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

November 15, 2022
The University of Chicago Rising Stars in Data Science Workshop
Motivation

**Goal:** Learn the intervention that has the largest positive causal effect on a variable of interest.
Motivation

**Goal:** Learn the intervention that has the largest positive causal effect on a variable of interest.

Impossible from observational data without unverifiable assumptions about the causal graph.
**Motivation**

**Goal:** Learn the intervention that has the largest positive causal effect on a variable of interest.

Impossible from observational data without unverifiable assumptions about the causal graph.

Instead, we can intervene...**but this is expensive.**
Motivation

**Goal:** Learn the intervention that has the largest positive causal effect on a variable of interest.

Impossible from observational data without unverifiable assumptions about the causal graph.

Instead, we can intervene...but this is expensive.

How can we most efficiently select which interventions to perform?
Motivation

**Goal:** Learn the intervention that has the largest positive causal effect on a variable of interest.

Impossible from observational data without unverifiable assumptions about the causal graph.

Instead, we can intervene...but this is expensive.

How can we most efficiently select which interventions to perform?

**Existing causal algorithms:**

Causal assumptions **hold** $\implies$ More efficient interventions!

Causal assumptions **fail** $\implies$ Learn biased estimates.
Motivation

**Goal:** Learn the intervention that has the largest positive causal effect on a variable of interest.

Impossible from observational data without unverifiable assumptions about the causal graph.

Instead, we can intervene...but this is expensive.

How can we most efficiently select which interventions to perform?

Existing causal algorithms:
Causal assumptions $\text{hold} \implies$ More efficient interventions!
Causal assumptions $\text{fail} \implies$ Learn biased estimates.

Our novel method:
Causal assumptions $\text{hold} \implies$ Optimally efficient interventions!
Causal assumptions $\text{fail} \implies$ Still learn causal effects!
Motivation

**Goal:** Learn the intervention that has the largest positive causal effect on a variable of interest.

Impossible from observational data without unverifiable assumptions about the causal graph.

Instead, we can intervene...but this is expensive.

How can we most efficiently select which interventions to perform?

Existing causal algorithms:
- Causal assumptions hold $\implies$ More efficient interventions!
- Causal assumptions fail $\implies$ Learn biased estimates.

Our novel method:
- Causal assumptions hold $\implies$ Optimally efficient interventions!
- Causal assumptions fail $\implies$ Still learn causal effects!

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

**Goal:** Learn the intervention that has the largest positive causal effect on a variable of interest.

Impossible from observational data without unverifiable assumptions about the causal graph.

Instead, we can intervene...but this is expensive.

How can we most efficiently select which interventions to perform?

**Existing causal algorithms:**

- Causal assumptions hold $\implies$ More efficient interventions!
- Causal assumptions fail $\implies$ Learn biased estimates.

**Our novel method:**

- Causal assumptions hold $\implies$ Optimally efficient interventions!
- Causal assumptions fail $\implies$ Still learn causal effects!

That is, our method *adapts* to the presence of causal structure.
Causal Inference with Interventions via Multi-Armed Bandits

Standard Multi-Armed Bandits

- Sequentially pick intervention $A_t \in \mathcal{A}$
- Observe reward $Y_t$
- Goal is to learn optimal intervention $\arg \max_{a \in \mathcal{A}} \mathbb{E}_a Y$

Without more structure, this can be necessarily inefficient.

In practice, we also observe other information when we take an intervention. Multi-Armed Bandits with Post-Action Contexts: Also observe $Z_t \in \mathcal{Z}$.

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

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

A policy $\pi$ maps the observed history $(A_1, Z_1, Y_1, ..., A_{t-1}, Z_{t-1}, Y_{t-1})$ to a distribution over $A_t$.

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

3/8
Causal Inference with Interventions via Multi-Armed Bandits

**Standard Multi-Armed Bandits**

- Sequentially pick intervention $A_t \in \mathcal{A}$
- Observe reward $Y_t$
- Goal is to learn optimal intervention $\arg \max_{a \in \mathcal{A}} \mathbb{E}_a Y$

*Without more structure, this can be necessarily inefficient.*
Causal Inference with Interventions via Multi-Armed Bandits

Standard Multi-Armed Bandits

- Sequentially pick intervention $A_t \in \mathcal{A}$
- Observe reward $Y_t$
- Goal is to learn optimal intervention $\arg \max_{a \in \mathcal{A}} \mathbb{E}_a Y$

Without more structure, this can be necessarily inefficient.

In practice, we also observe other information when we take an intervention.

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

A policy $\pi$ maps the observed history $(A_1, Z_1, Y_1, \ldots, A_{t-1}, Z_{t-1}, Y_{t-1})$ to a distribution over $A_t$. 

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

3/8
Causal Inference with Interventions via Multi-Armed Bandits

Standard Multi-Armed Bandits

- Sequentially pick intervention $A_t \in \mathcal{A}$
- Observe reward $Y_t$
- Goal is to learn optimal intervention $\arg\max_{a \in \mathcal{A}} \mathbb{E}_a Y$

Without more structure, this can be necessarily inefficient. In practice, we also observe other information when we take an intervention.

Multi-Armed Bandits with Post-Action Contexts: Also observe $Z_t \in \mathcal{Z}$.

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

Standard Multi-Armed Bandits
- Sequentially pick intervention $A_t \in \mathcal{A}$
- Observe reward $Y_t$
- Goal is to learn optimal intervention $\arg \max_{a \in \mathcal{A}} \mathbb{E}_a Y$

Without more structure, this can be necessarily inefficient.
In practice, we also observe other information when we take an intervention.

Multi-Armed Bandits with Post-Action Contexts: Also observe $Z_t \in \mathcal{Z}$.
We have no guarantees that observing $Z_t$ will help us...but we would like to exploit it when we can.

An environment $\nu$ is a fixed collection of distributions on $(\mathcal{Z}, \mathcal{Y})$: one for each $a \in \mathcal{A}$. 
Causal Inference with Interventions via Multi-Armed Bandits

Standard Multi-Armed Bandits

- Sequentially pick intervention $A_t \in \mathcal{A}$
- Observe reward $Y_t$
- Goal is to learn optimal intervention $\arg\max_{a \in \mathcal{A}} \mathbb{E}_a Y$

Without more structure, this can be necessarily inefficient. In practice, we also observe other information when we take an intervention.

Multi-Armed Bandits with Post-Action Contexts: Also observe $Z_t \in \mathcal{Z}$.

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

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

A policy $\pi$ maps the observed history $(A_1, Z_1, Y_1, \ldots, A_{t-1}, Z_{t-1}, Y_{t-1})$ to a distribution over $A_t$. 
Causal Inference with Interventions via Multi-Armed Bandits

Standard Multi-Armed Bandits

- Sequentially pick intervention $A_t \in \mathcal{A}$
- Observe reward $Y_t$
- Goal is to learn optimal intervention $\arg\max_{a \in \mathcal{A}} \mathbb{E}_a Y$

Without more structure, this can be necessarily inefficient.
In practice, we also observe other information when we take an intervention.

Multi-Armed Bandits with Post-Action Contexts: Also observe $Z_t \in \mathcal{Z}$.

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

An environment $\nu$ is a fixed collection of distributions on $(\mathcal{Z}, \mathcal{Y})$: one for each $a \in \mathcal{A}$.
A policy $\pi$ maps the observed history $(A_1, Z_1, Y_1, \ldots, A_{t-1}, Z_{t-1}, Y_{t-1})$ to a distribution over $A_t$.

Regret: $R_{\nu, \pi}(T) = T \cdot \max_{a \in \mathcal{A}} \mathbb{E}_{\nu_a}[Y] - \mathbb{E}_{\nu, \pi}[\sum_{t=1}^T Y_t]$. 
Causal Inference with Interventions via Multi-Armed Bandits

**Standard Multi-Armed Bandits**

- Sequentially pick intervention \( A_t \in \mathcal{A} \)
- Observe reward \( Y_t \)
- Goal is to learn optimal intervention \( \arg \max_{a \in \mathcal{A}} \mathbb{E}_a Y \)

Without more structure, this can be necessarily inefficient.

In practice, we also observe other information when we take an intervention.

**Multi-Armed Bandits with Post-Action Contexts:** Also observe \( Z_t \in \mathcal{Z} \).

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

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

A **policy** \( \pi \) maps the observed history \((A_1, Z_1, Y_1, \ldots, A_{t-1}, Z_{t-1}, Y_{t-1})\) to a distribution over \( A_t \).

**Regret:**

\[
R_{\nu,\pi}(T) = T \cdot \max_{a \in \mathcal{A}} \mathbb{E}_{\nu_a} [Y] - \mathbb{E}_{\nu,\pi} \left[ \sum_{t=1}^{T} Y_t \right].
\]
We formalize when $Z_t$ is helpful: conditionally benign environments. Existing causal algorithms have regret depending on $|Z|$ instead of $|A|$. Existing algorithms can have regret linear in $T$ in the worst case. This means they don't even have a consistent estimate of the causal effect!

We formalize adaptive minimax optimality for the conditionally benign property. Optimality is impossible: efficient interventions necessarily sacrifice worst-case performance.

We provide a new algorithm with:

a) optimal performance for conditionally benign environments and

b) sublinear regret (always learns causal effects).
We formalize when $Z_t$ is helpful: conditionally benign environments.
The Punchline: High-Level Overview of Results

We formalize when $Z_t$ is helpful: conditionally benign environments. Existing causal algorithms have regret depending on $|Z|$ instead of $|A|$. Existing algorithms can have regret linear in $T$ in the worst case. This means they don't even have a consistent estimate of the causal effect! We formalize adaptive minimax optimality for the conditionally benign property. Optimality is impossible: efficient interventions necessarily sacrifice worst-case performance. We provide a new algorithm with:

a) optimal performance for conditionally benign environments
b) sublinear regret (always learns causal effects).
We formalize when $Z_t$ is helpful: **conditionally benign environments**. Existing causal algorithms have regret depending on $|Z|$ instead of $|A|$.

Existing algorithms can have regret linear in $T$ in the worst case. *This means they don’t even have a consistent estimate of the causal effect!*
The Punchline: High-Level Overview of Results

We formalize when $Z_t$ is helpful: **conditionally benign environments.**
Existing causal algorithms have regret depending on $|Z|$ instead of $|A|$.

**Existing algorithms can have regret linear in $T$ in the worst case.**
This means they don’t even have a consistent estimate of the causal effect!

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

**Optimality is impossible:** efficient interventions necessarily sacrifice worst-case performance.
The Punchline: High-Level Overview of Results

We formalize when $Z_t$ is helpful: **conditionally benign environments**. Existing causal algorithms have regret depending on $|Z|$ instead of $|A|$.

**Existing algorithms can have regret linear in $T$ in the worst case.**
This means they don’t even have a consistent estimate of the causal effect!

We formalize *adaptive minimax optimality* for the **conditionally benign property**.
**Optimality is impossible: efficient interventions necessarily sacrifice worst-case performance.**

We provide a new algorithm with:
   a) **optimal performance for conditionally benign environments** and
   b) **sublinear regret** (always learns causal effects).
We formalize when $Z_t$ is helpful: conditionally benign environments. Existing causal algorithms have regret depending on $|Z|$ instead of $|A|$. Existing algorithms can have regret linear in $T$ in the worst case. This means they don’t even have a consistent estimate of the causal effect!

We formalize adaptive minimax optimality for the conditionally benign property. Optimality is impossible: efficient interventions necessarily sacrifice worst-case performance.

We provide a new algorithm with:

a) optimal performance for conditionally benign environments and
b) sublinear regret (always learns causal effects).
Impossibility Result

Without any assumptions beyond IID, UCB (Auer at al. 2002):

\[ R_{\nu, UCB}(T) = \tilde{\Theta}(\sqrt{|A|T}) \]

Definition (informal)

An environment \( \nu \) is conditionally benign if and only if \( \nu(a | Y | Z) \) is constant as a function of \( a \in A \).

When the environment \( \nu \) is conditionally benign and the marginal distributions \( \nu(Z) \) are known, C-UCB (Lu et al. 2020; BWR Thm 4.3):

\[ R_{\nu, C-UCB}(T) = \tilde{\Theta}(\sqrt{|Z|T}) \]

But in the worst case,

\[ R_{\nu, C-UCB}(T) \geq \Omega(T) \]

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

If \( \pi \) is such that \( R_{\nu, \pi}(T) \leq O(\sqrt{|A|T}) \) for all \( \nu \), there exists \( \nu \) that is conditionally benign but \( R_{\nu, \pi}(T) \geq \Omega(\sqrt{|A|T}) \).

Can we adapt at all?
Without any assumptions beyond IID, 

**UCB (Auer at al. 2002):** 

\[ R_{\nu,\text{UCB}}(T) = \tilde{\Theta}(\sqrt{|A|T}) \]
Impossibility Result

Without any assumptions beyond IID,

**UCB (Auer et al. 2002):** $R_{\nu,\text{UCB}}(T) = \tilde{\Theta}(\sqrt{|A|T})$

**Definition (informal)**

An environment $\nu$ is conditionally benign if and only if $\nu_a(Y | Z)$ is constant as a function of $a \in A$. 

Can we adapt at all?
Without any assumptions beyond IID,

**UCB (Auer at al. 2002):** \( R_{\nu,\text{UCB}}(T) = \tilde{\Theta}(\sqrt{|A|T}) \)

**Definition (informal)**

An environment \( \nu \) is conditionally benign if and only if \( \nu_a(Y \mid Z) \) is constant as a function of \( a \in A \).

When the environment \( \nu \) is conditionally benign and the marginal distributions \( \nu(Z) \) are known,

**C-UCB (Lu et al. 2020; BWR Thm 4.3):** \( R_{\nu,\text{C-UCB}}(T) = \tilde{\Theta}(\sqrt{|Z|T}) \)

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

If \( \pi \) is such that \( R_{\nu,\pi}(T) \leq O(\sqrt{|A|T}) \) for all \( \nu \), there exists \( \nu \) that is conditionally benign but \( R_{\nu,\pi}(T) \geq \Omega(\sqrt{|A|T}) \).

Can we adapt at all?
Without any assumptions beyond IID, 

**UCB (Auer et al. 2002):** \( R_{\nu,\text{UCB}}(T) = \tilde{\Theta}(\sqrt{|A|T}) \)

**Definition (informal):**

An environment \( \nu \) is conditionally benign if and only if \( \nu_a(Y \mid Z) \) is constant as a function of \( a \in A \).

When the environment \( \nu \) is conditionally benign and the marginal distributions \( \nu(Z) \) are known, 

**C-UCB (Lu et al. 2020; BWR Thm 4.3):** \( R_{\nu,\text{C-UCB}}(T) = \tilde{\Theta}(\sqrt{|Z|T}) \)

But in the worst case, \( R_{\nu,\text{C-UCB}}(T) \geq \Omega(T) \)
Impossibility Result

Without any assumptions beyond IID,

**UCB (Auer et al. 2002):** \( R_{\nu,\text{UCB}}(T) = \tilde{\Theta}(\sqrt{|A|T}) \)

**Definition (informal)**

An environment \( \nu \) is conditionally benign if and only if \( \nu_a(Y | Z) \) is constant as a function of \( a \in A \).

When the environment \( \nu \) is conditionally benign and the marginal distributions \( \nu(Z) \) are known,

**C-UCB (Lu et al. 2020; BWR Thm 4.3):** \( R_{\nu,\text{C-UCB}}(T) = \tilde{\Theta}(\sqrt{|Z|T}) \)

But in the worst case, \( R_{\nu,\text{C-UCB}}(T) \geq \Omega(T) \)

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

If \( \pi \) is such that \( R_{\nu,\pi}(T) \leq O(\sqrt{|A|T}) \) for all \( \nu \),

there exists \( \nu \) that is conditionally benign but \( R_{\nu,\pi}(T) \geq \Omega(\sqrt{|A|T}) \).
Impossibility Result

Without any assumptions beyond IID, 

\[ R_{\nu, \text{UCB}}(T) = \tilde{\Theta}(\sqrt{|A|T}) \]

**Definition (informal)**

An environment \( \nu \) is conditionally benign if and only if \( \nu_a(Y \mid Z) \) is constant as a function of \( a \in A \).

When the environment \( \nu \) is conditionally benign and the marginal distributions \( \nu(Z) \) are known, 

\[ R_{\nu, \text{C-UCB}}(T) = \tilde{\Theta}(\sqrt{|Z|T}) \]

But in the worst case, \( R_{\nu, \text{C-UCB}}(T) \geq \Omega(T) \)

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

If \( \pi \) is such that \( R_{\nu, \pi}(T) \leq O(\sqrt{|A|T}) \) for all \( \nu \),

there exists \( \nu \) that is conditionally benign but \( R_{\nu, \pi}(T) \geq \Omega(\sqrt{|A|T}) \).

Can we adapt at all?
Impossibility Result

Without any assumptions beyond IID,

**UCB (Auer at al. 2002):**

\[ R_{\nu,\text{UCB}}(T) = \tilde{\Theta}(\sqrt{|A|T}) \]

**Definition (informal):**

An environment \( \nu \) is conditionally benign if and only if \( \nu_a(Y \mid Z) \) is constant as a function of \( a \in A \).

When the environment \( \nu \) is conditionally benign and the marginal distributions \( \nu(Z) \) are known,

**C-UCB (Lu et al. 2020; BWR Thm 4.3):**

\[ R_{\nu,\text{C-UCB}}(T) = \tilde{\Theta}(\sqrt{|Z|T}) \]

But in the worst case, \( R_{\nu,\text{C-UCB}}(T) \geq \Omega(T) \)

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

If \( \pi \) is such that \( R_{\nu,\pi}(T) \leq O(\sqrt{|A|T}) \) for all \( \nu \),

there exists \( \nu \) that is conditionally benign but \( R_{\nu,\pi}(T) \geq \Omega(\sqrt{|A|T}) \).

Can we adapt at all?
Previous work requires that we know $\nu(Z) = \{\nu_a(Z) : a \in A\}$ in advance.

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

Hypothesis-Tested Adaptive C-UCB (HAC-UCB)

1. Optimistically suppose environment is conditionally benign and play C-UCB.
2. On each round, perform a hypothesis test for whether to switch to UCB.

We don't have to accurately identify failure of conditionally benign...

...just when that failure causes bad decision making.

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

For any $A$, $Z$, $T$, $\nu$, and $\tilde{\nu}$, $R_{\nu}^{HAC-UCB}(T) \leq \tilde{O}(T^{3/4})$.

Further, if $\nu$ is conditionally benign and $\|\nu(Z) - \tilde{\nu}(Z)\| \leq \varepsilon$, $R_{\nu}^{HAC-UCB}(T) \leq \tilde{O}(\sqrt{|Z|T} + \varepsilon T)$.
Previous work requires that we know $\nu(Z) = \{\nu_a(Z) : a \in A\}$ in advance.
Adaptive Results

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

Hypothesis-Tested Adaptive C-UCB (HAC-UCB)

1. Optimistically suppose environment is conditionally benign and play C-UCB.
2. On each round, perform a hypothesis test for whether to switch to UCB.

We don't have to accurately identify failure of conditionally benign... just when that failure causes bad decision making.

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

For any $A$, $Z$, $T$, $\nu$, and $\tilde{\nu}$, $R_{\nu, HAC-UCB}(T) \leq \tilde{O}(T^{3/4})$.

Further, if $\nu$ is conditionally benign and $\|\nu(Z) - \tilde{\nu}(Z)\| \leq \varepsilon$, $R_{\nu, HAC-UCB}(T) \leq \tilde{O}(\sqrt{|Z|T} + \varepsilon T)$.
Adaptive Results

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

**Hypothesis-Tested Adaptive C-UCB (HAC-UCB)**

1. Optimistically suppose environment is conditionally benign and play C-UCB.
2. On each round, perform a hypothesis test for whether to switch to UCB.

We don't have to accurately identify failure of conditionally benign... just when that failure causes bad decision making.

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

For any $A$, $Z$, $T$, $\nu$, and $\tilde{\nu}$, $R_{\nu}, HAC-UCB(T) \leq \tilde{O}(T^{3/4})$.

Further, if $\nu$ is conditionally benign and $\|\nu(Z) - \tilde{\nu}(Z)\| \leq \epsilon$, $R_{\nu}, HAC-UCB(T) \leq \tilde{O}(\sqrt{|Z|T} + \epsilon T)$. 
Adaptive Results

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

**Hypothesis-Tested Adaptive C-UCB (HAC-UCB)**

1. Optimistically suppose environment is conditionally benign and play C-UCB.
Adaptive Results

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

**Hypothesis-Tested Adaptive C-UCB (HAC-UCB)**

1. Optimistically suppose environment is conditionally benign and play C-UCB.
2. On each round, perform a hypothesis test for whether to switch to UCB.
Adaptive Results

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

**Hypothesis-Tested Adaptive C-UCB (HAC-UCB)**

1. Optimistically suppose environment is conditionally benign and play C-UCB.
2. On each round, perform a hypothesis test for whether to switch to UCB.

   We don't have to accurately identify failure of conditionally benign...
   ...just when that failure causes bad decision making.

Main Theorem: Our new algorithm HAC-UCB achieves non-trivial adaptivity. For any $A$, $Z$, $T$, $\nu$, and $\tilde{\nu}$, $R_\nu, HAC-UCB(T) \leq \tilde{O}(T^{3/4})$.

Further, if $\nu$ is conditionally benign and $\|\nu(Z) - \tilde{\nu}(Z)\| \leq \varepsilon$, $R_\nu, HAC-UCB(T) \leq \tilde{O}(\sqrt{|Z|T + \varepsilon T})$. 


Adaptive Results

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 \( \tilde{\nu}(Z) \).

**Hypothesis-Tested Adaptive C-UCB (HAC-UCB)**

1. Optimistically suppose environment is conditionally benign and play C-UCB.
2. On each round, perform a hypothesis test for whether to switch to UCB.

   We don't have to accurately identify failure of conditionally benign...

   ...just when that failure causes bad decision making.

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

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

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

Further, if \( \nu \) is conditionally benign and \( \| \nu(Z) - \tilde{\nu}(Z) \| \leq \varepsilon \),

\[
R_{\nu,\text{HAC-UCB}}(T) \leq \tilde{O}(\sqrt{|\mathcal{Z}|T + \varepsilon T}).
\]
Adaptive Results

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 \( \tilde{\nu}(Z) \).

Hypothesis-Tested Adaptive C-UCB (HAC-UCB)

1. Optimistically suppose environment is conditionally benign and play C-UCB.
2. On each round, perform a hypothesis test for whether to switch to UCB.
   
   We don't have to accurately identify failure of conditionally benign...
   ...just when that failure causes bad decision making.

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

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

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

Further, if \( \nu \) is conditionally benign and \( \| \nu(Z) - \tilde{\nu}(Z) \| \leq \varepsilon \),

\[
R_{\nu,HAC-UCB}(T) \leq \tilde{O}(\sqrt{|\mathcal{Z}|T + \varepsilon T}).
\]
Simulation Results

Conditionally Benign Environment (|A|=20, |Z|=2)

Regret(T) vs Time (T)

- UCB
- Corral
- C_UCB
- C_UCB_2
- HAC_UCB*

Graph showing the relationship between regret and time for different algorithms in a conditionally benign environment.
Simulation Results

Worst Case Environment (|A| = 20, |Z| = 2)

![Chart showing regret over time for different algorithms with legend: UCB, Corral, C_UCB, C_UCB_2, HAC_UCB*]
Pareto Frontier of Causal Bandits

Worst-case optimal: UCB (Auer at al. 2002), Conditionally benign optimal: C-UCB (Lu et al. 2020)

New algorithm: HAC-UCB (this work)

Conditionally Benign Regret

Worst-Case Regret

\[ \sqrt{|Z|T} \quad \sqrt{|A|T} \]

\[ T^{3/4} \]

\[ T \]
Pareto Frontier of Causal Bandits

Worst-case optimal: UCB (Auer at al. 2002), Conditionally benign optimal: C-UCB (Lu et al. 2020)

New algorithm: HAC-UCB (this work)

Conditionally Benign Regret

\( \sqrt{|Z|T} \)

Worst-Case Regret

\( \sqrt{|A|T} \)
Pareto Frontier of Causal Bandits

Worst-case optimal: UCB (Auer at al. 2002), Conditionally benign optimal: C-UCB (Lu et al. 2020)

New algorithm: HAC-UCB (this work)

Conditionally Benign Regret

\[ \sqrt{|Z|T} \]

Worst-Case Regret

\[ \sqrt{|A|T} \quad T^{3/4} \quad T \]

(Auer et al. 2002, Lu et al. 2020)

(Thms 4.3 and 4.5)
Pareto Frontier of Causal Bandits

Worst-case optimal: UCB (Auer et al. 2002), Conditionally benign optimal: C-UCB (Lu et al. 2020)

New algorithm: HAC-UCB (this work)

Conditionally Benign Regret

\[
\sqrt{|Z| T}
\]

\[
T^{3/4}
\]

Worst-Case Regret

\[
\sqrt{|A| T}
\]

\[
\sqrt{|Z| T}
\]
Worst-case optimal: UCB (Auer at al. 2002), Conditionally benign optimal: C-UCB (Lu et al. 2020)

New algorithm: HAC-UCB (this work)
Understanding UCB and C-UCB

Upper Confidence Bound (UCB) Algorithm:
• Maintain empirical mean estimate $\hat{\mu}_t(a)$ for each $t \in [T]$ and $a \in A$
• Use concentration inequality to construct confidence bound $\text{UCB}_t(a) = \hat{\mu}_t(a) + \sqrt{\log(T)/N_t(a)}$
• Play $A_t = \text{arg max}_{a \in A} \text{UCB}_t(a)$

Causal Upper Confidence Bound (C-UCB) Algorithm:
• Maintain empirical mean estimate $\hat{\mu}_t(z)$ for each $t \in [T]$ and $z \in Z$
• Use concentration inequality to construct confidence bound $\text{UCB}_t(z) = \hat{\mu}_t(z) + \sqrt{\log(T)/N_t(z)}$
• Play $A_t = \text{arg max}_{a \in A} \sum_{z \in Z} \text{UCB}_t(z) P_{\tilde{\nu}_a[Z=z]}$

Why does this work?
If all parents are observed (more generally, $\nu$ is conditionally benign) and $\tilde{\nu}_a[Z=z]$ is accurate, $\sum_{z \in Z} \text{UCB}_t(z) P_{\tilde{\nu}_a[Z=z]} \approx \text{UCB}_t(a)$, but concentration only requires a union bound of size $|Z|$ instead of size $|A|$. 
Understanding UCB and C-UCB

Upper Confidence Bound (UCB) Algorithm:

• Maintain empirical mean estimate $\hat{\mu}_t(a)$ for each $t \in [T]$ and $a \in A$
• Use concentration inequality to construct confidence bound $UCB_t(a) = \hat{\mu}_t(a) + \sqrt{\log(T)/N_t(a)}$
• Play $A_t = \arg \max_{a \in A} UCB_t(a)$

Causal Upper Confidence Bound (C-UCB) Algorithm:

• Maintain empirical mean estimate $\hat{\mu}_t(z)$ for each $t \in [T]$ and $z \in Z$
• Use concentration inequality to construct confidence bound $UCB_t(z) = \hat{\mu}_t(z) + \sqrt{\log(T)/N_t(z)}$
• Play $A_t = \arg \max_{a \in A} \sum_{z \in Z} UCB_t(z) P_{\tilde{\nu}[Z = z]}$

Why does this work? If all parents are observed (more generally, $\nu$ is conditionally benign) and $\tilde{\nu}(Z)$ is accurate, $\sum_{z \in Z} UCB_t(z) P_{\tilde{\nu}[Z = z]} \approx UCB_t(a)$, but concentration only requires a union bound of size $|Z|$ instead of size $|A|$.
Understanding UCB and C-UCB

Upper Confidence Bound (UCB) Algorithm:
- Maintain empirical mean estimate $\hat{\mu}_t(a)$ for each $t \in [T]$ and $a \in A$

Causal Upper Confidence Bound (C-UCB) Algorithm:
- Maintain empirical mean estimate $\hat{\mu}_t(z)$ for each $t \in [T]$ and $z \in Z$
- Use concentration inequality to construct confidence bound $UCB_t(z) = \hat{\mu}_t(z) + \sqrt{\log(T)/N_t(z)}$
- Play $A_t = \arg\max_{a \in A} \sum_{z \in Z} UCB_t(z) P_{\tilde{\nu}}[Z = z]$
Upper Confidence Bound (UCB) Algorithm:
- Maintain empirical mean estimate $\hat{\mu}_t(a)$ for each $t \in [T]$ and $a \in A$
- Use concentration inequality to construct confidence bound $UCB_t(a) = \hat{\mu}_t(a) + \sqrt{\log(T)/N_t(a)}$

Causal Upper Confidence Bound (C-UCB) Algorithm:
- Maintain empirical mean estimate $\hat{\mu}_t(z)$ for each $t \in [T]$ and $z \in Z$
- Use concentration inequality to construct confidence bound $UCB_t(z) = \hat{\mu}_t(z) + \sqrt{\log(T)/N_t(z)}$
- Play $A_t = \arg \max_a \sum_{z \in Z} UCB_t(z) P_{\tilde{\nu}}[Z = z]$
Understanding UCB and C-UCB

Upper Confidence Bound (UCB) Algorithm:
- Maintain empirical mean estimate $\hat{\mu}_t(a)$ for each $t \in [T]$ and $a \in A$
- Use concentration inequality to construct confidence bound $\text{UCB}_t(a) = \hat{\mu}_t(a) + \sqrt{\frac{\log(T)}{N_t(a)}}$
- Play $A_t = \arg\max_{a \in A} \text{UCB}_t(a)$

Causal Upper Confidence Bound (C-UCB) Algorithm:
- Maintain empirical mean estimate $\hat{\mu}_t(z)$ for each $t \in [T]$ and $z \in Z$
- Use concentration inequality to construct confidence bound $\text{UCB}_t(z) = \hat{\mu}_t(z) + \sqrt{\frac{\log(T)}{N_t(z)}}$
- Play $A_t = \arg\max_{a \in A} \sum_{z \in Z} \text{UCB}_t(z) \mathbb{P}_{\tilde{\nu}}[Z = z]$
Upper Confidence Bound (UCB) Algorithm:
- Maintain empirical mean estimate $\hat{\mu}_t(a)$ for each $t \in [T]$ and $a \in A$
- Use concentration inequality to construct confidence bound $UCB_t(a) = \hat{\mu}_t(a) + \sqrt{\frac{\log(T)}{N_t(a)}}$
- Play $A_t = \arg\max_{a \in A} UCB_t(a)$

Causal Upper Confidence Bound (C-UCB) Algorithm: 
- Maintain empirical mean estimate $\hat{\mu}_t(z)$ for each $t \in [T]$ and $z \in Z$
- Use concentration inequality to construct confidence bound $UCB_t(z) = \hat{\mu}_t(z) + \sqrt{\frac{\log(T)}{N_t(z)}}$
- Play $A_t = \arg\max_{a \in A} \sum_{z \in Z} UCB_t(z) P_{\tilde{\nu}}(Z = z)$

Why does this work? If all parents are observed (more generally, $\nu$ is conditionally benign) and $\tilde{\nu}(Z)$ is accurate, $\sum_{z \in Z} UCB_t(z) P_{\tilde{\nu}}(Z = z) \approx UCB_t(a)$, but concentration only requires a union bound of size $|Z|$ instead of size $|A|$. 
Understanding UCB and C-UCB

Upper Confidence Bound (UCB) Algorithm:
- Maintain empirical mean estimate $\hat{\mu}_t(a)$ for each $t \in [T]$ and $a \in A$
- Use concentration inequality to construct confidence bound $UCB_t(a) = \hat{\mu}_t(a) + \sqrt{\log(T)/N_t(a)}$
- Play $A_t = \arg\max_{a \in A} UCB_t(a)$

Causal Upper Confidence Bound (C-UCB) Algorithm:
- Maintain empirical mean estimate $\hat{\mu}_t(z)$ for each $t \in [T]$ and $z \in Z$
Understanding UCB and C-UCB

Upper Confidence Bound (UCB) Algorithm:

- Maintain empirical mean estimate $\hat{\mu}_t(a)$ for each $t \in [T]$ and $a \in A$
- Use concentration inequality to construct confidence bound $UCB_t(a) = \hat{\mu}_t(a) + \frac{\sqrt{\log(T) / N_t(a)}}{\sqrt{T}}$
- Play $A_t = \arg \max_{a \in A} UCB_t(a)$

Causal Upper Confidence Bound (C-UCB) Algorithm:

- Maintain empirical mean estimate $\hat{\mu}_t(z)$ for each $t \in [T]$ and $z \in Z$
- Use concentration inequality to construct confidence bound $UCB_t(z) = \hat{\mu}_t(z) + \frac{\sqrt{\log(T) / N_t(z)}}{\sqrt{T}}$

Why does this work?

If all parents are observed (more generally, $\nu$ is conditionally benign) and $\tilde{\nu}$ is accurate, $\sum_{z \in Z} UCB_t(z) P_{\tilde{\nu}}[Z = z] \approx UCB_t(a)$, but concentration only requires a union bound of size $|Z|$ instead of size $|A|$. 
Understanding UCB and C-UCB

Upper Confidence Bound (UCB) Algorithm:
- Maintain empirical mean estimate $\hat{\mu}_t(a)$ for each $t \in [T]$ and $a \in \mathcal{A}$
- Use concentration inequality to construct confidence bound $UCB_t(a) = \hat{\mu}_t(a) + \sqrt{\log(T)/N_t(a)}$
- Play $A_t = \arg \max_{a \in \mathcal{A}} UCB_t(a)$

Causal Upper Confidence Bound (C-UCB) Algorithm:
- Maintain empirical mean estimate $\hat{\mu}_t(z)$ for each $t \in [T]$ and $z \in \mathcal{Z}$
- Use concentration inequality to construct confidence bound $UCB_t(z) = \hat{\mu}_t(z) + \sqrt{\log(T)/N_t(z)}$
- Play $A_t = \arg \max_{a \in \mathcal{A}} \sum_{z \in \mathcal{Z}} UCB_t(z) P_{\nu_a}[Z = z]$

Why does this work?
If all parents are observed (more generally, $\nu$ is conditionally benign) and $\tilde{\nu}(Z)$ is accurate, $\sum_{z \in \mathcal{Z}} UCB_t(z) P_{\nu_a}[Z = z] \approx UCB_t(a)$, but concentration only requires a union bound of size $|\mathcal{Z}|$ instead of size $|\mathcal{A}|$. 
Upper Confidence Bound (UCB) Algorithm:
- Maintain empirical mean estimate $\hat{\mu}_t(a)$ for each $t \in [T]$ and $a \in A$
- Use concentration inequality to construct confidence bound $UCB_t(a) = \hat{\mu}_t(a) + \sqrt{\log(T)/N_t(a)}$
- Play $A_t = \arg\max_{a \in A} UCB_t(a)$

Causal Upper Confidence Bound (C-UCB) Algorithm:
- Maintain empirical mean estimate $\hat{\mu}_t(z)$ for each $t \in [T]$ and $z \in Z$
- Use concentration inequality to construct confidence bound $UCB_t(z) = \hat{\mu}_t(z) + \sqrt{\log(T)/N_t(z)}$
- Play $A_t = \arg\max_{a \in A} \sum_{z \in Z} UCB_t(z)P_{\tilde{\nu}_a}[Z = z]$

Why does this work?
Understanding UCB and C-UCB

Upper Confidence Bound (UCB) Algorithm:
- Maintain empirical mean estimate $\hat{\mu}_t(a)$ for each $t \in [T]$ and $a \in \mathcal{A}$
- Use concentration inequality to construct confidence bound $UCB_t(a) = \hat{\mu}_t(a) + \sqrt{\log(T)/N_t(a)}$
- Play $A_t = \arg \max_{a \in \mathcal{A}} UCB_t(a)$

Causal Upper Confidence Bound (C-UCB) Algorithm:
- Maintain empirical mean estimate $\hat{\mu}_t(z)$ for each $t \in [T]$ and $z \in \mathcal{Z}$
- Use concentration inequality to construct confidence bound $UCB_t(z) = \hat{\mu}_t(z) + \sqrt{\log(T)/N_t(z)}$
- Play $A_t = \arg \max_{a \in \mathcal{A}} \sum_{z \in \mathcal{Z}} UCB_t(z) \mathbb{P}_{\tilde{\nu}_a}[Z = z]$

Why does this work?
If all parents are observed (more generally, $\nu$ is conditionally benign) and $\tilde{\nu}(Z)$ is accurate,

$$\sum_{z \in \mathcal{Z}} UCB_t(z) \mathbb{P}_{\tilde{\nu}_a}[Z = z] \approx UCB_t(a),$$

but concentration only requires a union bound of size $|\mathcal{Z}|$ instead of size $|\mathcal{A}|$. 

9/8
Understanding UCB and C-UCB

Upper Confidence Bound (UCB) Algorithm:
• Maintain empirical mean estimate $\hat{\mu}_t(a)$ for each $t \in [T]$ and $a \in A$
• Use concentration inequality to construct confidence bound $UCB_t(a) = \hat{\mu}_t(a) + \sqrt{\log(T)/N_t(a)}$
• Play $A_t = \arg\max_{a \in A} UCB_t(a)$

Causal Upper Confidence Bound (C-UCB) Algorithm:
• Maintain empirical mean estimate $\hat{\mu}_t(z)$ for each $t \in [T]$ and $z \in Z$
• Use concentration inequality to construct confidence bound $UCB_t(z) = \hat{\mu}_t(z) + \sqrt{\log(T)/N_t(z)}$
• Play $A_t = \arg\max_{a \in A} \sum_{z \in Z} UCB_t(z) \mathbb{P}_{\tilde{\nu}_a}[Z = z]$

Why does this work?
If all parents are observed (more generally, $\nu$ is conditionally benign) and $\tilde{\nu}(Z)$ is accurate,

$$\sum_{z \in Z} UCB_t(z) \mathbb{P}_{\tilde{\nu}_a}[Z = z] \approx UCB_t(a),$$

but concentration only requires a union bound of size $|Z|$ instead of size $|A|$.
Adapting with Hypothesis Testing: HAC-UCB

**Intuition:**
Optimistically play C-UCB until a hypothesis test for conditionally benign fails, then play UCB.

1. **Initial Exploration**
   Uniformly sample $a \in A$ for $\sqrt{T/|A|}$ rounds.

   Compute MLE estimate $\hat{\nu}^a$ of $(\nu^a(Z))_{a \in A}$. If $\sup_{a \in A} \| \tilde{\nu}^a - \hat{\nu}^a \|_1 \gtrsim T^{-1/4}$, set $\tilde{\nu}^a \leftarrow \hat{\nu}^a$.

2. **Optimistic Phase:**
   For each round $t$, play
   
   $UCB_t(a) \approx \hat{E}_{\nu^a} [Y] + \sqrt{\log T/\nu^a(t)}$.

   $\tilde{UCB}_t(a) \approx \sum_{z \in Z} \hat{E}_{\nu^a(Z = z)} [Y] + \sqrt{\log T/\nu^a(Z = z)}$.

   If $UCB_t(a) \approx \tilde{UCB}_t(a)$, play $A_{t+1} = \arg\max_{a \in A} \tilde{UCB}_t(a)$.

   Otherwise, switch to **Pessimistic Phase**.

3. **Pessimistic Phase:**
   For remaining rounds $t$, play $A_{t+1} = \arg\max_{a \in A} UCB_t(a)$. 


Adapting with Hypothesis Testing: HAC-UCB

**Intuition:** Optimistically play C-UCB until a hypothesis test for conditionally benign fails, then play UCB.

---

1. **Initial Exploration**
   - Uniformly sample $a \in A$ for $\sqrt{\frac{T}{|A|}}$ rounds.
   - Compute MLE estimate $\hat{\nu}_a$ of $(\nu_a(Z))_{a \in A}$.
     - If $\sup_{a \in A} \| \tilde{\nu}_a - \hat{\nu}_a \|_1 \gtrsim \frac{T^{1/2}}{4}$, set $\tilde{\nu}_a \leftarrow \hat{\nu}_a$.

2. **Optimistic Phase:**
   - For each round $t$...
     - $\text{UCB}_t(a) \approx \hat{E}_\nu_a[Y] + \sqrt{\frac{\log T}{N_a(t)}}$.
     - $\tilde{\text{UCB}}_t(a) \approx \sum_{z \in Z} \hat{E}_{\nu_a(Z = z)}[Y] + \sqrt{\frac{\log T}{N_z(t)}} \tilde{\nu}_a(Z = z)$.
   - If $\text{UCB}_t(a) \approx \tilde{\text{UCB}}_t(a)$, play $A_{t+1} = \arg \max_{a \in A} \tilde{\text{UCB}}_t(a)$.
   - Otherwise, switch to Pessimistic Phase.

3. **Pessimistic Phase:**
   - For remaining rounds $t$,
     - Play $A_{t+1} = \arg \max_{a \in A} \text{UCB}_t(a)$.
Adapting with Hypothesis Testing: HAC-UCB

Intuition: Optimistically play C-UCB until a hypothesis test for conditionally benign fails, then play UCB.

(1) Initial Exploration
Adapting with Hypothesis Testing: HAC-UCB

**Intuition:** Optimistically play C-UCB until a hypothesis test for conditionally benign fails, then play UCB.

(1) **Initial Exploration**
Uniformly sample $a \in A$ for $\sqrt{T}/|A|$ rounds.
Compute MLE estimate $\hat{\nu}$ of $(\nu_a(Z))_{a \in A}$. If $\sup_{a \in A} \| \tilde{\nu}_a - \hat{\nu}_a \|_1 \gtrsim T^{-1/4}$, set $\tilde{\nu} \leftarrow \hat{\nu}$. 
Adapting with Hypothesis Testing: HAC-UCB

**Intuition:** Optimistically play C-UCB until a hypothesis test for conditionally benign fails, then play UCB.

**(1) Initial Exploration**
Uniformly sample $a \in \mathcal{A}$ for $\sqrt{T}/|\mathcal{A}|$ rounds.
Compute MLE estimate $\hat{\nu}$ of $(\nu_a(Z))_{a \in \mathcal{A}}$. If $\sup_{a \in \mathcal{A}} \|\tilde{\nu}_a - \hat{\nu}_a\|_1 \gtrsim T^{-1/4}$, set $\tilde{\nu} \leftarrow \hat{\nu}$.

**Optimistic Phase:** For each round $t$...
\[
\text{UCB}_t(a) \approx \tilde{E}_{\nu_a}[Y] + \sqrt{\log T}/N_a(t).
\]
\[
\text{UCB}_t(a) \approx \sum_{z \in \mathcal{Z}} \tilde{E}_{\nu}[Y \mid Z = z] + \sqrt{\log T}/N_z(t)\tilde{\nu}_a(Z = z).
\]
If $\text{UCB}_t(a) \approx \text{UCB}_t(a)$, play $A_{t+1} = \arg \max_{a \in \mathcal{A}} \text{UCB}_t(a)$.
Otherwise, switch to Pessimistic Phase.
Adapting with Hypothesis Testing: HAC-UCB

**Intuition:** Optimistically play C-UCB until a hypothesis test for conditionally benign fails, then play UCB.

(1) **Initial Exploration**
Uniformly sample \( a \in \mathcal{A} \) for \( \sqrt{T}/|\mathcal{A}| \) rounds.
Compute MLE estimate \( \hat{\nu} \) of \( \nu_a(Z) \) for all \( a \in \mathcal{A} \). If \( \sup_{a \in \mathcal{A}} \| \tilde{\nu}_a - \hat{\nu}_a \|_1 \gtrsim T^{-1/4} \), set \( \tilde{\nu} \leftarrow \hat{\nu} \).

**Optimistic Phase:** For each round \( t \)...
\[
\text{UCB}_t(a) \approx \hat{E}_{\nu_a}[Y] + \sqrt{(\log T)/N_a(t)}.
\]
\[
\text{UCB}_t(a) \approx \sum_{z \in \mathcal{Z}} \left[ \hat{E}_{\nu}[Y \mid Z = z] + \sqrt{(\log T)/N_z(t)} \right] \tilde{\nu}_a(Z = z).
\]
If \( \text{UCB}_t(a) \approx \text{UCB}_t(a) \), play \( A_{t+1} = \arg \max_{a \in \mathcal{A}} \text{UCB}_t(a) \).
Otherwise, switch to Pessimistic Phase.

**Pessimistic Phase:** For remaining rounds \( t \), play \( A_{t+1} = \arg \max_{a \in \mathcal{A}} \text{UCB}_t(a) \).
Intuition: Optimistically play C-UCB until a hypothesis test for conditionally benign fails, then play UCB.

(1) Initial Exploration
Uniformly sample \( a \in \mathcal{A} \) for \( \sqrt{T}/|\mathcal{A}| \) rounds.
Compute MLE estimate \( \hat{\nu} \) of \((\nu_a(Z))_{a \in \mathcal{A}}\). If \( \sup_{a \in \mathcal{A}} \| \tilde{\nu}_a - \hat{\nu}_a \|_1 \gtrsim T^{-1/4} \), set \( \tilde{\nu} \leftarrow \hat{\nu} \).

Optimistic Phase: For each round \( t \)...
\[
\begin{align*}
\text{UCB}_t(a) & \approx \hat{E}_{\nu_a}[Y] + \sqrt{(\log T)/N_a(t)}. \\
\overline{\text{UCB}}_t(a) & \approx \sum_{z \in \mathcal{Z}} [\hat{E}_{\nu}[Y \mid Z = z] + \sqrt{(\log T)/N_z(t)}] \hat{\nu}_a(Z = z).
\end{align*}
\]
If \( \text{UCB}_t(a) \approx \overline{\text{UCB}}_t(a) \), play \( A_{t+1} = \arg \max_{a \in \mathcal{A}} \text{UCB}_t(a) \).
Otherwise, switch to Pessimistic Phase.

Pessimistic Phase: For remaining rounds \( t \), play \( A_{t+1} = \arg \max_{a \in \mathcal{A}} \text{UCB}_t(a) \).
Proof Sketch for HAC-UCB

(1) Exploration Rounds

In the worst case, C-UCB never plays the optimal $a \in A$. To circumvent this, we explore each $a \in A$ for an initial $\sqrt{T/|A|}$ rounds. This is fine from a minimax perspective since even conditionally benign forces $\sqrt{T}$ regret.

Estimating a multinomial to scale $\varepsilon$ takes $\approx 1/\varepsilon^2$ samples, so we also use the initial exploration to obtain an $\varepsilon = T^{-1/4}$ accurate estimate of $\nu(Z)$.

(2) Optimistic Rounds

a) If the conditionally benign assumption holds, $\text{UCB}_t(a) \approx \tilde{\text{UCB}}_t(a)$ and the algorithm correctly plays optimistically.

b) If the conditionally benign assumption fails, either $\text{UCB}_t(a) \neq \tilde{\text{UCB}}_t(a)$ and the algorithm correctly plays pessimistically, or the regret incurred from playing optimistically is still sufficiently small.
Proof Sketch for HAC-UCB

(1) Exploration Rounds

In the worst case, C-UCB never plays the optimal action $a \in A$. To circumvent this, we explore each action $a \in A$ for an initial $\sqrt{T/|A|}$ rounds. This is fine from a minimax perspective since even conditionally benign forces $\sqrt{T}$ regret.

Estimating a multinomial to scale takes $\approx 1/\varepsilon^2$ samples, so we also use the initial exploration to obtain an $\varepsilon = T - 1/4$ accurate estimate of $\nu(Z)$.

(2) Optimistic Rounds

a) If the conditionally benign assumption holds, $\text{UCB}_t(a) \approx \tilde{\text{UCB}}_t(a)$ and the algorithm correctly plays optimistically.

b) If the conditionally benign assumption fails, either $\text{UCB}_t(a) \not\approx \tilde{\text{UCB}}_t(a)$ and the algorithm correctly plays pessimistically, or the regret incurred from playing optimistically is still sufficiently small.
(1) Exploration Rounds

In the worst case, \( \text{C-UCB} \) never plays the optimal \( a \in \mathcal{A} \).
Proof Sketch for HAC-UCB

(1) Exploration Rounds
In the worst case, C-UCB never plays the optimal \( a \in A \).
To circumvent this, we explore each \( a \in A \) for an initial \( \sqrt{T}/|A| \) rounds.

(2) Optimistic Rounds
a) If the conditionally benign assumption holds, \( \text{UCB}_t(a) \approx \tilde{\text{UCB}}_t(a) \) and the algorithm correctly plays optimistically.
b) If the conditionally benign assumption fails, either \( \text{UCB}_t(a) \not\approx \tilde{\text{UCB}}_t(a) \) and the algorithm correctly plays pessimistically, or the regret incurred from playing optimistically is still sufficiently small.

\[ 11/8 \]
Proof Sketch for HAC-UCB

(1) Exploration Rounds
In the worst case, C-UCB never plays the optimal $a \in A$.
To circumvent this, we explore each $a \in A$ for an initial $\sqrt{T/|A|}$ rounds.
This is fine from a minimax perspective since even conditionally benign forces $\sqrt{T}$ regret.

(2) Optimistic Rounds
a) If the conditionally benign assumption holds, $\text{UCB}_t(a) \approx \tilde{\text{UCB}}_t(a)$ and the algorithm correctly plays optimistically.
b) If the conditionally benign assumption fails, either $\text{UCB}_t(a) \not\approx \tilde{\text{UCB}}_t(a)$ and the algorithm correctly plays pessimistically, or the regret incurred from playing optimistically is still sufficiently small.

$11/8$
Proof Sketch for HAC-UCB

(1) Exploration Rounds
In the worst case, C-UCB never plays the optimal \( a \in \mathcal{A} \).
To circumvent this, we explore each \( a \in \mathcal{A} \) for an initial \( \sqrt{T/|\mathcal{A}|} \) rounds.
This is fine from a minimax perspective since even conditionally benign forces \( \sqrt{T} \) regret.
Estimating a multinomial to scale \( \varepsilon \) takes \( \approx 1/\varepsilon^2 \) samples,
so we also use the initial exploration to obtain an \( \varepsilon = T^{-1/4} \) accurate estimate of \( \nu(\mathcal{Z}) \).
Proof Sketch for HAC-UCB

(1) Exploration Rounds
In the worst case, C-UCB never plays the optimal \( a \in A \).
To circumvent this, we explore each \( a \in A \) for an initial \( \sqrt{T/|A|} \) rounds.
This is fine from a minimax perspective since even conditionally benign forces \( \sqrt{T} \) regret.
Estimating a multinomial to scale \( \varepsilon \) takes \( \approx 1/\varepsilon^2 \) samples,
so we also use the initial exploration to obtain an \( \varepsilon = T^{-1/4} \) accurate estimate of \( \nu(Z) \).

(2) Optimistic Rounds
Proof Sketch for HAC-UCB

(1) Exploration Rounds
In the worst case, C-UCB never plays the optimal \( a \in A \).
To circumvent this, we explore each \( a \in A \) for an initial \( \sqrt{T/|A|} \) rounds.
This is fine from a minimax perspective since even conditionally benign forces \( \sqrt{T} \) regret.
Estimating a multinomial to scale \( \varepsilon \) takes \( \approx 1/\varepsilon^2 \) samples,
so we also use the initial exploration to obtain an \( \varepsilon = T^{-1/4} \) accurate estimate of \( \nu(Z) \).

(2) Optimistic Rounds
a) If the conditionally benign assumption holds,
\( \text{UCB}_t(a) \approx \tilde{\text{UCB}}_t(a) \) and the algorithm correctly plays optimistically.

b) If the conditionally benign assumption fails,
either \( \text{UCB}_t(a) \not\approx \tilde{\text{UCB}}_t(a) \) and the algorithm correctly plays pessimistically,
or the regret incurred from playing optimistically is still sufficiently small.
Proof Sketch for HAC-UCB

(1) Exploration Rounds
In the worst case, C-UCB never plays the optimal \( a \in A \).
To circumvent this, we explore each \( a \in A \) for an initial \( \sqrt{T/|A|} \) rounds.
This is fine from a minimax perspective since even conditionally benign forces \( \sqrt{T} \) regret.
Estimating a multinomial to scale \( \varepsilon \) takes \( \approx 1/\varepsilon^2 \) samples,
so we also use the initial exploration to obtain an \( \varepsilon = T^{-1/4} \) accurate estimate of \( \nu(Z) \).

(2) Optimistic Rounds
a) If the conditionally benign assumption holds,
\[ \text{UCB}_t(a) \approx \text{UCB}_t(a) \] and the algorithm correctly plays optimistically.

b) If the conditionally benign assumption fails,
either \( \text{UCB}_t(a) \not\approx \text{UCB}_t(a) \) and the algorithm correctly plays pessimistically,
or the regret incurred from playing optimistically is still sufficiently small.
More on Conditionally Benign

Suppose we have a fixed DAG $G$ on $(A \times Z \times Y)$.

(a) conditionally benign and $d$-separated
(b) not conditionally benign
(c) conditionally benign through front-door, not $d$-separated
(d) no adjustment possible, not conditionally benign
More on Conditionally Benign

Suppose we have a fixed DAG $G$ on $(\mathcal{A} \times \mathcal{Z} \times \mathcal{Y})$. 
More on Conditionally Benign

Suppose we have a fixed DAG $G$ on $(A \times Z \times Y)$.

(a) conditionally benign and $d$-separated
More on Conditionally Benign

Suppose we have a fixed DAG $G$ on $\mathcal{A} \times \mathcal{Z} \times \mathcal{Y}$.

(a) conditionally benign and $d$-separated
(b) not conditionally benign
More on Conditionally Benign

Suppose we have a fixed DAG $G$ on $(A \times Z \times Y)$.

(a) conditionally benign and $d$-separated
(b) not conditionally benign
(c) conditionally benign through front-door, not $d$-separated
Suppose we have a fixed DAG $G$ on $(A \times Z \times Y)$.

(a) conditionally benign and $d$-separated
(b) not conditionally benign
(c) conditionally benign through front-door, not $d$-separated
(d) no adjustment possible, not conditionally benign
More on Conditionally Benign

Suppose we have a fixed DAG \( G \) on \((A \times Z \times Y)\).

Let \( G_A \) denote the graph with edges into \( A \) removed.

Theorem
Let \( A \) be all hard interventions. \( Z \) \( d \)-separates \( Y \) from \( A \) on \( G \) if and only if every Markov relative \( \nu \) on \( G \) is conditionally benign on \( A \).

Theorem
Let \( A_0 \) be all hard interventions except the null (observational) intervention. \( Z \) \( d \)-separates \( Y \) from \( A \) on \( G_A \) if and only if every Markov relative \( \nu \) on \( G \) is conditionally benign on \( A_0 \).

Proposition
If \( Z \) satisfies the front-door criterion with respect to \((A,Y)\) on \( G \) then \( Z \) \( d \)-separates \( Y \) from \( A \) on \( G_A \).
More on Conditionally Benign

Suppose we have a fixed DAG $G$ on $(\mathcal{A} \times \mathcal{Z} \times \mathcal{Y})$. 
More on Conditionally Benign

Suppose we have a fixed DAG $G$ on $(A \times Z \times Y)$.

**Theorem**
Let $A$ be all hard interventions. $Z d$-separates $Y$ from $A$ on $G$ if and only if every Markov relative $\nu$ on $G$ is conditionally benign on $A$. 

**Proposition**
If $Z$ satisfies the front-door criterion with respect to $(A,Y)$ on $G$ then $Z d$-separates $Y$ from $A$ on $G A$. 

Suppose we have a fixed DAG $G$ on $(A \times Z \times Y)$. Let $G_A$ denote the graph with edges into $A$ removed.

**Theorem**

Let $A$ be all hard interventions.

$Z$ $d$-separates $Y$ from $A$ on $G$ if and only if every Markov relative $\nu$ on $G$ is conditionally benign on $A$. 
More on Conditionally Benign

Suppose we have a fixed DAG $G$ on $(A \times Z \times Y)$. Let $G_A$ denote the graph with edges into $A$ removed.

**Theorem**

Let $A$ be all hard interventions.

$Z$ $d$-separates $Y$ from $A$ on $G$ if and only if every Markov relative $\nu$ on $G$ is conditionally benign on $A$.

**Theorem**

Let $A_0$ be all hard interventions except the null (observational) intervention.

$Z$ $d$-separates $Y$ from $A$ on $G_A$ if and only if every Markov relative $\nu$ on $G$ is conditionally benign on $A_0$. 
More on Conditionally Benign

Suppose we have a fixed DAG $\mathcal{G}$ on $(A \times Z \times Y)$. Let $\mathcal{G}_{\overline{A}}$ denote the graph with edges into $A$ removed.

**Theorem**

Let $\mathcal{A}$ be all hard interventions.

$Z$ $d$-separates $Y$ from $A$ on $\mathcal{G}$ if and only if every Markov relative $\nu$ on $\mathcal{G}$ is conditionally benign on $\mathcal{A}$.

**Theorem**

Let $\mathcal{A}_0$ be all hard interventions except the null (observational) intervention.

$Z$ $d$-separates $Y$ from $A$ on $\mathcal{G}_{\overline{A}}$ if and only if every Markov relative $\nu$ on $\mathcal{G}$ is conditionally benign on $\mathcal{A}_0$.

**Proposition**

If $Z$ satisfies the front-door criterion with respect to $(A, Y)$ on $\mathcal{G}$ then $Z$ $d$-separates $Y$ from $A$ on $\mathcal{G}_{\overline{A}}$. 