Contextual Bandit Exploration

John Langford, Microsoft Research, NYC

Machine Learning the Future, March 13, 2017

Reminder: Contextual Bandit Setting

For $t = 1, \ldots, T$:

1. The world produces some context $x \in X$
2. The learner chooses an action $a \in A$
3. The world reacts with reward $r_a \in [0, 1]$

Goal: Learn a good policy for choosing actions given context.

What does learning mean?

Reminder: Contextual Bandit Setting

For $t = 1, \ldots, T$:

1. The world produces some context $x \in X$
2. The learner chooses an action $a \in A$
3. The world reacts with reward $r_a \in [0, 1]$

Goal: Learn a good policy for choosing actions given context.

What does learning mean? Efficiently competing with some large reference class of policies $\Pi = \{\pi : X \to A\}$:

$$\text{Regret} = \max_{\pi \in \Pi} \text{average}_t(r_{\pi(x)} - r_a)$$

What is exploration?

Exploration = Choosing not-obviously best actions to gather information for better performance in the future.

What is exploration?

Exploration = Choosing not-obviously best actions to gather information for better performance in the future.
There are two kinds:

1. Deterministic. Choose action A, then B, then C, then A, then B, ...
2. Randomized. Choose random actions according to some distribution over actions.

What is exploration?

Exploration = Choosing not-obviously best actions to gather information for better performance in the future.
There are two kinds:

1. Deterministic. Choose action A, then B, then C, then A, then B, ...
2. Randomized. Choose random actions according to some distribution over actions.

We discuss Randomized here.

1. There are no good deterministic exploration algorithms in this setting.
2. Supports off-policy evaluation.
3. Randomize = robust to delayed updates, which are very common in practice.

Explore τ then Follow the Leader (Explore-τ)

Explore τ then Follow the Leader (Explore-τ)

Initially, $h = \emptyset$

For the first τ rounds

1. Observe x.
2. Choose a uniform randomly.
3. Observe r, and add $(x, a, r, 1/|A|)$ to h.

For the next $T - \tau$ rounds, use empirical best.

Explore τ then Follow the Leader (Explore-τ)

Initially, $h = \emptyset$

For the first τ rounds

1. Observe x.
2. Choose a uniform randomly.
3. Observe r, and add $(x, a, r, 1/|A|)$ to h.

For the next $T - \tau$ rounds, use empirical best.

Suppose all examples are drawn from a fixed distribution $D(x, \vec{r})$.

Theorem: For all D, Π, Explore-τ has regret $O\left(\frac{\tau}{T} + \sqrt{\frac{|A| \ln |\Pi|}{\tau}}\right)$ with high probability.

Explore τ then Follow the Leader (Explore-τ)

Initially, $h = \emptyset$

For the first τ rounds

1. Observe x.
2. Choose a uniform randomly.
3. Observe r, and add $(x, a, r, 1/|A|)$ to h.

For the next $T - \tau$ rounds, use empirical best.

Suppose all examples are drawn from a fixed distribution $D(x, \vec{r})$.

Theorem: For all D, Π, Explore-τ has regret $O\left(\frac{\tau}{T} + \sqrt{\frac{|A| \ln |\Pi|}{\tau}}\right)$ with high probability.

Proof: After τ rounds, large deviation bound

$$\Rightarrow \left| V(\pi) - E_{(x,\vec{r})\sim D}[r_{\pi(x)}\right| \leq \sqrt{\frac{|A| \ln(|\Pi|/\delta)}{\tau}}$$

Explore τ then Follow the Leader (Explore-τ)

Initially, $h = \emptyset$

For the first τ rounds

1. Observe x.
2. Choose a uniform randomly.
3. Observe r, and add $(x, a, r, 1/|A|)$ to h.

For the next $T - \tau$ rounds, use empirical best.

Suppose all examples are drawn from a fixed distribution $D(x, \vec{r})$.

Theorem: For all D, Π, Explore-τ has regret $O\left(\frac{\tau}{T} + \sqrt{\frac{|A| \ln |\Pi|}{\tau}}\right)$ with high probability.

Proof: After τ rounds, large deviation bound

$$\Rightarrow \left| V(\pi) - E_{(x,\vec{r})\sim D}[r_{\pi(x)}\right| \leq \sqrt{\frac{|A| \ln(|\Pi|/\delta)}{\tau}}$$

so regret bounded by $\frac{\tau}{T} + \frac{T-\tau}{T}\sqrt{\frac{|A| \ln(|\Pi|/\delta)}{\tau}}$

Explore τ then Follow the Leader (Explore-τ)

Initially, $h = \emptyset$

For the first τ rounds

1. Observe x.
2. Choose a uniform randomly.
3. Observe r, and add $(x, a, r, 1/|A|)$ to h.

For the next $T - \tau$ rounds, use empirical best.

Suppose all examples are drawn from a fixed distribution $D(x, \vec{r})$.

Theorem: For all D, Π, Explore-τ has regret $O\left(\frac{\tau}{T} + \sqrt{\frac{|A| \ln |\Pi|}{\tau}}\right)$ with high probability.

Proof: After τ rounds, large deviation bound

$$\Rightarrow \left| V(\pi) - E_{(x,\vec{r})\sim D}[r_{\pi(x)} \right| \leq \sqrt{\frac{|A| \ln(|\Pi|/\delta)}{\tau}}$$

so regret bounded by $\frac{\tau}{T} + \frac{T-\tau}{T}\sqrt{\frac{|A| \ln(|\Pi|/\delta)}{\tau}} = O\left(\left(\frac{|A| \ln |\Pi|}{T}\right)^{1/3}\right)$ at best τ.

Explore-τ summary

1. +Easiest approach: offline prerecorded exploration can feed into any learning algorithm.
2. -Doesn't adapt when world changes.
3. -Underexploration common. A clinical trial problem: no new information after initial exploration.
4. -Overexploration common. Explores obviously suboptimal choices.

Explore-τ summary

1. +Easiest approach: offline prerecorded exploration can feed into any learning algorithm.
2. -Doesn't adapt when world changes.
3. -Underexploration common. A clinical trial problem: no new information after initial exploration.
4. -Overexploration common. Explores obviously suboptimal choices.

Can we do better?

ϵ-Greedy

1. Observe x.
2. With probability $1 - \epsilon$
   1. Choose learned a
   2. Observe r, and learn with $(x, a, r, 1 - \epsilon)$.

ϵ-Greedy

1. Observe x.
2. With probability $1 - \epsilon$
   1. Choose learned a
   2. Observe r, and learn with $(x, a, r, 1 - \epsilon)$.

   With probability ϵ
   1. Choose Uniform random other a
   2. Observe r, and learn with $(x, a, r, \epsilon/(|A| - 1))$.

ϵ-Greedy

1. Observe x.
2. With probability $1 - \epsilon$
   1. Choose learned a
   2. Observe r, and learn with $(x, a, r, 1 - \epsilon)$.

   With probability ϵ
   1. Choose Uniform random other a
   2. Observe r, and learn with $(x, a, r, \epsilon/(|A| - 1))$.

Theorem: ϵ-Greedy has regret $O\left(\epsilon + \sqrt{\frac{|A| \ln |\Pi|}{T\epsilon}}\right)$

ϵ-Greedy

1. Observe x.
2. With probability $1 - \epsilon$
   1. Choose learned a
   2. Observe r, and learn with $(x, a, r, 1 - \epsilon)$.

   With probability ϵ
   1. Choose Uniform random other a
   2. Observe r, and learn with $(x, a, r, \epsilon/(|A| - 1))$.

Theorem: ϵ-Greedy has regret $O\left(\epsilon + \sqrt{\frac{|A| \ln |\Pi|}{T\epsilon}}\right)$

For optimal epsilon?

ϵ-Greedy

1. Observe x.
2. With probability $1 - \epsilon$
   1. Choose learned a
   2. Observe r, and learn with $(x, a, r, 1 - \epsilon)$.

   With probability ϵ
   1. Choose Uniform random other a
   2. Observe r, and learn with $(x, a, r, \epsilon/(|A| - 1))$.

Theorem: ϵ-Greedy has regret $O\left(\epsilon + \sqrt{\frac{|A| \ln |\Pi|}{T\epsilon}}\right)$

For optimal epsilon? $O\left(\left(\frac{|A| \ln |\Pi|}{T}\right)^{1/3}\right)$

ϵ-Greedy summary

1. -Harder Approach: Need online learning algorithm to use.
2. +Adapts when world changes.
3. -Overexploration common. Bad possibilities keep being explored.
4. +Can be adaptive. Epoch Greedy = Epsilon Greedy with adaptive ϵ.

ϵ-Greedy summary

1. -Harder Approach: Need online learning algorithm to use.
2. +Adapts when world changes.
3. -Overexploration common. Bad possibilities keep being explored.
4. +Can be adaptive. Epoch Greedy = Epsilon Greedy with adaptive ϵ.

Can we do better?

Better 1: Bagging Thompson Sampling

Maintain Bayesian posterior over policies.
On each round sample policy from posterior, act with policy.

Better 1: Bagging Thompson Sampling

Maintain Bayesian posterior over policies.
On each round sample policy from posterior, act with policy.
Problem: Posteriors are intractable.

Better 1: Bagging Thompson Sampling

Maintain Bayesian posterior over policies.
On each round sample policy from posterior, act with policy.
Problem: Posteriors are intractable.
Solution: Treat bootstrap as posterior.

Better 1: Bagging Thompson Sampling

Maintain Bayesian posterior over policies.
On each round sample policy from posterior, act with policy.
Problem: Posteriors are intractable.
Solution: Treat bootstrap as posterior.
Problem: Too much variance

Better 1: Bagging Thompson Sampling

Maintain Bayesian posterior over policies.
On each round sample policy from posterior, act with policy.
Problem: Posteriors are intractable.
Solution: Treat bootstrap as posterior.
Problem: Too much variance
Solution: Using bagging to compute probability of action

Better 1: Bagging Thompson Sampling

Maintain Bayesian posterior over policies.
On each round sample policy from posterior, act with policy.
Problem: Posteriors are intractable.
Solution: Treat bootstrap as posterior.
Problem: Too much variance
Solution: Using bagging to compute probability of action

Bagging Thompson Sampling

For each $t = 1, 2, \ldots$

1. Observe x
2. Let $p(a|x) = \Pr_\pi(\pi(x) = a)$
3. Choose $a \sim p(a|x)$
4. Observe reward r.
5. For each π update Poisson(1) times with $(x, a, r, p(a|x))$.

What does it mean?

1. +Avoids unnecessary exploration
2. +Known to work well empirically sometimes.
3. -Heuristic relatively weak theoretical guarantees.
4. -Underexplores sometimes.

Better 2: A Cover Algorithm

For $t = 1, \ldots, T$:

1. The world produces some context $x \in X$

3. The learner chooses an action $a \in A$
4. The world reacts with reward $r_a \in [0, 1]$

Better 2: A Cover Algorithm

Let Q_1 = uniform distribution
For $t = 1, \ldots, T$:

1. The world produces some context $x \in X$
2. Draw $\pi \sim Q_t$
3. The learner chooses an action $a \in A$ using $\pi(x)$.
4. The world reacts with reward $r_a \in [0, 1]$
5. Update Q_{t+1}

What is good Q_t?

- **Exploration:** Q_t allows discovery of good policies
- **Exploitation:** Q_t small on bad policies

What is good Q_t?

- **Exploration:** Q_t allows discovery of good policies
- **Exploitation:** Q_t small on bad policies

At time t define

- Empirical regret $\widehat{\text{Regret}}_t(\pi) = \max_{\pi'} V(\pi') - V(\pi)$
- Minimum probability $\mu = \sqrt{\frac{\ln \Pi}{t|A|}}$
- Regret certainty $b_\pi = \widehat{\text{Regret}}_t(\pi)/(100\mu)$

What is good Q_t?

- **Exploration:** Q_t allows discovery of good policies
- **Exploitation:** Q_t small on bad policies

At time t define

- Empirical regret $\widehat{\text{Regret}}_t(\pi) = \max_{\pi'} V(\pi') - V(\pi)$
- Minimum probability $\mu = \sqrt{\frac{\ln \Pi}{t|A|}}$
- Regret certainty $b_\pi = \widehat{\text{Regret}}_t(\pi)/(100\mu)$

Find a distribution Q over policies π satisfying

$$\text{Estimated regret} \leq \text{(small)}$$

$$\text{Estimated variance} \leq \text{(small for good } \pi\text{)}$$

What is good Q_t?

- **Exploration:** Q_t allows discovery of good policies
- **Exploitation:** Q_t small on bad policies

At time t define

- Empirical regret $\widehat{\text{Regret}}_t(\pi) = \max_{\pi'} V(\pi') - V(\pi)$
- Minimum probability $\mu = \sqrt{\frac{\ln \Pi}{t|A|}}$
- Regret certainty $b_\pi = \widehat{\text{Regret}}_t(\pi)/(100\mu)$

Find a distribution Q over policies π satisfying

$$\sum_{\pi \in \Pi} Q(\pi) b_\pi \;\leq\; 2|A|$$

$$\forall \pi \in \Pi \;:\; \frac{1}{t} \sum_{\tau=1}^{t} \left[\frac{1}{Q^\mu(\pi(x_\tau)|x_\tau)} \right] \;\leq\; 2|A| + b_\pi$$

How do you find Q_t?

How do you find Q_t?

by Reduction to ArgMax Oracle (AMO).

Definition

Given a set of policies Π and data $(x_1, v_1), \ldots, (x_t, v_t)$, AMO returns

$$\arg\max_{\pi \in \Pi} \sum_{\tau=1}^{t} v_\tau(\pi(x_\tau))$$

How do you find Q_t?

by Reduction to ArgMax Oracle (AMO).

Definition

Given a set of policies Π and data $(x_1, v_1), \ldots, (x_t, v_t)$, AMO returns

$$\arg\max_{\pi \in \Pi} \sum_{\tau=1}^{t} v_\tau(\pi(x_\tau))$$

Use:

$$v_t(a) = \text{Value}_t(a) + \frac{100\mu}{Q^\mu(a|x_t)}$$

to get worst constraint violater.

Mixture of successive constraint violaters $= Q_t$. Very fun to prove!

Theorem: Optimal in all ways

Regret: $\tilde{O}\left(\sqrt{\frac{|A| \ln |\Pi|}{T}}\right)$

Calls to Cost sensitive classification oracle: $\tilde{O}\left(T^{0.5}\right)$ $(< T!)$

Lower bound: $\Omega\left(T^{0.5}\right)$ calls to oracle

Running time: $\tilde{O}\left(T^{1.5}\right)$

Trying it out

Change rcv1 CCAT-or-not to be classes 1 and 2

vw --cbify 2 rcv1.train.multiclass.vw -c --epsilon 0.1
  Progressive 0/1 loss: 0.156
vw --cbify 2 rcv1.train.multiclass.vw -c --first 20000
  Progressive 0/1 loss: 0.082
vw --cbify 2 rcv1.train.multiclass.vw -c --bag 16 -b 22
  Progressive 0/1 loss: 0.059
vw --cbify 2 rcv1.train.multiclass.vw -c --cover 1
  Progressive 0/1 loss: 0.053

Trying it out

Change rcv1 CCAT-or-not to be classes 1 and 2

vw --cbify 2 rcv1.train.multiclass.vw -c --epsilon 0.1
  Progressive 0/1 loss: 0.156
vw --cbify 2 rcv1.train.multiclass.vw -c --first 20000
  Progressive 0/1 loss: 0.082
vw --cbify 2 rcv1.train.multiclass.vw -c --bag 16 -b 22
  Progressive 0/1 loss: 0.059
vw --cbify 2 rcv1.train.multiclass.vw -c --cover 1
  Progressive 0/1 loss: 0.053

	ϵ-greedy	Initial	Bagging	LinUCB	Online Cover	Supervised
Loss	0.148	0.081	0.059	0.128	0.053	0.051
time	17s	2.6s	275s	60h	12s	5.3s

Bibliography

Tau-first Unclear first use?

ϵ-Greedy Unclear first use?

Online Bag N. Oza and S. Russell, Online bagging and boosting, AI&Stat 2001.

Epoch J. Langford and T. Zhang, The Epoch-Greedy Algorithm for Contextual Multi-armed Bandits, NIPS 2007.

Thompson W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika, 25(3-4):285294, 1933.

Cover/Bag A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, R. Schapire, Taming the Monster: A Fast and Simple Algorithm for Contextual Bandits, ICML 2014.

Bootstrap D. Eckles and M. Kaptein, Thompson Sampling with Online Bootstrap, arxiv.org/1410.4009