A}\}$.

Regret Guarantee of Bayes Selector

Ideal Policy: at time t, $A^t_{\mathrm{BS}} \in \operatorname{argmin}\{\mathbb{P}[Q(t, S^t_{\mathrm{BS}}, a)] : a \in \mathcal{A}\}$.

Meta-Algorithm: Based on estimators $\hat{q}(t, s, a) \geq \mathbb{P}[Q(t, s, a)]$

Regret Guarantee of Bayes Selector

Ideal Policy: at time t, $A^t_{\mathrm{BS}} \in \operatorname{argmin}\{\mathbb{P}[Q(t, S^t_{\mathrm{BS}}, a)] : a \in \mathcal{A}\}$.

Meta-Algorithm: Based on estimators $\hat{q}(t, s, a) \geq \mathbb{P}[Q(t, s, a)]$ at time t, $A^t_{\mathrm{BS}} \in \operatorname{argmin}\{\hat{q}(t, S^t_{\mathrm{BS}}, a) : a \in \mathcal{A}\}$.

Regret Guarantee of Bayes Selector

Ideal Policy: at time t, $A^t_{\text{BS}} \in \text{argmin}\{\mathbb{P}[Q(t, S^t_{\text{BS}}, a)] : a \in \mathcal{A}\}$.

Meta-Algorithm: Based on estimators $\hat{q}(t, s, a) \geq \mathbb{P}[Q(t, s, a)]$ at time t, $A^t_{\text{BS}} \in \text{argmin}\{\hat{q}(t, S^t_{\text{BS}}, a) : a \in \mathcal{A}\}$.

Theorem

Let $\bar{r}$ be an upper bound on the compensations. The Bayes Selector achieves

$$\mathbb{E}[Regret] \leq \bar{r}\mathbb{E}\left[\sum_t \hat{q}(t, S^t_{BS}, A^t_{BS})\right]$$

Regret Guarantee of Bayes Selector

Ideal Policy: at time t, $A^t_{\text{BS}} \in \text{argmin}\{\mathbb{P}[Q(t, S^t_{\text{BS}}, a)] : a \in \mathcal{A}\}$.

Meta-Algorithm: Based on estimators $\hat{q}(t, s, a) \geq \mathbb{P}[Q(t, s, a)]$ at time t, $A^t_{\text{BS}} \in \text{argmin}\{\hat{q}(t, S^t_{\text{BS}}, a) : a \in \mathcal{A}\}$.

Theorem

Let $\bar{r}$ be an upper bound on the compensations. The Bayes Selector achieves

$$\mathbb{E}[Regret] \leq \bar{r}\mathbb{E}\left[\sum_t \hat{q}(t, S^t_{BS}, A^t_{BS})\right]$$

Theorem

The Bayes Selector achieves constant regret for online packing and matching problems.

Estimators for Online Packing

$\hat{q}(t, s, a) \geq \mathbb{P}[Q(t, s, a)]$, with $Q(t, s, a)$ is the disagreement set.

- Given ONLINE's budget B^t, OFFLINE solves $\max\{r'x : Ax \leq B^t, 0 \leq x \leq Z(t)\}$.

Estimators for Online Packing

$\hat{q}(t, s, a) \geq \mathbb{P}[Q(t, s, a)]$, with $Q(t, s, a)$ is the disagreement set.

- Given ONLINE's budget B^t, OFFLINE solves $\max\{r'x : Ax \leq B^t, 0 \leq x \leq Z(t)\}$.
- Let X solve $\max\{r'x : Ax \leq B^t, 0 \leq x \leq \mathbb{E}[Z(t)]\}$

Estimators for Online Packing

$\hat{q}(t, s, a) \geq \mathbb{P}[Q(t, s, a)]$, with $Q(t, s, a)$ is the disagreement set.

- Given ONLINE's budget B^t, OFFLINE solves $\max\{r'x : Ax \leq B^t, 0 \leq x \leq Z(t)\}$.
- Let X solve $\max\{r'x : Ax \leq B^t, 0 \leq x \leq \mathbb{E}[Z(t)]\}$
- $X_j/\mathbb{E}[Z_j(t)] \in [0, 1]$ for all j.

Estimators for Online Packing

$\hat{q}(t, s, a) \geq \mathbb{P}[Q(t, s, a)]$, with $Q(t, s, a)$ is the disagreement set.

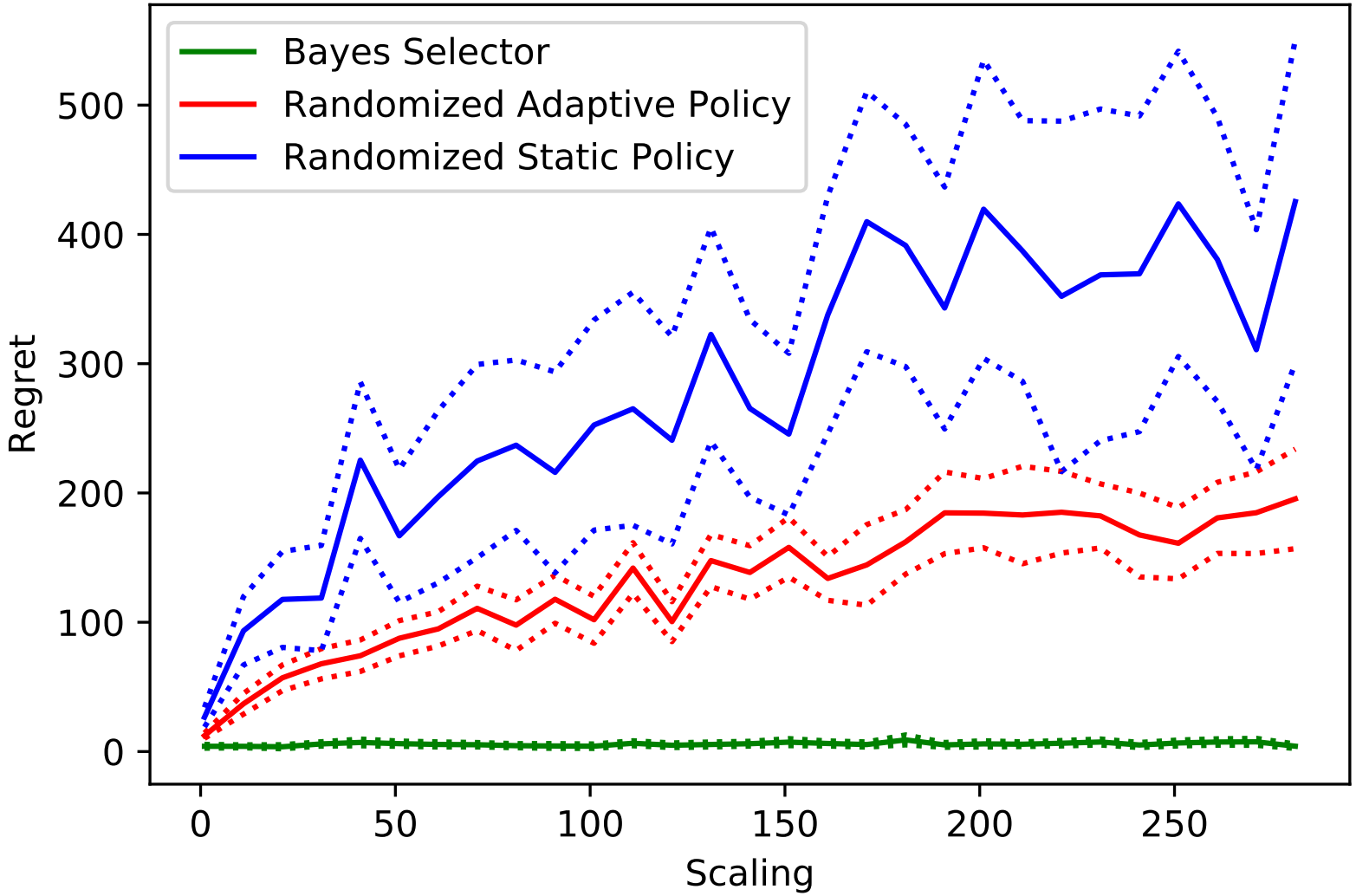
- Given ONLINE's budget B^t, OFFLINE solves $\max\{r'x : Ax \leq B^t, 0 \leq x \leq Z(t)\}$.
- Let X solve $\max\{r'x : Ax \leq B^t, 0 \leq x \leq \mathbb{E}[Z(t)]\}$
- $X_j/\mathbb{E}[Z_j(t)] \in [0, 1]$ for all j.
- Accept j iff $X_j/\mathbb{E}[Z_j(t)] \geq 1/2$.

Estimators for Online Packing

$\hat{q}(t, s, a) \geq \mathbb{P}[Q(t, s, a)]$, with $Q(t, s, a)$ is the disagreement set.

- Given ONLINE's budget B^t, OFFLINE solves $\max\{r'x : Ax \leq B^t, 0 \leq x \leq Z(t)\}$.
- Let X solve $\max\{r'x : Ax \leq B^t, 0 \leq x \leq \mathbb{E}[Z(t)]\}$
- $X_j/\mathbb{E}[Z_j(t)] \in [0, 1]$ for all j.
- Accept j iff $X_j/\mathbb{E}[Z_j(t)] \geq 1/2$.
- $\min\{\hat{q}(t, B^t, \text{accept}), \hat{q}(t, B^t, \text{reject})\} \leq \exp(-t(p_j/2\kappa)^2/25)$

Numerical Results



Conclusions and Future Directions

- Compensated coupling: A way to understand online decision-making

Conclusions and Future Directions

- Compensated coupling: A way to understand online decision-making
- Bayes Selector:
  - Natural policy with a performance guarantee
  - Bridge between estimation and optimization

Conclusions and Future Directions

- Compensated coupling: A way to understand online decision-making
- Bayes Selector:
  - Natural policy with a performance guarantee
  - Bridge between estimation and optimization
- Online packing and matching: the Bayes Selector achieves constant regret
- Future work: learning and multi-stage problems
- Promising directions: online convex optimization

Appendix: Details for Numerical Results

	Type j					
	1	2	3	4	5	6
Valuation	10	6	10	5	9	8
Resource	1	1	2	2	1,2	1,2
p_j	0.2	0.2	0.2	0.2	0.1	0.1