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

$$\max \ r'x$$

s.t. $Ax \leq B$ (Budget Consumption)

$x \leq Z(T)$ (Respect Arrivals)

$x \in \mathbb{N}^n$.

Paradigm: two agents solving this problem, ONLINE and OFFLINE
Traditional Approach and Our Contributions

- Compensated Coupling: for analyzing algorithms
- Bayes Selector: meta-algorithm based on estimators
- Applications: constant regret
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 := $V^{off} - V^{on}$. 
Application: Constant Regret for Online Packing

\[ V^{\text{on}} \text{ and } V^{\text{off}} \text{ collected values, Regret } := V^{\text{off}} - V^{\text{on}}. \]
\[ V^{\text{off}} = \max\{r'x : Ax \leq B, 0 \leq x \leq Z(T)\}. \]
Application: Constant Regret for Online Packing

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

**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)
- Prophet: worst case distribution (competitive ratio)
  - Ratio $1 - e^{-1}$ for maximum of i.i.d.  Hill and Kertz (1982)
  - Best ratio $\approx 0.745$, used in posted pricing  Correa et al. (2017)
- 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 $J^t \in J$ presented at $t = T, T - 1, \ldots, 1$
- States $s \in S$ and actions $a \in A$
Framework for Online Decision-Making

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

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

\[
\begin{array}{ccccccc}
6 & 10 & 10 & 8 & 7 & 8 \\
\end{array}
\]

\[
t = 6 \quad 5 \quad 4 \quad 3 \quad 2 \quad 1
\]

\( t \) 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 \( \omega \).
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 \(\omega\).

\[
\begin{array}{cccccccc}
 t & 6 & 10 & 10 & 8 & 7 & 8 \\
 B = 1 & 10 & 10 & 10 & 8 & 8 & 8 \\
 B = 2 & 10+10 & 10+10 & 10+8 & 8+8 & 8+7 & 8 \\
\end{array}
\]
**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 $\omega$.

```
  6  10  10  8  7  8
```

<table>
<thead>
<tr>
<th>$t$</th>
<th>6</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B = 1$</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>$B = 2$</td>
<td>10+10</td>
<td>10+10</td>
<td>10+8</td>
<td>8+8</td>
<td>8+7</td>
<td>8</td>
</tr>
</tbody>
</table>

It satisfies the Bellman Equation

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

ONLINE uses a policy $\pi^{on}$. We want OFFLINE to follow $\pi^{on}$
Offline Follows Online

**ONLINE** uses a policy $\pi^{on}$. We want **OFFLINE** to follow $\pi^{on}$

What if **OFFLINE** is not satisfied with $a = \pi^{on}(t, s, j)$?

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

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

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

We need to compensate him by paying

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

**ONLINE** uses a policy $\pi^{on}$. We want **OFFLINE** to follow $\pi^{on}$

What if **OFFLINE** is not satisfied with $a = \pi^{on}(t, s, j)$?

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

We need to compensate him by paying

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

How big are the compensations?
Compensations

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

How big are the compensations? $r_{\text{max}}$ for multi-secretary and $|A_j|r_{\text{max}} \leq d r_{\text{max}}$ 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?
\( r_{\text{max}} \) for multi-secretary and \( |A_j| r_{\text{max}} \leq d r_{\text{max}} \) 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*

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

Low-Regret for Online Decision-Making

Cornell University
Compensated Coupling View

- **State**: Offline, Online
- **Time**: $t_1, t_2, t_3, t_4$

**Diagrams**:
1. Offline 1
2. Offline 2
3. Offline 3

**Legend**:
- $S_1, S_2, S_3, S_4$
- Compensation

**Equations**:
- Introduction
- Compensated Coupling
- Bayes Selector
- Conclusion
Compensated Coupling View

State

$S_1$

$S_2$

$S_3$

$S_4$

Time

$t_1$

$t_2$

$t_3$

$t_4$

Offline

Online

Compensation

Low-Regret for Online Decision-Making Cornell University
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 \textbf{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

$$A^t_{BS} \in \text{argmin}\{\mathbb{P}[Q(t, S^t_{BS}, a)] : a \in \mathcal{A}\}.$$
Regret Guarantee of Bayes Selector

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

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

**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 $t$, $A^t_{BS} \in \text{argmin}\{P[Q(t, S^t_{BS}, a)] : a \in A\}$.

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

**Ideal Policy:** at time $t$, $A_{BS}^t \in \arg\min \{ P[Q(t, S_{BS}^t, a)] : a \in A \}$.

**Meta-Algorithm:** Based on estimators $\hat{q}(t, s, a) \geq P[Q(t, s, a)]$ at time $t$, $A_{BS}^t \in \arg\min \{ \hat{q}(t, S_{BS}^t, a) : a \in A \}$.

**Theorem**

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

$$\mathbb{E}[\text{Regret}] \leq \bar{r} \mathbb{E} \left[ \sum_t \hat{q}(t, S_{BS}^t, A_{BS}^t) \right]$$
Regret Guarantee of Bayes Selector

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

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

**Theorem**

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

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

**Theorem**

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

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

- 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

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

- 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

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

- 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

![Graph showing numerical results comparing Bayes Selector, Randomized Adaptive Policy, and Randomized Static Policy with respect to regret and scaling. The graph illustrates the performance of each policy as the scaling parameter increases.]
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

<table>
<thead>
<tr>
<th></th>
<th>Type 1</th>
<th>Type 2</th>
<th>Type 3</th>
<th>Type 4</th>
<th>Type 5</th>
<th>Type 6</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Valuation</strong></td>
<td>10</td>
<td>6</td>
<td>10</td>
<td>5</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td><strong>Resource</strong></td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1,2</td>
<td>1,2</td>
</tr>
<tr>
<td>$p_j$</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
</tr>
</tbody>
</table>