Algorithms for Easy and Hard Online Learning Environments

Yevgeny Seldin

University of Copenhagen
“Classical” (Batch) Machine Learning

Assumption
The samples are independent identically distributed (i.i.d.)
How Online is different from “batch”?

Batch Learning:
- Collect Data
- Analyze
- Act

Online Learning:
- Act
- Analyze
- Get More Data
Examples

• Investment in the stock market
• Online advertising/personalization
• Online routing
• Games
• Robotics
• ...
When do we need Online Learning?

Until recently, statistical theory has been restricted to the design and analysis of sampling experiments in which the size and composition of the samples are completely determined before the experimentation begins. The reasons for this are partly historical, dating back to the time when the statistician was consulted, if at all, only after the experiment was over, and partly intrinsic in the mathematical difficulty of working with anything but a fixed number of independent random variables. A major advance now appears to be in the making with the creation of a theory of the sequential design of experiments, in which the size and composition of the samples are not fixed in advance but are functions of the observations themselves.

(Robbins, 1952)
When do we need Online Learning?

• Interactive learning
• “Adversarial” game-theoretic settings
  ▪ No i.i.d. assumption
• Intelligent data collection / experiment design
• Large-scale data analysis
The Space of Online Learning Problems
The Space of Online Learning Problems
The Space of Online Learning Problems

- stock market
- limited (bandit) feedback
- medical treatments advertising
The Space of Online Learning Problems

Exploration-Exploitation Trade-off

- stock market
- medical treatments advertising
- limited (bandit) feedback

full feedback
Exploration-Exploitation Trade-off

What drug to give to a new patient

When there are more patients to come...

We are building the dataset for ourselves
The Space of Online Learning Problems

- spam filtering
- stock market
- medical treatments
- weather prediction

environmental resistance
adversarial
i.i.d.

full
bandit

feedback
The Space of Online Learning Problems

- spam filtering
- stock market
- medical treatments
- weather prediction

- adversarial
- environmental resistance
- i.i.d.
- full
- bandit
- feedback
Learning in Adversarial Environments

• Game theoretic setting

• Cannot be treated in batch learning

• Evaluation measure: *regret*
  - Difference in performance compared to the best choice in hindsight (out of a limited set)
  - E.g. investment revenue vs. the best stock in hindsight
The Space of Online Learning Problems

- Medical records of different patients
- Subsequent treatments of the same patient
- Environmental resistance
- Adversarial
- I.i.d.
- Stateless
- Contextual
- Markov Decision Processes (MDP)
- Structural complexity

Feedback

Full bandit
medical records of different patients

subsequent treatments of the same patient

Markov Decision Processes (MDP)

environmental resistance

adversarial

i.i.d.

stateless

contextual

structural complexity

Delayed Feedback
Delayed Feedback

Would you like to have a drink?

Yes

No

Would you like to have a drink?

Yes!!!

No

The next morning....
The Space of Online Learning Problems

- Reinforcement Learning
- Markov Decision Processes (MDP)
- Delayed Feedback

Environmental resistance
Adversarial
I.i.d.
Stateless
Contextual
Structural complexity
Subsequent treatments of the same patient
The Space of Online Learning Problems

- environmental resistance
- adversarial
- i.i.d.
- stateless
- contextual
- MDPs
- structural complexity

"Classical" batch ML would be here
Simplicities along the axes

environmental resistance
adversarial
i.i.d.
stateless
contextual
MDPs
structural complexity
full
bandit
feedback
The more the simpler
Simplicities along the axes

- **Gaps** (between action outcomes)
- **Variance/Range** (of action outcomes)

Environmental resistance

Adversarial

I.i.d.

Stateless

Contextual

MDPs

Structural complexity

Feedback

Full

Bandit
Simplicities along the axes

- Reducibility of the state space (context relevance)
- **Mixing time** (in MDPs)
Some “standard” algorithms

- Prediction with expert advice
- Adversarial bandits
- Stochastic bandits

- Environmental resistance
- Adversarial
- i.i.d.
- Stateless
- Full
- Bandit
- Contextual
- MDPs
- Structural complexity

feedback
Typical performance scaling

- $T \ln K$ - the number of actions
- $\sqrt{T \ln K}$ - the number of rounds
- $\Delta(a)$ - suboptimality gap

$K$ – the number of actions
$T$ – the number of rounds
$\Delta(a)$ – suboptimality gap

Diagram showing the relationship between environmental resistance, adversarial resistance, i.i.d. stateless, contextual, MDPs, and feedback with $\sqrt{KT}$, $\sum \frac{\ln T}{\Delta(a)}$, and $\sqrt{T \ln K}$.
"Standard" algorithms

Assume a certain form of simplicity and exploit it

\[ \sqrt{KT} \]

\[ \sum_{a \neq a^*} \frac{\ln T}{\Delta(a)} \]

environmental resistance

adversarial

stateless

contextual

MDPs

structural

Isolated points!

What if the assumptions are violated?
What are we busy with?

Algorithms for ranges of problems

Minimize the assumptions
Prediction with limited advice / Bandits with paid observations
[Seldin, Bartlett, Cramer, Abbasi-Yadkori, ICML, 2014]

Examples:
• Multiple algorithms or/and parametrizations, but limited computational resources. Only a subset can be executed.
• Expensive resources
Environmental resistance in full info


Examples:
- Small deviations from i.i.d.
- Not full adversariality
  ✓ Adaptation to i.i.d.
  ✓ Adaptation to effective outcome range
Environmental resistance in bandits


Examples:
- Contaminated observations
- Adaptation to i.i.d.
- Adaptation to effective outcome range is impossible!

[Gerthinovitz & Lattimore, NIPS, 2016]
Prediction with Limited Advice
[Thune & Seldin, NIPS, 2018]

Starting from at least 2 observations
✓ Adaptation to i.i.d.
✓ Adaptation to effective outcome range
Prediction with Limited Advice
Problem Setting

• Adversarial
  ▪ $\ell_t^i$ arbitrary in $[0,1]$

• I.I.D.
  ▪ $\mathbb{E}[\ell_t^i] = \mu_i$
  ▪ Gaps: $\Delta_i = \mu_i - \min_j \mu_j$

• Effective Loss Range $\varepsilon$
  ▪ $\forall i, j, t: |\ell_t^i - \ell_t^j| \leq \varepsilon$

• Evaluation measure: pseudo-regret
  ▪ $R_T = \mathbb{E}[\sum_{t=1}^T \ell_t^{A_t}] - \min_{i \in \{1, \ldots, K\}} \mathbb{E}[\sum_{t=1}^T \ell_t^i]$
SODA: Second Order Difference Adjustment
[Thune & Seldin, NIPS, 2018]

- The primary action $A_t$ sampled according to
  \[ p_t(i) \propto e^{-\eta_t D_{t-1}^i - \eta_t^2 S_{t-1}^i} \]
- The secondary observation $B_t$ sampled
  \[ \text{uniformly from } \{1, \ldots, K\}\backslash\{A_t\} \]
- Unbiased loss difference estimates
  \[ \Delta \ell_s^i = (K - 1) \mathbb{I}(B_t = i)(\ell_t^{B_t} - \ell_t^{A_t}) \]
- Loss difference estimator and its second moment
  \[ D_t^i = \sum_{s=1}^{t} \Delta \ell_s^i \quad S_t^i = \sum_{s=1}^{t} \left( \Delta \ell_s^i \right)^2 \]
- The learning rate
  \[ \eta_t \approx \sqrt{\frac{\ln K}{\max_{i} S_{t-1}^i}} \]
SODA: Regret Guarantees
[Thune & Seldin, NIPS, 2018]

• Adversarial
  ▪ $R_T = O\left(\varepsilon \sqrt{KT \ln K}\right)$

• Stochastic
  ▪ $R_T = O\left(\frac{K \varepsilon^2}{\ln K} \sum_{i: \Delta_i > 0} \frac{1}{\Delta_i}\right)$

• Knowledge of i.i.d./adversarial not required!
• Knowledge of $\varepsilon$ not required!
• Simultaneous adaptation to two types of easiness!
An optimal algorithm for i.i.d. and adversarial bandits

[Zimmert & Seldin, AISTATS, 2019]
I.I.D. and adversarial bandits

Problem Setting

- Evaluation measure: pseudo-regret (as before)
  \[ R_T = \mathbb{E}[\sum_{t=1}^{T} \ell_t^{A_t}] - \min_{i \in \{1, \ldots, K\}} \mathbb{E}[\sum_{t=1}^{T} \ell_t^i] \]

- Adversarial – as before

- Stochastically Constrained Adversary [Wei & Luo, 2018]
  - \( i^* \) — best action
  - \( \mathbb{E}[\ell_t^i - \ell_t^{i^*}] = \Delta_i \geq 0 \)

- I.I.D.: special case when
  - \( \mathbb{E}[\ell_t^i] = \mu_i \)

- Stochastically Constrained Adversary [Wei & Luo, 2018]
\( \alpha \)-Tsallis-INF

[Zimmert & Seldin, AISTATS, 2019]

- Tsallis entropy
  \[ H_\alpha(x) = \frac{1}{1-\alpha} (1 - \sum_i x_i^\alpha) \]

- INF – Implicitly Normalized Forecaster
  \[ \text{[Audibert & Bubeck, 2009]} \]
\( \alpha \)-Tsallis-INF

[Zimmert & Seldin, AISTATS, 2019]

- \( w_t = \arg \min_{w \in \Delta^k} \langle w, \tilde{L}_{t-1} \rangle + \sum_i \frac{(w^i)^{\alpha}}{\alpha \eta_{t,i}} \)
- Sample \( A_t \sim w_t \)
- Update \( \tilde{L}^i_t = \tilde{L}^i_{t-1} + \frac{\ell^i_t \mathbb{I}(A_t=i)}{w^i_t} \)

- \( \alpha = 1 \) corresponds to entropic regularization
  - EXP3 algorithm
- \( \alpha = 0 \) corresponds to log-barrier
\[ \frac{1}{2} - \text{Tsallis-INF: Regret Guarantees} \]

[Zimmert & Seldin, AISTATS, 2019]

• \( \eta_t = \sqrt{1/t} \)

• Adversarial
  - \( R_T \leq 4\sqrt{KT} + 1 \)

• Stochastically Constrained Adversary
  - \( R_T \leq \sum_{i \neq i^*} \frac{4 \ln T + 20}{\Delta_i} + 4\sqrt{K} \)
  - i.i.d. is a special case

• Both results match the lower bounds (up to constants)!
• No knowledge of i.i.d./adversarial is required!
<table>
<thead>
<tr>
<th>Regime</th>
<th>Upper Bound (\text{Lower Bound})</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BROAD [Wei&amp;Luo,2018]</strong>&lt;br&gt;Log-barrier + doubling (\alpha = 0)</td>
<td>i.i.d.</td>
</tr>
<tr>
<td></td>
<td>adversarial</td>
</tr>
<tr>
<td>(\alpha = \frac{1}{2})</td>
<td>i.i.d. &amp; adversarial</td>
</tr>
<tr>
<td><strong>EXP3++ [Seldin&amp;Lugosi,2017]</strong>&lt;br&gt;Entropic regularization + mix in extra exploration (\alpha = 1)</td>
<td>i.i.d.</td>
</tr>
<tr>
<td></td>
<td>adversarial</td>
</tr>
</tbody>
</table>
1/2-Tsallis-INF: Experiments

[Zimmert & Seldin, AISTATS, 2019]

i.i.d.

Stochastically Constrained Adversary
\frac{1}{2} - Tsallis-INF: Additional Results

[Zimmert & Seldin, AISTATS, 2019]

- Optimality in the moderately contaminated stochastic regime
- I.I.D. and adversarial optimality in utility-based dueling bandits
Summary

The Space of Online Learning Problems

- i.i.d.
- stateless
- contextual
- MDPs
- structural complexity
- environmental resistance
- adversarial
- full
- bandit
- feedback
Summary

“Standard” algorithms – isolated points

Real-world problems – cover the space
We design algorithms for ranges of problems.
If you want to learn more

• Check my homepage
  ▪ [https://sites.google.com/site/yevgenyseldin/](https://sites.google.com/site/yevgenyseldin/)
  ▪ (or Google me)
  ▪ Papers, tutorials, lecture notes, etc...

• Join our Advanced Topics in Machine Learning course
  ▪ Co-taught by Christian Igel
  ▪ September-October 2019
  ▪ WARNING: Math-heavy course
\(\alpha\)-Tsallis-INF: Some Considerations

[Zimmert & Seldin, AISTATS, 2019]

- Exploration rate \(\left(\eta_t^i (\bar{L}_t^i - \bar{L}_t^i*)\right)^{-\frac{1}{1-\alpha}}\)

- For i.i.d. regret of \(\frac{\ln T}{\Delta_i}\) the exploration rate must be \(\frac{1}{\Delta_i^2 t}\)

- \(\mathbb{E}[\bar{L}_t^i - \bar{L}_t^i*] = \Delta_i t\)

- \(\left(\eta_t^i \Delta_i t\right)^{-\frac{1}{1-\alpha}} = \frac{1}{\Delta_i^2 t} \Rightarrow \eta_t^i = \Delta_i^{1-2\alpha} t^{-\alpha}\)

- \(\alpha = \frac{1}{2}\) is the only value for which \(\eta_t^i\) requires no tuning by (unknown) \(\Delta_i\)
\( \alpha \)-Tsallis-INF: Some Considerations

[Zimmert & Seldin, AISTATS, 2019]

- \( \mathbb{E} [\tilde{L}_t^i - \tilde{L}_t^{i*} ] = \Delta_i t \)
- The variance of \( \tilde{L}_t^i - \tilde{L}_t^{i*} \) is of order \( \Delta_i^2 t^2 \)
- \( \tilde{L}_t^i - \tilde{L}_t^{i*} \) cannot be efficiently controlled by Bernstein’s inequality
- Our analysis is based on a self-bounding property of the regret