Adaptively Exploiting d-Separators with Causal Bandits

Blair Bilodeau
(Joint work with Linbo Wang and Daniel M. Roy)
University of Toronto, Department of Statistical Sciences

May 25, 2022
University of Toronto DoSS Student Research Day

Motivation

Motivation

We want to learn the intervention that has the largest causal effect on a variable of interest.

Motivation

We want to learn the intervention that has the largest causal effect on a variable of interest.

E.g., causal effect of a gene on disease markers or permutations of treatment plans on injury recovery.

Motivation

We want to learn the intervention that has the largest causal effect on a variable of interest.

E.g., causal effect of a gene on disease markers or permutations of treatment plans on injury recovery.

To learn this from solely observational data, we must make strong assumptions about the causal graph.

Motivation

We want to learn the intervention that has the largest causal effect on a variable of interest.

E.g., causal effect of a gene on disease markers or permutations of treatment plans on injury recovery.

To learn this from solely observational data, we must make strong assumptions about the causal graph.

To circumvent such assumptions, we must be able to actually perform some interventions...

...but they are too numerous and too costly to repeatedly try all of them.

Motivation

We want to learn the intervention that has the largest causal effect on a variable of interest.

E.g., causal effect of a gene on disease markers or permutations of treatment plans on injury recovery.

To learn this from solely observational data, we must make strong assumptions about the causal graph.

To circumvent such assumptions, we must be able to actually perform some interventions...

...but they are too numerous and too costly to repeatedly try all of them.

How can we most efficiently select which interventions to perform?

Motivation

We want to learn the intervention that has the largest causal effect on a variable of interest.

E.g., causal effect of a gene on disease markers or permutations of treatment plans on injury recovery.

To learn this from solely observational data, we must make strong assumptions about the causal graph.

To circumvent such assumptions, we must be able to actually perform some interventions...

...but they are too numerous and too costly to repeatedly try all of them.

How can we most efficiently select which interventions to perform?

It turns out that these same causal assumptions will allow for more efficient intervening...

...but if we rely on these assumptions and they fail, we may learn biased estimates.

Motivation

We want to learn the intervention that has the largest causal effect on a variable of interest.

E.g., causal effect of a gene on disease markers or permutations of treatment plans on injury recovery.

To learn this from solely observational data, we must make strong assumptions about the causal graph.

To circumvent such assumptions, we must be able to actually perform some interventions...

...but they are too numerous and too costly to repeatedly try all of them.

How can we most efficiently select which interventions to perform?

It turns out that these same causal assumptions will allow for more efficient intervening...

...but if we rely on these assumptions and they fail, we may learn biased estimates.

We provide a method to more efficiently perform interventions if causal structure exists...

...while still consistently learning the best intervention when it doesn't.

Motivation

We want to learn the intervention that has the largest causal effect on a variable of interest.
E.g., causal effect of a gene on disease markers or permutations of treatment plans on injury recovery.

To learn this from solely observational data, we must make strong assumptions about the causal graph.

To circumvent such assumptions, we must be able to actually perform some interventions...
...but they are too numerous and too costly to repeatedly try all of them.

How can we most efficiently select which interventions to perform?

It turns out that these same causal assumptions will allow for more efficient intervening...
...but if we rely on these assumptions and they fail, we may learn biased estimates.

We provide a method to more efficiently perform interventions if causal structure exists...
...while still consistently learning the best intervention when it doesn't.
That is, our method *adapts* to the presence of causal structure.

Motivation

We want to learn the intervention that has the largest causal effect on a variable of interest.

E.g., causal effect of a gene on disease markers or permutations of treatment plans on injury recovery.

To learn this from solely observational data, we must make strong assumptions about the causal graph.

To circumvent such assumptions, we must be able to actually perform some interventions...

...but they are too numerous and too costly to repeatedly try all of them.

How can we most efficiently select which interventions to perform?

It turns out that these same causal assumptions will allow for more efficient intervening...

...but if we rely on these assumptions and they fail, we may learn biased estimates.

We provide a method to more efficiently perform interventions if causal structure exists...

...while still consistently learning the best intervention when it doesn't.

That is, our method *adapts* to the presence of causal structure.

Let's formalize our setting.

Let's formalize our setting.

Multi-Armed Bandits

Multi-Armed Bandits

Fix a set of allowable *actions* $\mathcal{A}$.

Multi-Armed Bandits

Fix a set of allowable *actions* $\mathcal{A}$.

For each interaction $t \in [T]$:

Causal Inference with Interventions via Multi-Armed Bandits

Multi-Armed Bandits

Fix a set of allowable *actions* $\mathcal{A}$.

For each interaction $t \in [T]$:

- Player selects $A_t \in \mathcal{A}$

Causal Inference with Interventions via Multi-Armed Bandits

Multi-Armed Bandits

Fix a set of allowable *actions* $\mathcal{A}$.

For each interaction $t \in [T]$:

- Player selects $A_t \in \mathcal{A}$
- Environment generates $Y_t \in [0, 1]$

Causal Inference with Interventions via Multi-Armed Bandits

Multi-Armed Bandits

Fix a set of allowable *actions* $\mathcal{A}$.

For each interaction $t \in [T]$:

- Player selects $A_t \in \mathcal{A}$
- Environment generates $Y_t \in [0, 1]$
- Player observes Y_t

Causal Inference with Interventions via Multi-Armed Bandits

Multi-Armed Bandits

Fix a set of allowable *actions* $\mathcal{A}$.

For each interaction $t \in [T]$:

- Player selects $A_t \in \mathcal{A}$
- Environment generates $Y_t \in [0, 1]$
- Player observes Y_t

We do not observe the counterfactual:

How would the environment have generated Y_t if the player picked a different A_t?

Causal Inference with Interventions via Multi-Armed Bandits

Multi-Armed Bandits

Fix a set of allowable *actions* $\mathcal{A}$.

For each interaction $t \in [T]$:

- Player selects $A_t \in \mathcal{A}$
- Environment generates $Y_t \in [0,1]$
- Player observes Y_t

We do not observe the counterfactual:

How would the environment have generated Y_t if the player picked a different A_t?

Suppose we also observe other variables that might help us learn this.

Causal Inference with Interventions via Multi-Armed Bandits

Multi-Armed Bandits with Post-Action Contexts

Fix a set of allowable *actions* $\mathcal{A}$ and *post-action contexts* $\mathcal{Z}$.

For each interaction $t \in [T]$:

- Player selects $A_t \in \mathcal{A}$
- Environment generates $Y_t \in [0, 1]$ and $Z_t \in \mathcal{Z}$
- Player observes Y_t and Z_t

We do not observe the counterfactual:

How would the environment have generated Y_t and Z_t if the player picked a different A_t?

Suppose we also observe other variables that might help us learn this.

Causal Inference with Interventions via Multi-Armed Bandits

Multi-Armed Bandits with Post-Action Contexts

Fix a set of allowable *actions* $\mathcal{A}$ and *post-action contexts* $\mathcal{Z}$.

For each interaction $t \in [T]$:

- Player selects $A_t \in \mathcal{A}$
- Environment generates $Y_t \in [0, 1]$ and $Z_t \in \mathcal{Z}$
- Player observes Y_t and Z_t

We do not observe the counterfactual:

How would the environment have generated Y_t and Z_t if the player picked a different A_t?

Suppose we also observe other variables that might help us learn this.

Causal Inference with Interventions via Multi-Armed Bandits

Multi-Armed Bandits with Post-Action Contexts

Fix a set of allowable *actions* $\mathcal{A}$ and *post-action contexts* $\mathcal{Z}$.

For each interaction $t \in [T]$:

- Player selects $A_t \in \mathcal{A}$
- Environment generates $Y_t \in [0, 1]$ and $Z_t \in \mathcal{Z}$
- Player observes Y_t and Z_t

We do not observe the counterfactual:

How would the environment have generated Y_t and Z_t if the player picked a different A_t?

Suppose we also observe other variables that might help us learn this.

We have no guarantees that observing Z_t will help us...but we would like to exploit it when we can.

Causal Inference with Interventions via Multi-Armed Bandits

Multi-Armed Bandits with Post-Action Contexts

Fix a set of allowable *actions* $\mathcal{A}$ and *post-action contexts* $\mathcal{Z}$.

For each interaction $t \in [T]$:

- Player selects $A_t \in \mathcal{A}$
- Environment generates $Y_t \in [0,1]$ and $Z_t \in \mathcal{Z}$
- Player observes Y_t and Z_t

We do not observe the counterfactual:

How would the environment have generated Y_t and Z_t if the player picked a different A_t?

Suppose we also observe other variables that might help us learn this.

We have no guarantees that observing Z_t will help us...but we would like to exploit it when we can.

An **environment** ν is a *fixed* collection of distributions on $(\mathcal{Z}, \mathcal{Y})$: one for each $a \in \mathcal{A}$.

Multi-Armed Bandits with Post-Action Contexts

Fix a set of allowable *actions* $\mathcal{A}$ and *post-action contexts* $\mathcal{Z}$.

For each interaction $t \in [T]$:

- Player selects $A_t \in \mathcal{A}$
- Environment generates $Y_t \in [0, 1]$ and $Z_t \in \mathcal{Z}$
- Player observes Y_t and Z_t

We do not observe the counterfactual:

How would the environment have generated Y_t and Z_t if the player picked a different A_t?

Suppose we also observe other variables that might help us learn this.

We have no guarantees that observing Z_t will help us...but we would like to exploit it when we can.

An **environment** ν is a *fixed* collection of distributions on $(\mathcal{Z}, \mathcal{Y})$: one for each $a \in \mathcal{A}$.

A **policy** π maps the observed history (not counterfactuals) to actions.

Causal Inference with Interventions via Multi-Armed Bandits

Multi-Armed Bandits with Post-Action Contexts

Fix a set of allowable *actions* $\mathcal{A}$ and *post-action contexts* $\mathcal{Z}$.

For each interaction $t \in [T]$:

- Player selects $A_t \in \mathcal{A}$
- Environment generates $Y_t \in [0, 1]$ and $Z_t \in \mathcal{Z}$
- Player observes Y_t and Z_t

We do not observe the counterfactual:

How would the environment have generated Y_t and Z_t if the player picked a different A_t?

Suppose we also observe other variables that might help us learn this.

We have no guarantees that observing Z_t will help us...but we would like to exploit it when we can.

An **environment** ν is a *fixed* collection of distributions on $(\mathcal{Z}, \mathcal{Y})$: one for each $a \in \mathcal{A}$.

A **policy** π maps the observed history (not counterfactuals) to actions.

Regret: $R_{\nu,\pi}(T) = T \cdot \max_{a \in \mathcal{A}} \mathbb{E}_{\nu_a}\left[Y\right] - \mathbb{E}_{\nu,\pi}\left[\sum_{t=1}^{T} Y_t\right].$

What do we actually achieve?

What do we actually achieve?

The Punchline: High-Level Overview of Results

The Punchline: High-Level Overview of Results

We introduce the *conditionally benign property*. Informally, $\nu_a(Y \mid Z)$ does not depend on a.

The Punchline: High-Level Overview of Results

We introduce the *conditionally benign property*. Informally, $\nu_a(Y \mid Z)$ does not depend on a.

Generalizes causal structure that previous work exploits non-adaptively.

The Punchline: High-Level Overview of Results

We introduce the *conditionally benign property*. Informally, $\nu_a(Y \mid Z)$ does not depend on a.

Generalizes causal structure that previous work exploits non-adaptively.

We show existing methods can incur *worst-case regret* (linear in T) when assumptions fail.

The Punchline: High-Level Overview of Results

We introduce the *conditionally benign property*. Informally, $\nu_a(Y \mid Z)$ does not depend on a.
Generalizes causal structure that previous work exploits non-adaptively.

We show existing methods can incur *worst-case regret* (linear in T) when assumptions fail.

We formalize *adaptive minimax optimality* for the conditionally benign property.

The Punchline: High-Level Overview of Results

We introduce the *conditionally benign property*. Informally, $\nu_a(Y \mid Z)$ does not depend on a.
Generalizes causal structure that previous work exploits non-adaptively.

We show existing methods can incur *worst-case regret* (linear in T) when assumptions fail.

We formalize *adaptive minimax optimality* for the conditionally benign property.
We show optimality is impossible: realizing the benefits necessarily sacrifices worst-case performance.

The Punchline: High-Level Overview of Results

We introduce the *conditionally benign property*. Informally, $\nu_a(Y \mid Z)$ does not depend on a.

Generalizes causal structure that previous work exploits non-adaptively.

We show existing methods can incur *worst-case regret* (linear in T) when assumptions fail.

We formalize *adaptive minimax optimality* for the conditionally benign property.

We show optimality is impossible: realizing the benefits necessarily sacrifices worst-case performance.

We provide a new algorithm that:

The Punchline: High-Level Overview of Results

We introduce the *conditionally benign property*. Informally, $\nu_a(Y \mid Z)$ does not depend on a.
Generalizes causal structure that previous work exploits non-adaptively.

We show existing methods can incur *worst-case regret* (linear in T) when assumptions fail.

We formalize *adaptive minimax optimality* for the conditionally benign property.
We show optimality is impossible: realizing the benefits necessarily sacrifices worst-case performance.

We provide a new algorithm that:
a) achieves optimal performance when the environment is conditionally benign;

The Punchline: High-Level Overview of Results

We introduce the *conditionally benign property*. Informally, $\nu_a(Y \mid Z)$ does not depend on a.
Generalizes causal structure that previous work exploits non-adaptively.

We show existing methods can incur *worst-case regret* (linear in T) when assumptions fail.

We formalize *adaptive minimax optimality* for the conditionally benign property.
We show optimality is impossible: realizing the benefits necessarily sacrifices worst-case performance.

We provide a new algorithm that:

a) achieves optimal performance when the environment is conditionally benign;
b) always achieves sublinear regret;

The Punchline: High-Level Overview of Results

We introduce the *conditionally benign property*. Informally, $\nu_a(Y \mid Z)$ does not depend on a.
Generalizes causal structure that previous work exploits non-adaptively.

We show existing methods can incur *worst-case regret* (linear in T) when assumptions fail.

We formalize *adaptive minimax optimality* for the conditionally benign property.
We show optimality is impossible: realizing the benefits necessarily sacrifices worst-case performance.

We provide a new algorithm that:

a) achieves optimal performance when the environment is conditionally benign;
b) always achieves sublinear regret;
c) incurs optimal $\sqrt{T}$ regret for some cases where existing algorithms incur linear regret.

The Punchline: High-Level Overview of Results

We introduce the *conditionally benign property*. Informally, $\nu_a(Y \mid Z)$ does not depend on a.
Generalizes causal structure that previous work exploits non-adaptively.

We show existing methods can incur *worst-case regret* (linear in T) when assumptions fail.

We formalize *adaptive minimax optimality* for the conditionally benign property.
We show optimality is impossible: realizing the benefits necessarily sacrifices worst-case performance.

We provide a new algorithm that:

a) achieves optimal performance when the environment is conditionally benign;
b) always achieves sublinear regret;
c) incurs optimal $\sqrt{T}$ regret for some cases where existing algorithms incur linear regret.

What are we precisely trying to adapt to?

What are we precisely trying to adapt to?

Existing Results

Existing Results

Classically known (Auer et al. 2002) that the minimax regret is

$$R_{\nu,\pi}(T) = \Theta\left(\sqrt{|\mathcal{A}|\, T}\right).$$

Existing Results

Classically known (Auer et al. 2002) that the minimax regret is

$$R_{\nu,\pi}(T) = \Theta\left(\sqrt{|\mathcal{A}|\, T}\right).$$

This is achieved by, for example, the UCB algorithm.

Existing Results

Classically known (Auer et al. 2002) that the minimax regret is

$$R_{\nu,\pi}(T) = \Theta\left(\sqrt{|\mathcal{A}|\, T}\right).$$

This is achieved by, for example, the UCB algorithm.

Recently shown (Lu et al. 2020) that if Z is all causal parents of Y (under ν) and we know $(\nu_a(Z))_{a\in\mathcal{A}}$,

$$R_{\nu,\pi}(T) = \tilde{\Theta}\left(\sqrt{|\mathcal{Z}|\, T}\right).$$

Existing Results

Classically known (Auer et al. 2002) that the minimax regret is

$$R_{\nu,\pi}(T) = \Theta\left(\sqrt{|\mathcal{A}|\, T}\right).$$

This is achieved by, for example, the UCB algorithm.

Recently shown (Lu et al. 2020) that if Z is all causal parents of Y (under ν) and we know $(\nu_a(Z))_{a\in\mathcal{A}}$,

$$R_{\nu,\pi}(T) = \tilde{\Theta}\left(\sqrt{|\mathcal{Z}|\, T}\right).$$

They provide an algorithm that achieves this: C-UCB.

Existing Results

Classically known (Auer et al. 2002) that the minimax regret is

$$R_{\nu,\pi}(T) = \Theta\left(\sqrt{|\mathcal{A}|\, T}\right).$$

This is achieved by, for example, the UCB algorithm.

Recently shown (Lu et al. 2020) that if Z is all causal parents of Y (under ν) and we know $(\nu_a(Z))_{a\in\mathcal{A}}$,

$$R_{\nu,\pi}(T) = \tilde{\Theta}\left(\sqrt{|\mathcal{Z}|\, T}\right).$$

They provide an algorithm that achieves this: C-UCB.

A natural question: Can these regret bounds be achieved *adaptively*?

Existing Results

Classically known (Auer et al. 2002) that the minimax regret is

$$R_{\nu,\pi}(T) = \Theta\left(\sqrt{|\mathcal{A}|\,T}\right).$$

This is achieved by, for example, the UCB algorithm.

Recently shown (Lu et al. 2020) that if Z is all causal parents of Y (under ν) and we know $(\nu_a(Z))_{a\in\mathcal{A}}$,

$$R_{\nu,\pi}(T) = \tilde{\Theta}\left(\sqrt{|\mathcal{Z}|\,T}\right).$$

They provide an algorithm that achieves this: C-UCB.

A natural question: Can these regret bounds be achieved *adaptively*?

Adapting to Causal Structure

Theorem: Existing algorithms do not adapt to failure of assumptions.

For every $\mathcal{A}$ and $\mathcal{Z}$, there exists ν such that

$$\lim_{T\to\infty} \frac{R_{\nu,\text{C-UCB}}(T)}{T} \geq 1/120.$$

Adapting to Causal Structure

Theorem: Existing algorithms do not adapt to failure of assumptions.

For every $\mathcal{A}$ and $\mathcal{Z}$, there exists ν such that

$$\lim_{T\to\infty} \frac{R_{\nu,\text{C-UCB}}(T)}{T} \geq 1/120.$$

When can we use Z to learn counterfactuals?

Adapting to Causal Structure

Theorem: Existing algorithms do not adapt to failure of assumptions.

For every $\mathcal{A}$ and $\mathcal{Z}$, there exists ν such that

$$\lim_{T\to\infty} \frac{R_{\nu,\text{C-UCB}}(T)}{T} \geq 1/120.$$

When can we use Z to learn counterfactuals?

When the choice of A doesn't affect the outcome of Y conditional on Z.

Adapting to Causal Structure

Theorem: Existing algorithms do not adapt to failure of assumptions.

For every $\mathcal{A}$ and $\mathcal{Z}$, there exists ν such that

$$\lim_{T\to\infty} \frac{R_{\nu,\text{C-UCB}}(T)}{T} \geq 1/120.$$

When can we use Z to learn counterfactuals?

When the choice of A doesn't affect the outcome of Y conditional on Z.

Definition (informal)

An environment ν is conditionally benign if and only if $\nu_a(Y \mid Z)$ is constant as a function of $a \in \mathcal{A}$.

Adapting to Causal Structure

Theorem: Existing algorithms do not adapt to failure of assumptions.

For every $\mathcal{A}$ and $\mathcal{Z}$, there exists ν such that

$$\lim_{T\to\infty} \frac{R_{\nu,\text{C-UCB}}(T)}{T} \geq 1/120.$$

When can we use Z to learn counterfactuals?

When the choice of A doesn't affect the outcome of Y conditional on Z.

Definition (informal)

An environment ν is conditionally benign if and only if $\nu_a(Y \mid Z)$ is constant as a function of $a \in \mathcal{A}$.

Theorem: Strict adaptation to the conditionally benign property is impossible.

If $\mathfrak{a}$ always satisfies $R_{\nu,\mathfrak{a}}(T) \leq \sqrt{|\mathcal{A}|T}$,
there exists a conditionally benign ν such that $R_{\nu,\mathfrak{a}}(T) \geq \sqrt{|\mathcal{A}|T}$.

Adapting to Causal Structure

Theorem: Existing algorithms do not adapt to failure of assumptions.

For every $\mathcal{A}$ and $\mathcal{Z}$, there exists ν such that

$$\lim_{T \to \infty} \frac{R_{\nu,\text{C-UCB}}(T)}{T} \geq 1/120.$$

When can we use Z to learn counterfactuals?

When the choice of A doesn't affect the outcome of Y conditional on Z.

Definition (informal)

An environment ν is conditionally benign if and only if $\nu_a(Y \mid Z)$ is constant as a function of $a \in \mathcal{A}$.

Theorem: Strict adaptation to the conditionally benign property is impossible.

If $\mathfrak{a}$ always satisfies $R_{\nu,\mathfrak{a}}(T) \leq \sqrt{|\mathcal{A}|T}$,
there exists a conditionally benign ν such that $R_{\nu,\mathfrak{a}}(T) \geq \sqrt{|\mathcal{A}|T}$.

But it is still possible to achieve some adaptivity!

But it is still possible to achieve some adaptivity!

Main Result: Upper Bounds for HAC-UCB

Main Result: Upper Bounds for HAC-UCB

Previous work requires that we know $\nu(Z) = \{\nu_a(Z) : a \in \mathcal{A}\}$ in advance.

Main Result: Upper Bounds for HAC-UCB

Previous work requires that we know $\nu(Z) = \{\nu_a(Z) : a \in \mathcal{A}\}$ in advance.

Instead suppose that we have access to an estimate $\tilde{\nu}(Z)$ (learned offline).

Main Result: Upper Bounds for HAC-UCB

Previous work requires that we know $\nu(Z) = \{\nu_a(Z) : a \in \mathcal{A}\}$ in advance.
Instead suppose that we have access to an estimate $\tilde{\nu}(Z)$ (learned offline).

Theorem: Our new algorithm HAC-UCB achieves non-trivial adaptivity.

For any $\mathcal{A}$, $\mathcal{Z}$, T, ν, and $\tilde{\nu}(Z)$,

$$R_{\nu,\text{HAC-UCB}}(T) \leq \tilde{O}(T^{3/4}).$$

Further, if ν is conditionally benign and $\sup_{a\in\mathcal{A}} d_{\mathrm{TV}}(\tilde{\nu}_a(Z), \nu_a(Z)) \leq \varepsilon$,

$$R_{\nu,\text{HAC-UCB}}(T) \leq \tilde{O}\left(\sqrt{|\mathcal{Z}|T} + \varepsilon T\right).$$

Main Result: Upper Bounds for HAC-UCB

Previous work requires that we know $\nu(Z) = \{\nu_a(Z) : a \in \mathcal{A}\}$ in advance.
Instead suppose that we have access to an estimate $\tilde{\nu}(Z)$ (learned offline).

Theorem: Our new algorithm HAC-UCB achieves non-trivial adaptivity.

For any $\mathcal{A}$, $\mathcal{Z}$, T, ν, and $\tilde{\nu}(Z)$,

$$R_{\nu,\text{HAC-UCB}}(T) \leq \tilde{O}(T^{3/4}).$$

Further, if ν is conditionally benign and $\sup_{a\in\mathcal{A}} d_{\mathrm{TV}}(\tilde{\nu}_a(Z), \nu_a(Z)) \leq \varepsilon$,

$$R_{\nu,\text{HAC-UCB}}(T) \leq \tilde{O}\left(\sqrt{|\mathcal{Z}|T} + \varepsilon T\right).$$

Theorem

For any $\mathcal{A}$ and $\mathcal{Z}$, there exists ν such that

$$R_{\nu,\text{C-UCB}}(T) \geq \Omega(T) \quad \text{but} \quad R_{\nu,\text{HAC-UCB}}(T) \leq \tilde{O}(\sqrt{|\mathcal{A}|T}).$$

Main Result: Upper Bounds for HAC-UCB

Previous work requires that we know $\nu(Z) = \{\nu_a(Z) : a \in \mathcal{A}\}$ in advance.
Instead suppose that we have access to an estimate $\tilde{\nu}(Z)$ (learned offline).

Theorem: Our new algorithm HAC-UCB achieves non-trivial adaptivity.

For any $\mathcal{A}$, $\mathcal{Z}$, T, ν, and $\tilde{\nu}(Z)$,

$$R_{\nu,\text{HAC-UCB}}(T) \leq \tilde{O}(T^{3/4}).$$

Further, if ν is conditionally benign and $\sup_{a\in\mathcal{A}} d_{\mathrm{TV}}(\tilde{\nu}_a(Z), \nu_a(Z)) \leq \varepsilon$,

$$R_{\nu,\text{HAC-UCB}}(T) \leq \tilde{O}\left(\sqrt{|\mathcal{Z}|T} + \varepsilon T\right).$$

Theorem

For any $\mathcal{A}$ and $\mathcal{Z}$, there exists ν such that

$$R_{\nu,\text{C-UCB}}(T) \geq \Omega(T) \quad \text{but} \quad R_{\nu,\text{HAC-UCB}}(T) \leq \tilde{O}(\sqrt{|\mathcal{A}|T}).$$

Conclusions

- How can most efficiently learn the best intervention?
- Can we exploit benign structure when it exists while still learning consistent estimates when it doesn't?
- We formalize and generalize what structure makes the task easier.
- We show existing algorithms fail strongly when assumptions fail.
- We show optimal adaptivity is impossible and formalize the Pareto trade-off for conditionally benign.
- We introduce HAC-UCB, which has non-trivial adaptivity.
- Many wide-open and interesting problems to study in bandit adaptivity!

Conclusions

- How can most efficiently learn the best intervention?
- Can we exploit benign structure when it exists while still learning consistent estimates when it doesn't?
- We formalize and generalize what structure makes the task easier.
- We show existing algorithms fail strongly when assumptions fail.
- We show optimal adaptivity is impossible and formalize the Pareto trade-off for conditionally benign.
- We introduce HAC-UCB, which has non-trivial adaptivity.
- Many wide-open and interesting problems to study in bandit adaptivity!