Doing Exploration

John Langford, Microsoft Research, NYC

Large Scale Learning Class, April 9, 2013

(Post presentation version)

Reminder: Contextual Bandit Setting

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

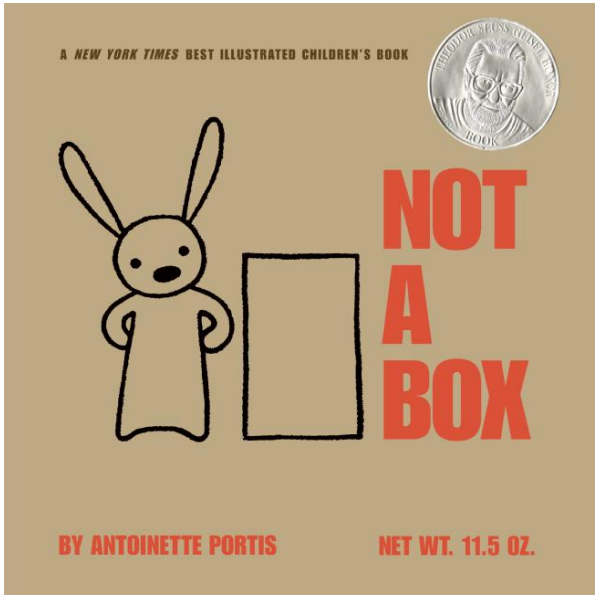
1. The world produces some context $x \in X$
2. The learner chooses an action $a \in A$
3. The world reacts with reward $r_a \in [0, 1]$

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

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

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

A Basic Observation



This is not a supervised learning problem:

- We don't know the reward of actions not taken—loss function is unknown even at training time.
- Exploration is required to succeed. No exploration $\Rightarrow$ no basis for sound decision making.

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. (See first half.)
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) to h.

For the next T 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) to h.

For the next T 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) to h.

For the next T 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, a large deviation bound implies empirical average value of a policy deviates from expectation $E_{(x,\vec{r})\sim D}[r_{\pi(x)}]$ by at most $\sqrt{\frac{|A| \ln(|\Pi|/\delta)}{\tau}}$, so regret is bounded by $\frac{\tau}{T} + \frac{T}{T}\sqrt{\frac{|A| \ln(|\Pi|/\delta)}{\tau}}$.

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) to h.

For the next T 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, a large deviation bound implies empirical average value of a policy deviates from expectation $E_{(x,\vec{r})\sim D}[r_{\pi(x)}]$ by at most $\sqrt{\frac{|A| \ln(|\Pi|/\delta)}{\tau}}$, so regret is bounded by $\frac{\tau}{T} + \frac{T}{T}\sqrt{\frac{|A| \ln(|\Pi|/\delta)}{\tau}}$.

At optimal 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) to h.

For the next T 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, a large deviation bound implies empirical average value of a policy deviates from expectation $E_{(x,\vec{r})\sim D}[r_{\pi(x)}]$ by at most $\sqrt{\frac{|A| \ln(|\Pi|/\delta)}{\tau}}$, so regret is bounded by $\frac{\tau}{T} + \frac{T}{T}\sqrt{\frac{|A| \ln(|\Pi|/\delta)}{\tau}}$.

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

What does this mean?

1. +Easiest approach: offline prerecorded exploration can feed into any learning algorithm. See first half.
2. -Doesn't adapt when world changes.
3. -Underexploration common. Think of clinical trials.

What does this mean?

1. +Easiest approach: offline prerecorded exploration can feed into any learning algorithm. See first half.
2. -Doesn't adapt when world changes.
3. -Underexploration common. Think of clinical trials.

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

What does this mean?

1. -Harder Approach: Need online learning algorithm to use.
2. +Adapts when world changes.
3. -Overexploration common. Bad possibilities keep being explored.

What does this mean?

1. -Harder Approach: Need online learning algorithm to use.
2. +Adapts when world changes.
3. -Overexploration common. Bad possibilities keep being explored.

Can we do better?

Epoch Greedy

At every timestep t, the learned policy has an empirical performance known up to some precision ϵ_t which can be estimated.

Epoch Greedy

At every timestep t, the learned policy has an empirical performance known up to some precision ϵ_t which can be estimated.

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

Epoch Greedy

At every timestep t, the learned policy has an empirical performance known up to some precision ϵ_t which can be estimated.

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

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

Epoch Greedy

At every timestep t, the learned policy has an empirical performance known up to some precision ϵ_t which can be estimated.

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

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

Theorem: Epoch Greedy has regret $O\left(\left(\frac{|A| \ln |\Pi|}{T}\right)^{1/3}\right)$ with high probability.

Autotuning!

What does this mean?

1. -Harder Approach: Need online learning algorithm to use + keeping track of deviation bound.
2. +Adapts when world changes.
3. +Neither under nor over exploration.

What does this mean?

1. -Harder Approach: Need online learning algorithm to use + keeping track of deviation bound.
2. +Adapts when world changes.
3. +Neither under nor over exploration.

Is it possible to do better?

Better 1: Policy Elimination

Policy_Elimination

Let $\Pi_0 = \Pi =$ initial set of policies.
For each $t = 1, 2, \ldots$

1. Choose distribution P over remaining policies Π_{t-1} so every remaining policy π has small expected variance in value estimate.

Better 1: Policy Elimination

Policy_Elimination

Let $\Pi_0 = \Pi =$ initial set of policies.
For each $t = 1, 2, \ldots$

1. Choose distribution P over remaining policies Π_{t-1} so every remaining policy π has small expected variance in value estimate.
2. observe x

Better 1: Policy Elimination

Policy_Elimination

Let $\Pi_0 = \Pi =$ initial set of policies.
For each $t = 1, 2, \ldots$

1. Choose distribution P over remaining policies Π_{t-1} so every remaining policy π has small expected variance in value estimate.
2. observe x
3. Let $p(a) =$ fraction of policies from P choosing a given x.

Better 1: Policy Elimination

Policy_Elimination

Let $\Pi_0 = \Pi =$ initial set of policies.
For each $t = 1, 2, \ldots$

1. Choose distribution P over remaining policies Π_{t-1} so every remaining policy π has small expected variance in value estimate.
2. observe x
3. Let $p(a) =$ fraction of policies from P choosing a given x.
4. Choose $a \sim p$ and observe reward r.

Better 1: Policy Elimination

Policy_Elimination

Let $\Pi_0 = \Pi =$ initial set of policies.
For each $t = 1, 2, \ldots$

1. Choose distribution P over remaining policies Π_{t-1} so every remaining policy π has small expected variance in value estimate.
2. observe x
3. Let $p(a) =$ fraction of policies from P choosing a given x.
4. Choose $a \sim p$ and observe reward r.
5. Let $\Pi_t =$ remaining near empirical best policies.

Better 1: Policy Elimination

Policy_Elimination

Let $\Pi_0 = \Pi =$ initial set of policies.
For each $t = 1, 2, \ldots$

1. Choose distribution P over remaining policies Π_{t-1} so every remaining policy π has small expected variance in value estimate.
2. observe x
3. Let $p(a) =$ fraction of policies from P choosing a given x.
4. Choose $a \sim p$ and observe reward r.
5. Let $\Pi_t =$ remaining near empirical best policies.

Theorem: With high probability Policy_Elimination has expected regret

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

What does this mean?

1. -Doesn't adapt when world changes.
2. ++Much more efficient exploration. Only efficient in special cases.
3. - -Much Harder Approach: Need to keep track of policies, which is often intractable.

What does this mean?

1. -Doesn't adapt when world changes.
2. ++Much more efficient exploration. Only efficient in special cases.
3. - -Much Harder Approach: Need to keep track of policies, which is often intractable.

Adapting algorithms exist (EXP4).
More efficient versions exist (RUCB), but not yet efficient enough.

Better 2: Thompson Sampling

Always maintain a Bayesian posterior over policies.
On each round sample policy from posterior, and act according to it.

Better 2: Thompson Sampling

Always maintain a Bayesian posterior over policies.
On each round sample policy from posterior, and act according to it.
An efficient special case: Gaussian Posterior.

Thompson Sampling

Let w = mean 0 multivariate gaussian.
For each $t = 1, 2, \ldots$

1. Draw $w' \sim w$
2. Observe x
3. Choose $a = \max_{a'} w' x_{a'}$
4. Observe reward r.
5. Bayesian update w with (x, a, r).

What does it mean?

1. +Efficient special cases for Gaussian posteriors.
2. +Known to work well empirically sometimes.
3. -Not robust to model misspecification.

The current state

Starter	
Baseline	
Purring	
Shiny	
Something to try	

The current state

Explore-τ	Simplest Possible
Baseline	
Purring	
Shiny	
Something to try	

The current state

Explore-τ	Simplest Possible
ϵ-Greedy	Simplest Adaptive
Purring	
Shiny	
Something to try	

The current state

Explore-τ	Simplest Possible
ϵ-Greedy	Simplest Adaptive
Epoch Greedy	Unequivocal Improvement
Shiny	
Something to try	

The current state

Explore-τ	Simplest Possible
ϵ-Greedy	Simplest Adaptive
Epoch Greedy	Unequivocal Improvement
Policy Elimination	Intriguing Impracticality
Something to try	

The current state

Explore-τ	Simplest Possible
ϵ-Greedy	Simplest Adaptive
Epoch Greedy	Unequivocal Improvement
Policy Elimination	Intriguing Impracticality
Thompson Sampling	Sometimes Excellent

The current state

Explore-τ	Simplest Possible
ϵ-Greedy	Simplest Adaptive
Epoch Greedy	Unequivocal Improvement
Policy Elimination	Intriguing Impracticality
Thompson Sampling	Sometimes Excellent

You can see the edge of the understood world in this lecture. We hope to see further soon.

Bibliography

Tau-first Unclear first use?

ϵ-Greedy Unclear first use?

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

EXP4 P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. SIAM Journal of Computing, 32(1):4877, 2002b.

PolyElim M. Dudik, D. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, T. Zhang, Efficient Optimal Learning for Contextual Bandits, UAI 2011.

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.