Influencing Bandits

D Manjunath

CMInDS and EE
IIT Bombay

February 27, 2024

Background

- Consider an ad placement system (APS) that has to display k out of m possible ads.
  - Value of an ad to the APS is the click-through proabability for the ad.
  - Further different ads could provide different revenues with a click-through.
  - There is an incentive to place items with low click-through probabilities.
- Not unreasonable that the click-through probability depends on the history of ads seen before
  - Some are annoyed by repetition; for others disinterest can turn to curiosity.
- If the APS is learning the interests of the user, and hence the value of each ad, the learning algorithm will explore
  - How will exploration shape the preferences of the population?
- General interest: Capture the effect of the history of ads placed on the preferences? Remarks on this later
- For now, assume an extreme case of the system wanting to **shape the population** preferences through the ads placed or, in the case of a recommendation system, the items that it recommends—Opinion shaping or opinion control.

Background

- Consider an ad placement system (APS) that has to display k out of m possible ads.
    - Value of an ad to the APS is the click-through proabability for the ad.
    - Further different ads could provide different revenues with a click-through.
    - There is an incentive to place items with low click-through probabilities.
- Not unreasonable that the click-through probability depends on the history of ads seen before
    - Some are annoyed by repetition; for others disinterest can turn to curiosity.
- If the APS is learning the interests of the user, and hence the value of each ad, the learning algorithm will explore
    - How will exploration shape the preferences of the population?
- General interest: Capture the effect of the history of ads placed on the preferences? Remarks on this later
- For now, assume an extreme case of the system wanting to **shape the population** preferences through the ads placed or, in the case of a recommendation system, the items that it recommends—Opinion shaping or opinion control.

Background

- Consider an ad placement system (APS) that has to display k out of m possible ads.
  - Value of an ad to the APS is the click-through proabability for the ad.
  - Further different ads could provide different revenues with a click-through.
  - There is an incentive to place items with low click-through probabilities.
- Not unreasonable that the click-through probability depends on the history of ads seen before
  - Some are annoyed by repetition; for others disinterest can turn to curiosity.
- If the APS is learning the interests of the user, and hence the value of each ad, the learning algorithm will explore
  - How will exploration shape the preferences of the population?
- General interest: Capture the effect of the history of ads placed on the preferences? Remarks on this later
- For now, assume an extreme case of the system wanting to **shape the population** preferences through the ads placed or, in the case of a recommendation system, the items that it recommends—Opinion shaping or opinion control.

Background

- Consider an ad placement system (APS) that has to display k out of m possible ads.
  - Value of an ad to the APS is the click-through proabability for the ad.
  - Further different ads could provide different revenues with a click-through.
  - There is an incentive to place items with low click-through probabilities.
- Not unreasonable that the click-through probability depends on the history of ads seen before
  - Some are annoyed by repetition; for others disinterest can turn to curiosity.
- If the APS is learning the interests of the user, and hence the value of each ad, the learning algorithm will explore
  - How will exploration shape the preferences of the population?
- General interest: Capture the effect of the history of ads placed on the preferences? Remarks on this later
- For now, assume an extreme case of the system wanting to **shape the population** preferences through the ads placed or, in the case of a recommendation system, the items that it recommends—Opinion shaping or opinion control.

Background

- Consider an ad placement system (APS) that has to display k out of m possible ads.
  - Value of an ad to the APS is the click-through proaability for the ad.
  - Further different ads could provide different revenues with a click-through.
  - There is an incentive to place items with low click-through probabilities.
- Not unreasonable that the click-through probability depends on the history of ads seen before
  - Some are annoyed by repetition; for others disinterest can turn to curiosity.
- If the APS is learning the interests of the user, and hence the value of each ad, the learning algorithm will explore
  - How will exploration shape the preferences of the population?
- General interest: Capture the effect of the history of ads placed on the preferences? Remarks on this later
- For now, assume an extreme case of the system wanting to **shape the population** preferences through the ads placed or, in the case of a recommendation system, the items that it recommends—Opinion shaping or opinion control.

Background

- Consider an ad placement system (APS) that has to display k out of m possible ads.
  - Value of an ad to the APS is the click-through proabability for the ad.
  - Further different ads could provide different revenues with a click-through.
  - There is an incentive to place items with low click-through probabilities.
- Not unreasonable that the click-through probability depends on the history of ads seen before
  - Some are annoyed by repetition; for others disinterest can turn to curiosity.
- If the APS is learning the interests of the user, and hence the value of each ad, the learning algorithm will explore
  - How will exploration shape the preferences of the population?
- General interest: Capture the effect of the history of ads placed on the preferences? Remarks on this later
- For now, assume an extreme case of the system wanting to **shape the population** preferences through the ads placed or, in the case of a recommendation system, the items that it recommends—Opinion shaping or opinion control.

Background

- Consider an ad placement system (APS) that has to display k out of m possible ads.
  - Value of an ad to the APS is the click-through proabability for the ad.
  - Further different ads could provide different revenues with a click-through.
  - There is an incentive to place items with low click-through probabilities.
- Not unreasonable that the click-through probability depends on the history of ads seen before
  - Some are annoyed by repetition; for others disinterest can turn to curiosity.
- If the APS is learning the interests of the user, and hence the value of each ad, the learning algorithm will explore
  - How will exploration shape the preferences of the population?
- General interest: Capture the effect of the history of ads placed on the preferences? Remarks on this later
- For now, assume an extreme case of the system wanting to **shape the population** preferences through the ads placed or, in the case of a recommendation system, the items that it recommends—Opinion shaping or opinion control.

Background

- Consider an ad placement system (APS) that has to display k out of m possible ads.
  - Value of an ad to the APS is the click-through proabability for the ad.
  - Further different ads could provide different revenues with a click-through.
  - There is an incentive to place items with low click-through probabilities.
- Not unreasonable that the click-through probability depends on the history of ads seen before
  - Some are annoyed by repetition; for others disinterest can turn to curiosity.
- If the APS is learning the interests of the user, and hence the value of each ad, the learning algorithm will explore
  - How will exploration shape the preferences of the population?
- General interest: Capture the effect of the history of ads placed on the preferences? Remarks on this later
- For now, assume an extreme case of the system wanting to **shape the population** preferences through the ads placed or, in the case of a recommendation system, the items that it recommends—Opinion shaping or opinion control.

Background

- Consider an ad placement system (APS) that has to display k out of m possible ads.
  - Value of an ad to the APS is the click-through proabability for the ad.
  - Further different ads could provide different revenues with a click-through.
  - There is an incentive to place items with low click-through probabilities.
- Not unreasonable that the click-through probability depends on the history of ads seen before
  - Some are annoyed by repetition; for others disinterest can turn to curiosity.
- If the APS is learning the interests of the user, and hence the value of each ad, the learning algorithm will explore
  - How will exploration shape the preferences of the population?
- General interest: Capture the effect of the history of ads placed on the preferences? Remarks on this later
- For now, assume an extreme case of the system wanting to **shape the population** preferences through the ads placed or, in the case of a recommendation system, the items that it recommends—Opinion shaping or opinion control.

Background

- Consider an ad placement system (APS) that has to display k out of m possible ads.
  - Value of an ad to the APS is the click-through proabability for the ad.
  - Further different ads could provide different revenues with a click-through.
  - There is an incentive to place items with low click-through probabilities.
- Not unreasonable that the click-through probability depends on the history of ads seen before
  - Some are annoyed by repetition; for others disinterest can turn to curiosity.
- If the APS is learning the interests of the user, and hence the value of each ad, the learning algorithm will explore
  - How will exploration shape the preferences of the population?
- General interest: Capture the effect of the history of ads placed on the preferences? Remarks on this later
- For now, assume an extreme case of the system wanting to **shape the population** preferences through the ads placed or, in the case of a recommendation system, the items that it recommends—Opinion shaping or opinion control.

Background

- Consider an ad placement system (APS) that has to display k out of m possible ads.
    - Value of an ad to the APS is the click-through proabability for the ad.
    - Further different ads could provide different revenues with a click-through.
    - There is an incentive to place items with low click-through probabilities.
- Not unreasonable that the click-through probability depends on the history of ads seen before
    - Some are annoyed by repetition; for others disinterest can turn to curiosity.
- If the APS is learning the interests of the user, and hence the value of each ad, the learning algorithm will explore
    - How will exploration shape the preferences of the population?
- General interest: Capture the effect of the history of ads placed on the preferences? Remarks on this later
- For now, assume an extreme case of the system wanting to **shape the population** preferences through the ads placed or, in the case of a recommendation system, the items that it recommends—Opinion shaping or opinion control.

Background

- Consider an ad placement system (APS) that has to display k out of m possible ads.
  - Value of an ad to the APS is the click-through proabability for the ad.
  - Further different ads could provide different revenues with a click-through.
  - There is an incentive to place items with low click-through probabilities.
- Not unreasonable that the click-through probability depends on the history of ads seen before
  - Some are annoyed by repetition; for others disinterest can turn to curiosity.
- If the APS is learning the interests of the user, and hence the value of each ad, the learning algorithm will explore
  - How will exploration shape the preferences of the population?
- General interest: Capture the effect of the history of ads placed on the preferences? Remarks on this later
- For now, assume an extreme case of the system wanting to **shape the population** preferences through the ads placed or, in the case of a recommendation system, the items that it recommends—Opinion shaping or opinion control.

Background

- Consider an ad placement system (APS) that has to display k out of m possible ads.
  - Value of an ad to the APS is the click-through proabability for the ad.
  - Further different ads could provide different revenues with a click-through.
  - There is an incentive to place items with low click-through probabilities.
- Not unreasonable that the click-through probability depends on the history of ads seen before
  - Some are annoyed by repetition; for others disinterest can turn to curiosity.
- If the APS is learning the interests of the user, and hence the value of each ad, the learning algorithm will explore
  - How will exploration shape the preferences of the population?
- General interest: Capture the effect of the history of ads placed on the preferences? Remarks on this later
- For now, assume an extreme case of the system wanting to **shape the population** preferences through the ads placed or, in the case of a recommendation system, the items that it recommends—Opinion shaping or opinion control.

Background

- Consider an ad placement system (APS) that has to display k out of m possible ads.
  - Value of an ad to the APS is the click-through proabability for the ad.
  - Further different ads could provide different revenues with a click-through.
  - There is an incentive to place items with low click-through probabilities.
- Not unreasonable that the click-through probability depends on the history of ads seen before
  - Some are annoyed by repetition; for others disinterest can turn to curiosity.
- If the APS is learning the interests of the user, and hence the value of each ad, the learning algorithm will explore
  - How will exploration shape the preferences of the population?
- General interest: Capture the effect of the history of ads placed on the preferences? Remarks on this later
- For now, assume an extreme case of the system wanting to **shape the population** preferences through the ads placed or, in the case of a recommendation system, the items that it recommends—Opinion shaping or opinion control.

Model

- Two types of users in the population distinguished by preferences; Recommendation system, or an APS, S serves the population.
- S recommends one of two arms to each arriving user.
- Time is discrete and takes values $t \in [1 : T]$.
- At time t, a user of type $X_t \in \{1, 2\}$ arrives, S observes the type and shows arm $A_t \in \{a_1, a_2\}$.
- Fraction of type 1 and type 2 users is tracked by an urn containing colored balls—colors 1 and 2 correspond to, respectively, types 1 and 2.
- Fraction of type 1 users in the population equals the fraction of type 1 balls in the urn.

Model

- Two types of users in the population distinguished by preferences; Recommendation system, or an APS, S serves the population.
- S recommends one of two arms to each arriving user.
- Time is discrete and takes values $t \in [1:T]$.
- At time t, a user of type $X_t \in \{1, 2\}$ arrives, S observes the type and shows arm $A_t \in \{a_1, a_2\}$.
- Fraction of type 1 and type 2 users is tracked by an urn containing colored balls—colors 1 and 2 correspond to, respectively, types 1 and 2.
- Fraction of type 1 users in the population equals the fraction of type 1 balls in the urn.

Model

- Two types of users in the population distinguished by preferences; Recommendation system, or an APS, S serves the population.
- S recommends one of two arms to each arriving user.
- Time is discrete and takes values $t \in [1 : T]$.
- At time t, a user of type $X_t \in \{1, 2\}$ arrives, S observes the type and shows arm $A_t \in \{a_1, a_2\}$.
- Fraction of type 1 and type 2 users is tracked by an urn containing colored balls—colors 1 and 2 correspond to, respectively, types 1 and 2.
- Fraction of type 1 users in the population equals the fraction of type 1 balls in the urn.

Model

- Two types of users in the population distinguished by preferences; Recommendation system, or an APS, S serves the population.
- S recommends one of two arms to each arriving user.
- Time is discrete and takes values $t \in [1 : T]$.
- At time t, a user of type $X_t \in \{1, 2\}$ arrives, S observes the type and shows arm $A_t \in \{a_1, a_2\}$.
- Fraction of type 1 and type 2 users is tracked by an urn containing colored balls—colors 1 and 2 correspond to, respectively, types 1 and 2.
- Fraction of type 1 users in the population equals the fraction of type 1 balls in the urn.

Model

- Two types of users in the population distinguished by preferences; Recommendation system, or an APS, S serves the population.
- S recommends one of two arms to each arriving user.
- Time is discrete and takes values $t \in [1 : T]$.
- At time t, a user of type $X_t \in \{1, 2\}$ arrives, S observes the type and shows arm $A_t \in \{a_1, a_2\}$.
- Fraction of type 1 and type 2 users is tracked by an urn containing colored balls—colors 1 and 2 correspond to, respectively, types 1 and 2.
- Fraction of type 1 users in the population equals the fraction of type 1 balls in the urn.

Model

- Two types of users in the population distinguished by preferences; Recommendation system, or an APS, S serves the population.
- S recommends one of two arms to each arriving user.
- Time is discrete and takes values $t \in [1 : T]$.
- At time t, a user of type $X_t \in \{1, 2\}$ arrives, S observes the type and shows arm $A_t \in \{a_1, a_2\}$.
- Fraction of type 1 and type 2 users is tracked by an urn containing colored balls—colors 1 and 2 correspond to, respectively, types 1 and 2.
- Fraction of type 1 users in the population equals the fraction of type 1 balls in the urn.

Model

- **Reward Structure**
  - Suppose $X_t = i$ and $A_t = j$
  - S gets a random Bernoulli reward $W_t \in \{0, 1\}$ with mean b_{ij}
  - $B = [[b_{ij}]]$ is the reward means matrix.
  - WLOG, assume b_{ii} is the maximum in row i of B.
- **Population Dynamics**
  - $Z_i(t)$ is the number of type i balls in urn at time t.
  - $N_0 = Z_1(0) + Z_2(0)$ is the total number of balls at $t = 0$.
  - User arriving at time t is of type i with probability $z_i(t) = Z_i(t)/(\sum_j Z_j(t))$.
  - Realization of reward at time t, W_t, causes urn to be updated.
  - Two evolution models for the urn.

Model

- **Reward Structure**
  - Suppose $X_t = i$ and $A_t = j$
  - S gets a random Bernoulli reward $W_t \in \{0, 1\}$ with mean b_{ij}
  - $B = [[b_{ij}]]$ is the reward means matrix.
  - WLOG, assume b_{ii} is the maximum in row i of B.
- **Population Dynamics**
  - $Z_i(t)$ is the number of type i balls in urn at time t.
  - $N_0 = Z_1(0) + Z_2(0)$ is the total number of balls at $t = 0$.
  - User arriving at time t is of type i with probability $z_i(t) = Z_i(t)/(\sum_j Z_j(t))$.
  - Realization of reward at time t, W_t, causes urn to be updated.
  - Two evolution models for the urn.

Model

- **Reward Structure**
  - Suppose $X_t = i$ and $A_t = j$
  - S gets a random Bernoulli reward $W_t \in \{0, 1\}$ with mean b_{ij}
  - $B = [[b_{ij}]]$ is the reward means matrix.
  - WLOG, assume b_{ii} is the maximum in row i of B.
- **Population Dynamics**
  - $Z_i(t)$ is the number of type i balls in urn at time t.
  - $N_0 = Z_1(0) + Z_2(0)$ is the total number of balls at $t = 0$.
  - User arriving at time t is of type i with probability $z_i(t) = Z_i(t)/(\sum_j Z_j(t))$.
  - Realization of reward at time t, W_t, causes urn to be updated.
  - Two evolution models for the urn.

Model

- **Reward Structure**
  - Suppose $X_t = i$ and $A_t = j$
  - S gets a random Bernoulli reward $W_t \in \{0, 1\}$ with mean b_{ij}
  - $B = [[b_{ij}]]$ is the reward means matrix.
  - WLOG, assume b_{ii} is the maximum in row i of B.
- **Population Dynamics**
  - $Z_i(t)$ is the number of type i balls in urn at time t.
  - $N_0 = Z_1(0) + Z_2(0)$ is the total number of balls at $t = 0$.
  - User arriving at time t is of type i with probability $z_i(t) = Z_i(t)/(\sum_j Z_j(t))$.
  - Realization of reward at time t, W_t, causes urn to be updated.
  - Two evolution models for the urn.

Model

- **Reward Structure**
  - Suppose $X_t = i$ and $A_t = j$
  - S gets a random Bernoulli reward $W_t \in \{0, 1\}$ with mean b_{ij}
  - $B = [[b_{ij}]]$ is the reward means matrix.
  - WLOG, assume b_{ii} is the maximum in row i of B.
- **Population Dynamics**
  - $Z_i(t)$ is the number of type i balls in urn at time t.
  - $N_0 = Z_1(0) + Z_2(0)$ is the total number of balls at $t = 0$.
  - User arriving at time t is of type i with probability $z_i(t) = Z_i(t)/(\sum_j Z_j(t))$.
  - Realization of reward at time t, W_t, causes urn to be updated.
  - Two evolution models for the urn.

Model

- **Reward Structure**
  - Suppose $X_t = i$ and $A_t = j$
  - S gets a random Bernoulli reward $W_t \in \{0, 1\}$ with mean b_{ij}
  - $B = [[b_{ij}]]$ is the reward means matrix.
  - WLOG, assume b_{ii} is the maximum in row i of B.
- **Population Dynamics**
  - $Z_i(t)$ is the number of type i balls in urn at time t.
  - $N_0 = Z_1(0) + Z_2(0)$ is the total number of balls at $t = 0$.
  - User arriving at time t is of type i with probability $z_i(t) = Z_i(t)/(\sum_j Z_j(t))$.
  - Realization of reward at time t, W_t, causes urn to be updated.
  - Two evolution models for the urn.

Model

- **Reward Structure**
  - Suppose $X_t = i$ and $A_t = j$
  - S gets a random Bernoulli reward $W_t \in \{0, 1\}$ with mean b_{ij}
  - $B = [[b_{ij}]]$ is the reward means matrix.
  - WLOG, assume b_{ii} is the maximum in row i of B.
- **Population Dynamics**
  - $Z_i(t)$ is the number of type i balls in urn at time t.
  - $N_0 = Z_1(0) + Z_2(0)$ is the total number of balls at $t = 0$.
  - User arriving at time t is of type i with probability $z_i(t) = Z_i(t)/(\sum_j Z_j(t))$.
  - Realization of reward at time t, W_t, causes urn to be updated.
  - Two evolution models for the urn.

Model

- **Reward Structure**
  - Suppose $X_t = i$ and $A_t = j$
  - S gets a random Bernoulli reward $W_t \in \{0, 1\}$ with mean b_{ij}
  - $B = [[b_{ij}]]$ is the reward means matrix.
  - WLOG, assume b_{ii} is the maximum in row i of B.
- **Population Dynamics**
  - $Z_i(t)$ is the number of type i balls in urn at time t.
  - $N_0 = Z_1(0) + Z_2(0)$ is the total number of balls at $t = 0$.
  - User arriving at time t is of type i with probability $z_i(t) = Z_i(t)/(\sum_j Z_j(t))$.
  - Realization of reward at time t, W_t, causes urn to be updated.
  - Two evolution models for the urn.

Model

- **Reward Structure**
  - Suppose $X_t = i$ and $A_t = j$
  - S gets a random Bernoulli reward $W_t \in \{0, 1\}$ with mean b_{ij}
  - $B = [[b_{ij}]]$ is the reward means matrix.
  - WLOG, assume b_{ii} is the maximum in row i of B.
- **Population Dynamics**
  - $Z_i(t)$ is the number of type i balls in urn at time t.
  - $N_0 = Z_1(0) + Z_2(0)$ is the total number of balls at $t = 0$.
  - User arriving at time t is of type i with probability $z_i(t) = Z_i(t)/(\sum_j Z_j(t))$.
  - Realization of reward at time t, W_t, causes urn to be updated.
  - Two evolution models for the urn.

Model

- **Reward Structure**
  - Suppose $X_t = i$ and $A_t = j$
  - S gets a random Bernoulli reward $W_t \in \{0, 1\}$ with mean b_{ij}
  - $B = [[b_{ij}]]$ is the reward means matrix.
  - WLOG, assume b_{ii} is the maximum in row i of B.
- **Population Dynamics**
  - $Z_i(t)$ is the number of type i balls in urn at time t.
  - $N_0 = Z_1(0) + Z_2(0)$ is the total number of balls at $t = 0$.
  - User arriving at time t is of type i with probability $z_i(t) = Z_i(t)/(\sum_j Z_j(t))$.
  - Realization of reward at time t, W_t, causes urn to be updated.
  - Two evolution models for the urn.

Model

- **Reward Structure**
  - Suppose $X_t = i$ and $A_t = j$
  - S gets a random Bernoulli reward $W_t \in \{0, 1\}$ with mean b_{ij}
  - $B = [[b_{ij}]]$ is the reward means matrix.
  - WLOG, assume b_{ii} is the maximum in row i of B.
- **Population Dynamics**
  - $Z_i(t)$ is the number of type i balls in urn at time t.
  - $N_0 = Z_1(0) + Z_2(0)$ is the total number of balls at $t = 0$.
  - User arriving at time t is of type i with probability $z_i(t) = Z_i(t)/(\sum_j Z_j(t))$.
  - Realization of reward at time t, W_t, causes urn to be updated.
  - Two evolution models for the urn.

Model

- **Decreasing influence dynamics (DID) model:** Population becomes less plastic with time.
  - Total number of balls in the urn increases by one each time.
  - If $W_t = 1$, add ball of type A_t to urn
  - If $W_t = 0$, add ball of type $-A_t$ to urn

$$Z_{A_t}(t+1) = Z_{A_t}(t) + W_t,$$
$$Z_{-A_t}(t+1) = Z_{-A_t}(t) + (1 - W_t).$$

- **Constant Influence Model: Voter model**
  - Total number of balls in the urn remains constant
  - If X_t is shown shown arm a_{-X_t} and $W_t = 1$, *OR* if it is shown arm a_{X_t} and $W_t = 0$ then one ball of type X_t switches colors.
  - No change otherwise.
  - Writing $\theta_t = \{A_t = a_{-X_t}\}$, the urn evolution will be

- Reiterate: B does not change with time. Only Z, and hence population preference, changes.

Model

- **Decreasing influence dynamics (DID) model:** Population becomes less plastic with time.
  - Total number of balls in the urn increases by one each time.
  - If $W_t = 1$, add ball of type A_t to urn
  - If $W_t = 0$, add ball of type $-A_t$ to urn

$$Z_{A_t}(t+1) = Z_{A_t}(t) + W_t,$$
$$Z_{-A_t}(t+1) = Z_{-A_t}(t) + (1 - W_t).$$

- **Constant Influence Model: Voter model**
  - Total number of balls in the urn remains constant
  - If X_t is shown shown arm a_{-X_t} and $W_t = 1$, *OR* if it is shown arm a_{X_t} and $W_t = 0$ then one ball of type X_t switches colors
  - No change otherwise
  - Writing $\theta_t = \mathbb{I}\{A_t = a_{-X_t}\}$, the urn evolution will be

- Reiterate: B does not change with time. Only Z, and hence population preference, changes.

Model

- **Decreasing influence dynamics (DID) model:** Population becomes less plastic with time.
  - Total number of balls in the urn increases by one each time.
  - If $W_t = 1$, add ball of type A_t to urn
  - If $W_t = 0$, add ball of type $-A_t$ to urn

$$Z_{A_t}(t+1) = Z_{A_t}(t) + W_t,$$
$$Z_{-A_t}(t+1) = Z_{-A_t}(t) + (1 - W_t).$$

- **Constant Influence Model: Voter model**
  - Total number of balls in the urn remains constant
  - If X_t is shown shown arm a_{-X_t} and $W_t = 1$,

- Reiterate: B does not change with time. Only Z_t and hence population preference, changes.

Model

- **Decreasing influence dynamics (DID) model:** Population becomes less plastic with time.
  - Total number of balls in the urn increases by one each time.
  - If $W_t = 1$, add ball of type A_t to urn
  - If $W_t = 0$, add ball of type $-A_t$ to urn

$$Z_{A_t}(t+1) = Z_{A_t}(t) + W_t,$$
$$Z_{-A_t}(t+1) = Z_{-A_t}(t) + (1 - W_t).$$

- **Constant Influence Model: Voter model**
  - Total number of balls in the urn remains constant
  - If X_t is shown shown arm a_{-X_t} and $W_t = 1$, *OR* if it is shown arm a_{X_t} and $W_t = 0$ then one ball of type X_t switches colour.
  - No change otherwise.
  - Writing [illegible], the urn evolution will be

[illegible]

- Reiterate: B does not change with time. Only Z, and hence population preference, changes.

Model

- **Decreasing influence dynamics (DID) model:** Population becomes less plastic with time.
  - Total number of balls in the urn increases by one each time.
  - If $W_t = 1$, add ball of type A_t to urn
  - If $W_t = 0$, add ball of type $-A_t$ to urn

$$Z_{A_t}(t+1) = Z_{A_t}(t) + W_t,$$
$$Z_{-A_t}(t+1) = Z_{-A_t}(t) + (1 - W_t).$$

- **Constant Influence Model: Voter model**
  - Total number of balls in the urn remains constant
  - If X_t is shown shown arm a_{-X_t} and $W_t = 1$, *OR* if it is shown arm a_{X_t} and $W_t = 0$ then one ball of type X_t switches colors
  - No change otherwise.
  - Writing $\theta_t = \{A_t = a_{-X_t}\}$, the urn evolution will be

$$Z_{A_t}(t+1) = Z_{A_t}(t) + (\mathbb{1}_{\theta_t} \oplus W_t),$$
$$Z_{-A_t}(t+1) = Z_{-A_t}(t) - (\mathbb{1}_{\theta_t} \oplus W_t).$$

- Reiterate: B does not change with time. Only Z, and hence population preference, changes.

Model

- **Decreasing influence dynamics (DID) model:** Population becomes less plastic with time.
  - Total number of balls in the urn increases by one each time.
  - If $W_t = 1$, add ball of type A_t to urn
  - If $W_t = 0$, add ball of type $-A_t$ to urn

$$Z_{A_t}(t+1) = Z_{A_t}(t) + W_t,$$
$$Z_{-A_t}(t+1) = Z_{-A_t}(t) + (1 - W_t).$$

- **Constant Influence Model: Voter model**
  - Total number of balls in the urn remains constant
  - If X_t is shown shown arm a_{-X_t} and $W_t = 1$, *OR* if it is shown arn a_{X_t} and $W_t = 0$ then one ball of type X_t switches colors
  - No change otherwise.
  - Writing $\theta_t = \{A_t = a_{-X_t}\}$, the urn evolution will be

$$Z_{A_t}(t+1) = Z_{A_t}(t) + (\mathbb{1}_{\theta_t} \oplus W_t),$$
$$Z_{-A_t}(t+1) = Z_{-A_t}(t) - (\mathbb{1}_{\theta_t} \oplus W_t).$$

- Reiterate: B does not change with time. Only Z, and hence population preference, changes.

Model

- **Decreasing influence dynamics (DID) model:** Population becomes less plastic with time.
  - Total number of balls in the urn increases by one each time.
  - If $W_t = 1$, add ball of type A_t to urn
  - If $W_t = 0$, add ball of type $-A_t$ to urn

$$Z_{A_t}(t+1) = Z_{A_t}(t) + W_t,$$
$$Z_{-A_t}(t+1) = Z_{-A_t}(t) + (1 - W_t).$$

- **Constant Influence Model: Voter model**
  - Total number of balls in the urn remains constant
  - If X_t is shown shown arm a_{-X_t} and $W_t = 1$, *OR* if it is shown arn a_{X_t} and $W_t = 0$ then one ball of type X_t switches colors
  - No change otherwise.
  - Writing $\theta_t = \{A_t = a_{-X_t}\}$, the urn evolution will be

$$Z_{A_t}(t+1) = Z_{A_t}(t) + (\mathbb{1}_{\theta_t} \oplus W_t),$$
$$Z_{-A_t}(t+1) = Z_{-A_t}(t) - (\mathbb{1}_{\theta_t} \oplus W_t).$$

- Reiterate: B does not change with time. Only Z, and hence population preference, changes.

Model

- **Decreasing influence dynamics (DID) model:** Population becomes less plastic with time.
  - Total number of balls in the urn increases by one each time.
  - If $W_t = 1$, add ball of type A_t to urn
  - If $W_t = 0$, add ball of type $-A_t$ to urn

$$Z_{A_t}(t+1) = Z_{A_t}(t) + W_t,$$
$$Z_{-A_t}(t+1) = Z_{-A_t}(t) + (1 - W_t).$$

- **Constant Influence Model: Voter model**
  - Total number of balls in the urn remains constant
  - If X_t is shown shown arm a_{-X_t} and $W_t = 1$, *OR* if it is shown arn a_{X_t} and $W_t = 0$ then one ball of type X_t switches colors
  - No change otherwise.
  - Writing $\theta_t = \{A_t = a_{-X_t}\}$, the urn evolution will be

$$Z_{A_t}(t+1) = Z_{A_t}(t) + (\mathbb{1}_{\theta_t} \oplus W_t),$$
$$Z_{-A_t}(t+1) = Z_{-A_t}(t) - (\mathbb{1}_{\theta_t} \oplus W_t).$$

- Reiterate: B does not change with time. Only Z, and hence population preference, changes.

Model

- **Decreasing influence dynamics (DID) model:** Population becomes less plastic with time.
  - Total number of balls in the urn increases by one each time.
  - If $W_t = 1$, add ball of type A_t to urn
  - If $W_t = 0$, add ball of type $-A_t$ to urn

$$Z_{A_t}(t+1) = Z_{A_t}(t) + W_t,$$
$$Z_{-A_t}(t+1) = Z_{-A_t}(t) + (1 - W_t).$$

- **Constant Influence Model: Voter model**
  - Total number of balls in the urn remains constant
  - If X_t is shown shown arm a_{-X_t} and $W_t = 1$, *OR* if it is shown arn a_{X_t} and $W_t = 0$ then one ball of type X_t switches colors
  - No change otherwise.
  - Writing $\theta_t = \{A_t = a_{-X_t}\}$, the urn evolution will be

$$Z_{A_t}(t+1) = Z_{A_t}(t) + (\mathbb{1}_{\theta_t} \oplus W_t),$$
$$Z_{-A_t}(t+1) = Z_{-A_t}(t) - (\mathbb{1}_{\theta_t} \oplus W_t).$$

- Reiterate: B does not change with time. Only Z, and hence population preference, changes.

Model

- **Decreasing influence dynamics (DID) model:** Population becomes less plastic with time.
  - Total number of balls in the urn increases by one each time.
  - If $W_t = 1$, add ball of type A_t to urn
  - If $W_t = 0$, add ball of type $-A_t$ to urn

$$Z_{A_t}(t+1) = Z_{A_t}(t) + W_t,$$
$$Z_{-A_t}(t+1) = Z_{-A_t}(t) + (1 - W_t).$$

- **Constant Influence Model: Voter model**
  - Total number of balls in the urn remains constant
  - If X_t is shown shown arm a_{-X_t} and $W_t = 1$, *OR* if it is shown arn a_{X_t} and $W_t = 0$ then one ball of type X_t switches colors
  - No change otherwise.
  - Writing $\theta_t = \{A_t = a_{-X_t}\}$, the urn evolution will be

$$Z_{A_t}(t+1) = Z_{A_t}(t) + \left(\mathbb{1}_{\theta_t} \oplus W_t\right),$$
$$Z_{-A_t}(t+1) = Z_{-A_t}(t) - \left(\mathbb{1}_{\theta_t} \oplus W_t\right).$$

- Reiterate: B does not change with time. Only Z, and hence population preference, changes.

Model

- **Decreasing influence dynamics (DID) model:** Population becomes less plastic with time.
  - Total number of balls in the urn increases by one each time.
  - If $W_t = 1$, add ball of type A_t to urn
  - If $W_t = 0$, add ball of type $-A_t$ to urn

$$Z_{A_t}(t+1) = Z_{A_t}(t) + W_t,$$
$$Z_{-A_t}(t+1) = Z_{-A_t}(t) + (1 - W_t).$$

- **Constant Influence Model: Voter model**
  - Total number of balls in the urn remains constant
  - If X_t is shown shown arm a_{-X_t} and $W_t = 1$, *OR* if it is shown arn a_{X_t} and $W_t = 0$ then one ball of type X_t switches colors
  - No change otherwise.
  - Writing $\theta_t = \{A_t = a_{-X_t}\}$, the urn evolution will be

$$Z_{A_t}(t+1) = Z_{A_t}(t) + (\mathbb{1}_{\theta_t} \oplus W_t)\,,$$
$$Z_{-A_t}(t+1) = Z_{-A_t}(t) - (\mathbb{1}_{\theta_t} \oplus W_t)\,.$$

- Reiterate: B does not change with time. Only Z, and hence population preference, changes.

Model

- **Objective:** Achieve maximum possible increase in $z_1(t) := Z_1(t)/(Z_1(t) + Z_2(t))$ at every step.
- **Policy:** A policy $\pi = (p_t, q_t)$ where, for all $t \in [1:T]$,

$$\begin{aligned} p_t &= P(A_t = a_1 | X_t = 1, \{X_\tau, A_\tau, W_\tau\}_{\tau<t}) \\ q_t &= P(A_t = a_2 | X_t = 2, \{X_\tau, A_\tau, W_\tau\}_{\tau<t}). \end{aligned} \tag{1}$$

- **Optimal policy:** (p_t, q_t) that maximizes expected increase in $z_1(t)$ given the population profile at time t,

$$(p_t^*, q_t^*) = \arg\max_{(p_t, q_t)} E\left[\Delta Z_1^\pi(t) | z_1(t)\right] \tag{2}$$

where $\Delta Z_1^\pi(t) = Z_1^\pi(t+1) - Z_1^\pi(t)$.
- **One-step regret:** Regret at time t, (R_t^π) for a policy π is

$$R_t^\pi := E\left[\Delta Z_1^*(t) - \Delta Z_1^\pi(t) \mid Z_1^*(t) = Z_1^\pi(t)\right]. \tag{3}$$

- **Cumulative regret:**

$$R_{[1:T]}^\pi = \sum_{t=1}^{T} R_t^\pi.$$

Model

- **Objective:** Achieve maximum possible increase in $z_1(t) := Z_1(t)/(Z_1(t) + Z_2(t))$ at every step.
- **Policy:** A policy $\pi = (p_t, q_t)$ where, for all $t \in [1 : T]$,

$$
\begin{aligned}
p_t &= P(A_t = a_1 | X_t = 1, \{X_\tau, A_\tau, W_\tau\}_{\tau<t}) \\
q_t &= P(A_t = a_2 | X_t = 2, \{X_\tau, A_\tau, W_\tau\}_{\tau<t}).
\end{aligned} \tag{1}
$$

- **Optimal policy:** (p_t, q_t) that maximizes expected increase in $z_1(t)$ given the population profile at time t,

$$(p_t^*, q_t^*) = \arg\max_{(p_t,q_t)} E\left[\Delta Z_1^\pi(t) | z_1(t)\right] \tag{2}$$

where $\Delta Z_1^\pi(t) = Z_1^\pi(t+1) - Z_1^\pi(t)$.
- **One-step regret:** Regret at time t, (R_t^π) for a policy π is

$$R_t^\pi := E\left[\Delta Z_1^*(t) - \Delta Z_1^\pi(t) \mid Z_1^*(t) = Z_1^\pi(t)\right]. \tag{3}$$

- **Cumulative regret:**

$$R_{[1:T]}^\pi = \sum_{t=1}^{T} R_t^\pi.$$

Model

- **Objective:** Achieve maximum possible increase in $z_1(t) := Z_1(t)/(Z_1(t) + Z_2(t))$ at every step.
- **Policy:** A policy $\pi = (p_t, q_t)$ where, for all $t \in [1 : T]$,

$$\begin{aligned} p_t &= P(A_t = a_1 | X_t = 1, \{X_\tau, A_\tau, W_\tau\}_{\tau<t}) \\ q_t &= P(A_t = a_2 | X_t = 2, \{X_\tau, A_\tau, W_\tau\}_{\tau<t}). \end{aligned} \quad (1)$$

- **Optimal policy:** (p_t, q_t) that maximizes expected increase in $z_1(t)$ given the population profile at time t,

$$(p_t^*, q_t^*) = \arg\max_{(p_t, q_t)} E\left[\Delta Z_1^\pi(t) | z_1(t)\right] \quad (2)$$

where $\Delta Z_1^\pi(t) = Z_1^\pi(t+1) - Z_1^\pi(t)$.

- **One-step regret:** Regret at time t, (R_t^π) for a policy π is

$$R_t^\pi := E\left[\Delta Z_1^*(t) - \Delta Z_1^\pi(t) \mid Z_1^*(t) = Z_1^\pi(t)\right]. \quad (3)$$

- **Cumulative regret:**

$$R_{[1:T]}^\pi = \sum_{t=1}^{T} R_t^\pi.$$

Model

- **Objective:** Achieve maximum possible increase in $z_1(t) := Z_1(t)/(Z_1(t) + Z_2(t))$ at every step.
- **Policy:** A policy $\pi = (p_t, q_t)$ where, for all $t \in [1 : T]$,

$$\begin{aligned} p_t &= P(A_t = a_1 | X_t = 1, \{X_\tau, A_\tau, W_\tau\}_{\tau<t}) \\ q_t &= P(A_t = a_2 | X_t = 2, \{X_\tau, A_\tau, W_\tau\}_{\tau<t}). \end{aligned} \quad (1)$$

- **Optimal policy:** (p_t, q_t) that maximizes expected increase in $z_1(t)$ given the population profile at time t,

$$(p_t^*, q_t^*) = \arg\max_{(p_t,q_t)} E\left[\Delta Z_1^\pi(t) | z_1(t)\right] \quad (2)$$

  where $\Delta Z_1^\pi(t) = Z_1^\pi(t+1) - Z_1^\pi(t)$.
- **One-step regret:** Regret at time t, (R_t^π) for a policy π is

$$R_t^\pi := E\left[\Delta Z_1^*(t) - \Delta Z_1^\pi(t) \mid Z_1^*(t) = Z_1^\pi(t)\right]. \quad (3)$$

- **Cumulative regret:**

$$R_{[1:T]}^\pi = \sum_{t=1}^{T} R_t^\pi.$$

Model

- **Objective:** Achieve maximum possible increase in $z_1(t) := Z_1(t)/(Z_1(t) + Z_2(t))$ at every step.
- **Policy:** A policy $\pi = (p_t, q_t)$ where, for all $t \in [1 : T]$,

$$\begin{aligned} p_t &= P(A_t = a_1 | X_t = 1, \{X_\tau, A_\tau, W_\tau\}_{\tau<t}) \\ q_t &= P(A_t = a_2 | X_t = 2, \{X_\tau, A_\tau, W_\tau\}_{\tau<t}). \end{aligned} \tag{1}$$

- **Optimal policy:** (p_t, q_t) that maximizes expected increase in $z_1(t)$ given the population profile at time t,

$$(p_t^*, q_t^*) = \arg\max_{(p_t, q_t)} E\left[\Delta Z_1^\pi(t) | z_1(t)\right] \tag{2}$$

  where $\Delta Z_1^\pi(t) = Z_1^\pi(t+1) - Z_1^\pi(t)$.
- **One-step regret:** Regret at time t, (R_t^π) for a policy π is

$$R_t^\pi := E\left[\Delta Z_1^*(t) - \Delta Z_1^\pi(t) \mid Z_1^*(t) = Z_1^\pi(t)\right]. \tag{3}$$

- **Cumulative regret:**

$$R_{[1:T]}^\pi = \sum_{t=1}^{T} R_t^\pi.$$

Decreasing Influence Dynamics Model

Preference Shaping with Known B in D

- Optimal policy is a simple stationary policy

Lemma

The optimal policy for the time slot t is

$$(p_t^*, q_t^*) = (\mathbb{1}_{\{b_{11}+b_{12}-1>0\}}, \mathbb{1}_{\{b_{21}+b_{22}-1<0\}}).$$

- Optimal policy for type 1 is to recommend a_1 if they like a_1 more than they 'dislike' a_2, and recommend a_2 otherwise.
- This is because of a negative reinforcement that can happen if a_1 is not liked or if a_2 is liked.
- S may recommend arms which are not preferred by the user. Example: For $B = (b_{11} = 0.9, b_{12} = 0.3, b_{21} = 0.4, b_{22} = 0.7)$, optimal policy is $(p^* = 1, q^* = 0)$.

Preference Shaping with Known B in D

- Optimal policy is a simple stationary policy

Lemma

The optimal policy for the time slot t is

$$(p_t^*, q_t^*) = \left(\mathbb{1}_{\{b_{11}+b_{12}-1>0\}}, \mathbb{1}_{\{b_{21}+b_{22}-1<0\}}\right).$$

- Optimal policy for type 1 is to recommend a_1 if they like a_1 more than they 'dislike' a_2, and recommend a_2 otherwise.
- This is because of a negative reinforcement that can happen if a_1 is not liked or if a_2 is liked.
- S may recommend arms which are not preferred by the user. Example: For $B = (b_{11} = 0.9, b_{12} = 0.3, b_{21} = 0.4, b_{22} = 0.7)$, optimal policy is $(p^* = 1, q^* = 0)$.

Preference Shaping with Known B in D

- Optimal policy is a simple stationary policy

Lemma

The optimal policy for the time slot t is

$$(p_t^*, q_t^*) = (\mathbb{1}_{\{b_{11}+b_{12}-1>0\}}, \mathbb{1}_{\{b_{21}+b_{22}-1<0\}}).$$

- Optimal policy for type 1 is to recommend a_1 if they like a_1 more than they 'dislike' a_2, and recommend a_2 otherwise.
- This is because of a negative reinforcement that can happen if a_1 is not liked or if a_2 is liked.
- S may recommend arms which are not preferred by the user. Example: For $B = (b_{11} = 0.9, b_{12} = 0.3, b_{21} = 0.4, b_{22} = 0.7)$, optimal policy is $(p^* = 1, q^* = 0)$.

Preference Shaping with Known B in D

- Optimal policy is a simple stationary policy

Lemma

The optimal policy for the time slot t is

$$(p_t^*, q_t^*) = \left(\mathbb{1}_{\{b_{11}+b_{12}-1>0\}}, \mathbb{1}_{\{b_{21}+b_{22}-1<0\}}\right).$$

- Optimal policy for type 1 is to recommend a_1 if they like a_1 more than they 'dislike' a_2, and recommend a_2 otherwise.
- This is because of a negative reinforcement that can happen if a_1 is not liked or if a_2 is liked.
- S may recommend arms which are not preferred by the user. Example: For $B = (b_{11} = 0.9, b_{12} = 0.3, b_{21} = 0.4, b_{22} = 0.7)$, optimal policy is $(p^* = 1, q^* = 0)$.

Preference Shaping with Known B in D

- Optimal policy is a simple stationary policy

Lemma

The optimal policy for the time slot t is

$$(p_t^*, q_t^*) = (\mathbb{1}_{\{b_{11}+b_{12}-1>0\}}, \mathbb{1}_{\{b_{21}+b_{22}-1<0\}}).$$

- Optimal policy for type 1 is to recommend a_1 if they like a_1 more than they 'dislike' a_2, and recommend a_2 otherwise.
- This is because of a negative reinforcement that can happen if a_1 is not liked or if a_2 is liked.
- S may recommend arms which are not preferred by the user. Example: For $B = (b_{11} = 0.9, b_{12} = 0.3, b_{21} = 0.4, b_{22} = 0.7)$, optimal policy is $(p^* = 1, q^* = 0)$.

Preference Shaping with Known B in D

- Optimal policy is a simple stationary policy

Lemma

The optimal policy for the time slot t is

$$(p_t^*, q_t^*) = \left(\mathbb{1}_{\{b_{11}+b_{12}-1>0\}}, \mathbb{1}_{\{b_{21}+b_{22}-1<0\}}\right).$$

- Optimal policy for type 1 is to recommend a_1 if they like a_1 more than they 'dislike' a_2, and recommend a_2 otherwise.
- This is because of a negative reinforcement that can happen if a_1 is not liked or if a_2 is liked.
- S may recommend arms which are not preferred by the user. Example: For $B = (b_{11} = 0.9, b_{12} = 0.3, b_{21} = 0.4, b_{22} = 0.7)$, optimal policy is $(p^* = 1, q^* = 0)$.

Preference Shaping with Known B in D

- Optimal policy is a simple stationary policy

Lemma

The optimal policy for the time slot t is

$$(p_t^*, q_t^*) = \left(\mathbb{1}_{\{b_{11}+b_{12}-1>0\}}, \mathbb{1}_{\{b_{21}+b_{22}-1<0\}}\right).$$

- Optimal policy for type 1 is to recommend a_1 if they like a_1 more than they 'dislike' a_2, and recommend a_2 otherwise.
- This is because of a negative reinforcement that can happen if a_1 is not liked or if a_2 is liked.
- S may recommend arms which are not preferred by the user. Example: For $B = (b_{11} = 0.9, b_{12} = 0.3, b_{21} = 0.4, b_{22} = 0.7)$, optimal policy is $(p^* = 1, q^* = 0)$.

Preference Shaping with Known B in D

- Optimal policy is a simple stationary policy

Lemma

The optimal policy for the time slot t is

$$(p_t^*, q_t^*) = (\mathbb{1}_{\{b_{11}+b_{12}-1>0\}}, \mathbb{1}_{\{b_{21}+b_{22}-1<0\}}).$$

- Optimal policy for type 1 is to recommend a_1 if they like a_1 more than they 'dislike' a_2, and recommend a_2 otherwise.
- This is because of a negative reinforcement that can happen if a_1 is not liked or if a_2 is liked.
- S may recommend arms which are not preferred by the user. Example: For $B = (b_{11} = 0.9, b_{12} = 0.3, b_{21} = 0.4, b_{22} = 0.7)$, optimal policy is $(p^* = 1, q^* = 0)$.

Preference Shaping with Known B in D

- Evolution of $z_1(t)$: Expectation monotonically approaches $d_2/(d_1+d_2)$, the maximum that can be achieved with a policy of the form $(p_t = p, q_t = q)$.

Lemma

For a policy π with $(p_t, q_t) = (p, q)$, the expected proportion of type 1 users at time t is

$$z_1(t) = \frac{d_2}{d_1+d_2} + \left(z_1(0) - \frac{d_2}{d_1+d_2}\right)\left(1+\frac{t}{N_0}\right)^{-(d_1+d_2)}.$$

Here $d_1 = p(1-b_{11}) + (1-p)b_{12}$, $d_2 = q(1-b_{22}) + (1-q)b_{21}$ and $z_1(0)$ is proportion of type 1 users at $t = 0$.

Preference Shaping with Known B in D

- Evolution of $z_1(t)$: Expectation monotonically approaches $d_2/(d_1 + d_2)$, the maximum that can be achieved with a policy of the form $(p_t = p, q_t = q)$.

Lemma

For a policy π with $(p_t, q_t) = (p, q)$, the expected proportion of type 1 users at time t is

$$z_1(t) = \frac{d_2}{d_1 + d_2} + \left(z_1(0) - \frac{d_2}{d_1 + d_2}\right)\left(1 + \frac{t}{N_0}\right)^{-(d_1+d_2)}.$$

Here $d_1 = p(1 - b_{11}) + (1 - p)b_{12}$, $d_2 = q(1 - b_{22}) + (1 - q)b_{21}$ and $z_1(0)$ is proportion of type 1 users at $t = 0$.

Preference Shaping with Known B in D

- The policy also maximizes z_1

Theorem

The optimal policy also maximizes expected asymptotic proportion of type 1 users at $\left(\frac{d_2}{d_1+d_2}\right)$.

- Can write the evolution equations for $Z_1(t)$ and $z_1(t)$ as follows

$$\begin{aligned} Z_1(t+1) &= Z_1(t) + \Delta Z_1(t) \\ &= Z_1(t) + (E[\Delta Z_1(t)|z_1(t)] \\ &\quad +(\Delta Z_1(t) - E[\Delta Z_1(t)|z_1(t)])), \end{aligned}$$

where $E[\Delta Z_1(t)|z_1(t)] = z_1(t)(1-d_1) + (1-z_1(t))d_2$.

$$z_1(t+1) = z_1(t) + \frac{1}{N_0+t+1}(d_2 - (d_1+d_2)z_1(t) + M_t)$$

- All sample paths followed by $z_1(t)$ converge asymptotically almost surely to $z_1 = d_2/(d_1+d_2)$.

Preference Shaping with Known B in D

- The policy also maximizes z_1

Theorem

The optimal policy also maximizes expected asymptotic proportion of type 1 users at $\left(\frac{d_2}{d_1+d_2}\right)$.

- Can write the evolution equations for $Z_1(t)$ and $z_1(t)$ as follows

$$\begin{aligned} Z_1(t+1) &= Z_1(t) + \Delta Z_1(t) \\ &= Z_1(t) + (E[\Delta Z_1(t)|z_1(t)] \\ &\quad + (\Delta Z_1(t) - E[\Delta Z_1(t)|z_1(t)])), \end{aligned}$$

where $E[\Delta Z_1(t)|z_1(t)] = z_1(t)(1-d_1) + (1-z_1(t))d_2$.

$$z_1(t+1) = z_1(t) + \frac{1}{N_0+t+1}(d_2 - (d_1+d_2)z_1(t) + M_t)$$

- All sample paths followed by $z_1(t)$ converge asymptotically almost surely to $z_1 = d_2/(d_1+d_2)$.

Preference Shaping with Known B in D

- The policy also maximizes z_1

Theorem

The optimal policy also maximizes expected asymptotic proportion of type 1 users at $\left(\frac{d_2}{d_1+d_2}\right)$.

- Can write the evolution equations for $Z_1(t)$ and $z_1(t)$ as follows

$$\begin{aligned} Z_1(t+1) &= Z_1(t) + \Delta Z_1(t) \\ &= Z_1(t) + (E[\Delta Z_1(t)|z_1(t)] \\ &\qquad + (\Delta Z_1(t) - E[\Delta Z_1(t)|z_1(t)])), \end{aligned}$$

where $E[\Delta Z_1(t)|z_1(t)] = z_1(t)(1-d_1) + (1-z_1(t))d_2$.

$$z_1(t+1) = z_1(t) + \frac{1}{N_0+t+1}(d_2 - (d_1+d_2)z_1(t) + M_t)$$

- All sample paths followed by $z_1(t)$ converge asymptotically almost surely to $z_1 = d_2/(d_1+d_2)$.

Preference Shaping with Known B in D

- The policy also maximizes z_1

Theorem

The optimal policy also maximizes expected asymptotic proportion of type 1 users at $\left(\frac{d_2}{d_1+d_2}\right)$.

- Can write the evolution equations for $Z_1(t)$ and $z_1(t)$ as follows

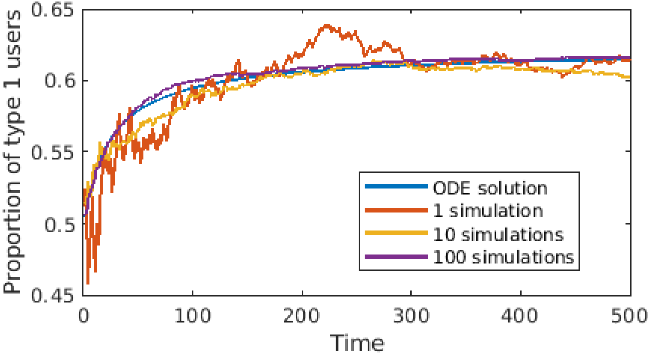
$$\begin{aligned} Z_1(t+1) &= Z_1(t) + \Delta Z_1(t) \\ &= Z_1(t) + (E[\Delta Z_1(t)|z_1(t)] \\ &\quad +(\Delta Z_1(t) - E[\Delta Z_1(t)|z_1(t)])), \end{aligned}$$

where $E[\Delta Z_1(t)|z_1(t)] = z_1(t)(1-d_1) + (1-z_1(t))d_2$.

$$z_1(t+1) = z_1(t) + \frac{1}{N_0+t+1}(d_2 - (d_1+d_2)z_1(t) + M_t)$$

- All sample paths followed by $z_1(t)$ converge asymptotically almost surely to $z_1 = d_2/(d_1+d_2)$.

Preference Shaping with Known B in DID



Trajectory $z_1(t)$. Comparison of o.d.e. solution and averages from 1, 10 amd 100 sample paths. B is $b_{00} = 0.9$, $b_{01} = 0.4$, $b_{10} = 0.2$, and $b_{11} = 0.6$.

Explore-Then-Commit

- Usual method: Explore by playing each arm uniformly for time m and estimate rewards matrix B. Exploit for time $T - m$.
- General analysis is elusive; consider the special case of $b_{11} = b_{22}$ and $b_{12} = b_{21}$.
- ETC is inherently inefficient because, exploitation is during the most plastic phase.
- Do not know how to obtain m

Preference Shaping with Unknown B

Explore-Then-Commit

- Usual method: Explore by playing each arm uniformly for time m and estimate rewards matrix B. Exploit for time $T - m$.
- General analysis is elusive; consider the special case of $b_{11} = b_{22}$ and $b_{12} = b_{21}$.
- ETC is inherently inefficient because, exploitation is during the most plastic phase.
- Do not know how to obtain m

Preference Shaping with Unknown B

Explore-Then-Commit

- Usual method: Explore by playing each arm uniformly for time m and estimate rewards matrix B. Exploit for time $T - m$.
- General analysis is elusive; consider the special case of $b_{11} = b_{22}$ and $b_{12} = b_{21}$.

Theorem

$$R^{ETC}_{[1:T]} \leq m\Delta_1/2 + (T-m)\Delta_1 e^{-m\Delta_1^2/8} \tag{4}$$

Further, using $m = 8\log(T)/\Delta_1^2$ (to bring the regret bound in terms of T and eliminate m), we get a logarithmic regret, i.e.,

$$R_{ETC} \leq \frac{4}{\Delta_1}\log(T) + \mathcal{O}(1/T). \tag{5}$$

- ETC is inherently inefficient because, exploitation is during the most plastic phase.
- Do not know how to obtain m

Preference Shaping with Unknown B

Explore-Then-Commit

- Usual method: Explore by playing each arm uniformly for time m and estimate rewards matrix B. Exploit for time $T - m$.
- General analysis is elusive; consider the special case of $b_{11} = b_{22}$ and $b_{12} = b_{21}$.

Theorem

$$R_{[1:T]}^{ETC} \leq m\Delta_1/2 + (T - m)\Delta_1 e^{-m\Delta_1^2/8} \tag{4}$$

Further, using $m = 8\log(T)/\Delta_1^2$ (to bring the regret bound in terms of T and eliminate m), we get a logarithmic regret, i.e.,

$$R_{ETC} \leq \frac{4}{\Delta_1}\log(T) + \mathcal{O}(1/T). \tag{5}$$

- ETC is inherently inefficient because, exploitation is during the most plastic phase.
- Do not know how to obtain m

Preference Shaping with Unknown B

Explore-Then-Commit

- Usual method: Explore by playing each arm uniformly for time m and estimate rewards matrix B. Exploit for time $T - m$.
- General analysis is elusive; consider the special case of $b_{11} = b_{22}$ and $b_{12} = b_{21}$.

Theorem

$$R^{ETC}_{[1:T]} \leq m\Delta_1/2 + (T - m)\Delta_1 e^{-m\Delta_1^2/8} \tag{4}$$

Further, using $m = 8\log(T)/\Delta_1^2$ (to bring the regret bound in terms of T and eliminate m), we get a logarithmic regret, i.e.,

$$R_{ETC} \leq \frac{4}{\Delta_1}\log(T) + \mathcal{O}(1/T). \tag{5}$$

- ETC is inherently inefficient because, exploitation is during the most plastic phase.
- Do not know how to obtain m

Preference Shaping with Unknown B

Explore-Then-Commit

- Usual method: Explore by playing each arm uniformly for time m and estimate rewards matrix B. Exploit for time $T - m$.
- General analysis is elusive; consider the special case of $b_{11} = b_{22}$ and $b_{12} = b_{21}$.

Theorem

$$R^{ETC}_{[1:T]} \leq m\Delta_1/2 + (T - m)\Delta_1 e^{-m\Delta_1^2/8} \tag{4}$$

Further, using $m = 8\log(T)/\Delta_1^2$ (to bring the regret bound in terms of T and eliminate m), we get a logarithmic regret, i.e.,

$$R_{ETC} \leq \frac{4}{\Delta_1}\log(T) + \mathcal{O}(1/T). \tag{5}$$

- ETC is inherently inefficient because, exploitation is during the most plastic phase.
- Do not know how to obtain m

Preference Shaping with Unknown B

Thompson sampling

1. Initialize $\alpha_{ij} = 1$, $\beta_{ij} = 1$ for all $i, j \in \{1, 2\}$
2. At time step t,
    1. Let i type of user
    2. Sample $\tilde{b}_{ij} \sim Beta(\alpha_{ij}, \beta_{ij})$ for all $i, j \in \{1, 2\}$
    3. If $i == 1$ then show arm 1 w.p. $\mathbb{1}_{\{\tilde{b}_{11}+\tilde{b}_{12}-1>0\}}$, else show arm 2
    4. If $i == 0$ then show arm 1 w.p. $\mathbb{1}_{\{\tilde{b}_{22}+\tilde{b}_{21}-1<0\}}$, else show arm 2
    5. $j =$Arm showed; $R_t =$ Reward obtained;
    6. $\alpha_{ij} = \alpha_{ij} + R_t$; $\beta_{ij} = \beta_{ij} + (1 - R_t)$.

Preference Shaping with Unknown B

Thompson sampling

1. Initialize $\alpha_{ij} = 1$, $\beta_{ij} = 1$ for all $i, j \in \{1, 2\}$
2. At time step t,
    1. Let i type of user
    2. Sample $\tilde{b}_{ij} \sim Beta(\alpha_{ij}, \beta_{ij})$ for all $i, j \in \{1, 2\}$
    3. If $i == 1$ then show arm 1 w.p. $\mathbb{1}_{\{\tilde{b}_{11}+\tilde{b}_{12}-1>0\}}$, else show arm 2
    4. If $i == 0$ then show arm 1 w.p. $\mathbb{1}_{\{\tilde{b}_{22}+\tilde{b}_{21}-1<0\}}$, else show arm 2
    5. $j =$Arm showed; $R_t =$ Reward obtained;
    6. $\alpha_{ij} = \alpha_{ij} + R_t$; $\beta_{ij} = \beta_{ij} + (1 - R_t)$.

Preference Shaping with Unknown B

Thompson sampling

1. Initialize $\alpha_{ij} = 1$, $\beta_{ij} = 1$ for all $i, j \in \{1, 2\}$
2. At time step t,
    1. Let i type of user
    2. Sample $\tilde{b}_{ij} \sim Beta(\alpha_{ij}, \beta_{ij})$ for all $i, j \in \{1, 2\}$
    3. If $i == 1$ then show arm 1 w.p. $\mathbb{1}_{\{\tilde{b}_{11}+\tilde{b}_{12}-1>0\}}$, else show arm 2
    4. If $i == 0$ then show arm 1 w.p. $\mathbb{1}_{\{\tilde{b}_{22}+\tilde{b}_{21}-1<0\}}$, else show arm 2
    5. $j =$Arm showed; $R_t =$ Reward obtained;
    6. $\alpha_{ij} = \alpha_{ij} + R_t$; $\beta_{ij} = \beta_{ij} + (1 - R_t)$.

Preference Shaping with Unknown B

Thompson sampling

1. Initialize $\alpha_{ij} = 1$, $\beta_{ij} = 1$ for all $i, j \in \{1, 2\}$
2. At time step t,
    1. Let i type of user
    2. Sample $\tilde{b}_{ij} \sim Beta(\alpha_{ij}, \beta_{ij})$ for all $i, j \in \{1, 2\}$
    3. If $i == 1$ then show arm 1 w.p. $\mathbb{1}_{\{\tilde{b}_{11}+\tilde{b}_{12}-1>0\}}$, else show arm 2
    4. If $i == 0$ then show arm 1 w.p. $\mathbb{1}_{\{\tilde{b}_{22}+\tilde{b}_{21}-1<0\}}$, else show arm 2
    5. $j =$Arm showed; $R_t =$ Reward obtained;
    6. $\alpha_{ij} = \alpha_{ij} + R_t$; $\beta_{ij} = \beta_{ij} + (1 - R_t)$.

Thompson sampling

1. Initialize $\alpha_{ij} = 1$, $\beta_{ij} = 1$ for all $i, j \in \{1, 2\}$
2. At time step t,
   1. Let i type of user
   2. Sample $\tilde{b}_{ij} \sim Beta(\alpha_{ij}, \beta_{ij})$ for all $i, j \in \{1, 2\}$
   3. If $i == 1$ then show arm 1 w.p. $\mathbb{1}_{\{\tilde{b}_{11}+\tilde{b}_{12}-1>0\}}$, else show arm 2
   4. If $i == 0$ then show arm 1 w.p. $\mathbb{1}_{\{\tilde{b}_{22}+\tilde{b}_{21}-1<0\}}$, else show arm 2
   5. $j =$ Arm showed; $R_t =$ Reward obtained;
   6. $\alpha_{ij} = \alpha_{ij} + R_t$; $\beta_{ij} = \beta_{ij} + (1 - R_t)$.

Preference Shaping with Unknown B

Thompson sampling

1. Initialize $\alpha_{ij} = 1$, $\beta_{ij} = 1$ for all $i, j \in \{1, 2\}$
2. At time step t,
    1. Let i type of user
    2. Sample $\tilde{b}_{ij} \sim Beta(\alpha_{ij}, \beta_{ij})$ for all $i, j \in \{1, 2\}$
    3. If $i == 1$ then show arm 1 w.p. $\mathbb{1}_{\{\tilde{b}_{11}+\tilde{b}_{12}-1>0\}}$, else show arm 2
    4. If $i == 0$ then show arm 1 w.p. $\mathbb{1}_{\{\tilde{b}_{22}+\tilde{b}_{21}-1<0\}}$, else show arm 2
    5. j =Arm showed; R_t = Reward obtained;
    6. $\alpha_{ij} = \alpha_{ij} + R_t$; $\beta_{ij} = \beta_{ij} + (1 - R_t)$.

Preference Shaping with Unknown B

Thompson sampling

1. Initialize $\alpha_{ij} = 1$, $\beta_{ij} = 1$ for all $i, j \in \{1, 2\}$
2. At time step t,
    1. Let i type of user
    2. Sample $\tilde{b}_{ij} \sim Beta(\alpha_{ij}, \beta_{ij})$ for all $i, j \in \{1, 2\}$
    3. If $i == 1$ then show arm 1 w.p. $\mathbb{1}_{\{\tilde{b}_{11}+\tilde{b}_{12}-1>0\}}$, else show arm 2
    4. If $i == 0$ then show arm 1 w.p. $\mathbb{1}_{\{\tilde{b}_{22}+\tilde{b}_{21}-1<0\}}$, else show arm 2
    5. $j =$Arm showed; $R_t =$ Reward obtained;
    6. $\alpha_{ij} = \alpha_{ij} + R_t$; $\beta_{ij} = \beta_{ij} + (1 - R_t)$.

Thompson sampling

1. Initialize $\alpha_{ij} = 1$, $\beta_{ij} = 1$ for all $i, j \in \{1, 2\}$
2. At time step t,
    1. Let i type of user
    2. Sample $\tilde{b}_{ij} \sim Beta(\alpha_{ij}, \beta_{ij})$ for all $i, j \in \{1, 2\}$
    3. If $i == 1$ then show arm 1 w.p. $\mathbb{1}_{\{\tilde{b}_{11}+\tilde{b}_{12}-1>0\}}$, else show arm 2
    4. If $i == 0$ then show arm 1 w.p. $\mathbb{1}_{\{\tilde{b}_{22}+\tilde{b}_{21}-1<0\}}$, else show arm 2
    5. $j =$Arm showed; $R_t =$ Reward obtained;
    6. $\alpha_{ij} = \alpha_{ij} + R_t$; $\beta_{ij} = \beta_{ij} + (1 - R_t)$.

Thompson Sampling

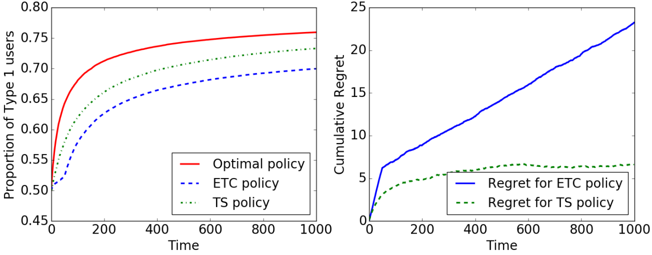
Theorem

Cumulative regret for the Thompson sampling policy is bounded above by

$$R_{[1:T]}^{Thomp} \leq \frac{(z^*)^2}{4}\left(\frac{1}{f_1(1-f_1)\Delta_1}+\frac{1}{f_2(1-f_2)\Delta_2}\right)\log(T). \tag{6}$$

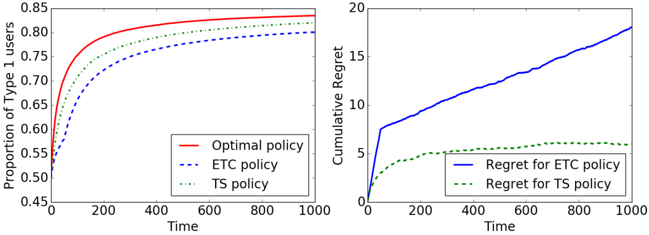
z^* *is asymptotic proportion from optimal policy;* $f_1, f_2 < 1$ *are constants*

Numerical Results



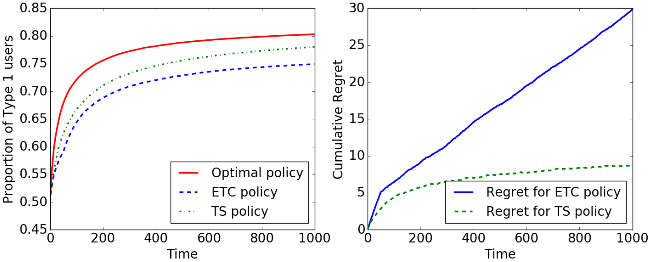
Expected population proportion vs time (left) and cumulative regret vs time (right) for the ETC, TS, and the optimal policy that knows B.
$B_1 = (b_{11} = 0.9, b_{12} = 0.4, b_{21} = 0.2, b_{22} = 0.6)$. Optimal policy is $(p = 1, q = 1)$.

Numerical Results



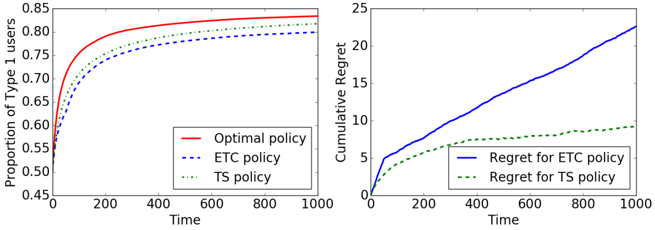
Expected population proportion vs time (left) and cumulative regret vs time (right) for the ETC, TS, and the optimal policy that knows B.
$B_2 = (b_{11} = 0.9, b_{12} = 0.4, b_{21} = 0.6, b_{22} = 0.7)$. Optimal policy is $(p = 1, q = 0)$.

Numerical Results



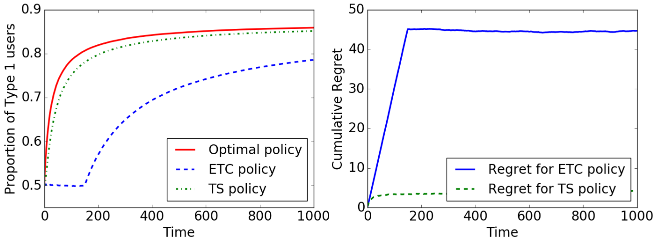
Expected population proportion vs time (left) and cumulative regret vs time (right) for the ETC, TS, and the optimal policy that knows B.
$B_3 = (b_{11} = 0.7, b_{12} = 0.1, b_{21} = 0.3, b_{22} = 0.5)$. Optimal policy is $(p = 0, q = 1)$.

Numerical Results



Expected population proportion vs time (left) and cumulative regret vs time (right) for the ETC, TS, and the optimal policy that knows B.
$B_4 = (b_{11} = 0.7, b_{12} = 0.1, b_{21} = 0.6, b_{22} = 0.6)$. Optimal policy is $(p = 0, q = 0)$.

Numerical Results



Expected population proportion vs time (left) and cumulative regret vs time (right) for the ETC, TS, and the optimal for $B_{sym} = (b_{11} = 0.9, b_{12} = 0.7, b_{21} = 0.7, b_{22} = 0.9)$

Constant Influence Dynamics Model

Results

- Recall that there is a difference in the population evolution: Ball of color i flips if type i gets reward when shown arm A_{-i} OR if it gets reward 0 when shown arm A_i.

Lemma

For a policy π such that $(p_t, q_t) = (p, q)$,

$$z_1(t) = \frac{d_2}{d_1 + d_2} + \left(z_1^0 - \frac{d_2}{d_1 + d_2} \right) e^{-t \frac{d_1 + d_2}{N_0}}.$$

Here $d_1 = p(1 - b_{11}) + (1 - p)b_{12}$, $d_2 = q(1 - b_{22}) + (1 - q)b_{21}$ and $z_1(0)$ is the initial proportion of type 1 users.

- With a fixed (p, q), the asymptotic fraction is the same as in the Decreasing Influence model; but rate is not the same.
- For $b_{11} = b_{22}$ and $b_{12} = b_{21}$, all the results from the DID model hold with no change. Rather surprising because the two-fold tradeoff of DID does not seem to have caused it additional damage!

Results

- Recall that there is a difference in the population evolution: Ball of color i flips if type i gets reward when shown arm A_{-i} OR if it gets reward 0 when shown arm A_i.

Lemma

For a policy π such that $(p_t, q_t) = (p, q)$,

$$z_1(t) = \frac{d_2}{d_1 + d_2} + \left(z_1^0 - \frac{d_2}{d_1 + d_2} \right) e^{-t \frac{d_1 + d_2}{N_0}}.$$

Here $d_1 = p(1 - b_{11}) + (1 - p)b_{12}$, $d_2 = q(1 - b_{22}) + (1 - q)b_{21}$ and $z_1(0)$ is the initial proportion of type 1 users.

- With a fixed (p, q), the asymptotic fraction is the same as in the Decreasing Influence model; but rate is not the same.
- For $b_{11} = b_{22}$ and $b_{12} = b_{21}$, all the results from the DID model hold with no change. Rather surprising because the two-fold tradeoff of DID does not seem to have caused it additional damage!

Results

- Recall that there is a difference in the population evolution: Ball of color i flips if type i gets reward when shown arm A_{-i} OR if it gets reward 0 when shown arm A_i.

Lemma

For a policy π such that $(p_t, q_t) = (p, q)$,

$$z_1(t) = \frac{d_2}{d_1 + d_2} + \left(z_1^0 - \frac{d_2}{d_1 + d_2} \right) e^{-t \frac{d_1 + d_2}{N_0}}.$$

Here $d_1 = p(1 - b_{11}) + (1 - p)b_{12}$, $d_2 = q(1 - b_{22}) + (1 - q)b_{21}$ and $z_1(0)$ is the initial proportion of type 1 users.

- With a fixed (p, q), the asymptotic fraction is the same as in the Decreasing Influence model; but rate is not the same.
- For $b_{11} = b_{22}$ and $b_{12} = b_{21}$, all the results from the DID model hold with no change. Rather surprising because the two-fold tradeoff of DID does not seem to have caused it additional damage!

Results

- Recall that there is a difference in the population evolution: Ball of color i flips if type i gets reward when shown arm A_{-i} OR if it gets reward 0 when shown arm A_i.

Lemma

For a policy π such that $(p_t, q_t) = (p, q)$,

$$z_1(t) = \frac{d_2}{d_1 + d_2} + \left(z_1^0 - \frac{d_2}{d_1 + d_2}\right) e^{-t\frac{d_1+d_2}{N_0}}.$$

Here $d_1 = p(1 - b_{11}) + (1 - p)b_{12}$, $d_2 = q(1 - b_{22}) + (1 - q)b_{21}$ and $z_1(0)$ is the initial proportion of type 1 users.

- With a fixed (p, q), the asymptotic fraction is the same as in the Decreasing Influence model; but rate is not the same.
- For $b_{11} = b_{22}$ and $b_{12} = b_{21}$, all the results from the DID model hold with no change. Rather surprising because the two-fold tradeoff of DID does not seem to have caused it additional damage!

Results

- Recall that there is a difference in the population evolution: Ball of color i flips if type i gets reward when shown arm A_{-i} OR if it gets reward 0 when shown arm A_i.

Lemma

For a policy π such that $(p_t, q_t) = (p, q)$,

$$z_1(t) = \frac{d_2}{d_1 + d_2} + \left(z_1^0 - \frac{d_2}{d_1 + d_2} \right) e^{-t \frac{d_1 + d_2}{N_0}}.$$

Here $d_1 = p(1 - b_{11}) + (1 - p)b_{12}$, $d_2 = q(1 - b_{22}) + (1 - q)b_{21}$ and $z_1(0)$ is the initial proportion of type 1 users.

- With a fixed (p, q), the asymptotic fraction is the same as in the Decreasing Influence model; but rate is not the same.
- For $b_{11} = b_{22}$ and $b_{12} = b_{21}$, all the results from the DID model hold with no change. Rather surprising because the two-fold tradeoff of DID does not seem to have caused it additional damage!

Results

- Recall that there is a difference in the population evolution: Ball of color i flips if type i gets reward when shown arm A_{-i} OR if it gets reward 0 when shown arm A_i.

Lemma

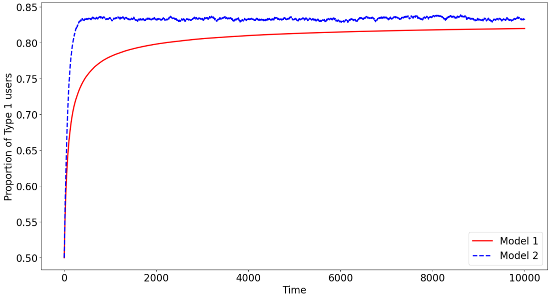
For a policy π such that $(p_t, q_t) = (p, q)$,

$$z_1(t) = \frac{d_2}{d_1 + d_2} + \left(z_1^0 - \frac{d_2}{d_1 + d_2}\right) e^{-t\frac{d_1+d_2}{N_0}}.$$

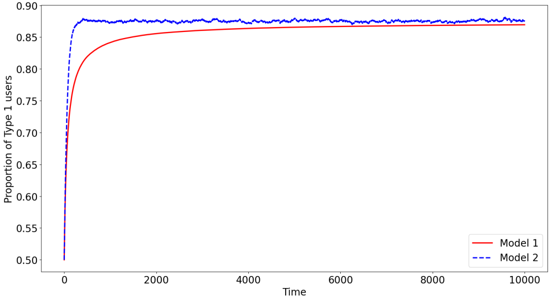
Here $d_1 = p(1 - b_{11}) + (1 - p)b_{12}$, $d_2 = q(1 - b_{22}) + (1 - q)b_{21}$ and $z_1(0)$ is the initial proportion of type 1 users.

- With a fixed (p, q), the asymptotic fraction is the same as in the Decreasing Influence model; but rate is not the same.
- For $b_{11} = b_{22}$ and $b_{12} = b_{21}$, all the results from the DID model hold with no change. Rather surprising because the two-fold tradeoff of DID does not seem to have caused it additional damage!

Numerical Results



Expected population proportion vs time for optimal policies that knows B. for Model 1 and Model 2.
$B_1 = (b_{11} = 0.7, b_{12} = 0.1, b_{21} = 0.2, b_{22} = 0.5)$. Optimal policy is $(p = 0, q = 0)$.



Expected population proportion vs time for optimal policies that knows B. for Model 1 and Model 2.
$B_2 = (b_{11} = 0.9, b_{12} = 0.7, b_{21} = 0.7, b_{22} = 0.9)$. Optimal policy is $(p = 1, q = 0)$.

Extensions and way forward

- Generalises to N arms and N user types. Optimal policies for some natural generalisations can be defined and analysed.
- Can have two competing systems influencing in opposite directions.
  - Here the users also have a choice of the RS that they choose and one needs to define the such a matrix P.
  - Equilibrium, not surprisingly, outcome depends on the matrix B (polarised population or a uniform population), and on P.
  - Simulation results show interesting behavour!

Extensions and way forward

- Generalises to N arms and N user types. Optimal policies for some natural generalisations can be defined and analysed.
- Can have two competing systems influencing in opposite directions.
  - Here the users also have a choice of the RS that they choose and one needs to define the such a matrix P.
  - Equilibrium, not surprisingly, outcome depends on the matrix B (polarised population or a uniform population), and on P.
  - Simulation results show interesting behavour!

Extensions and way forward

- Generalises to N arms and N user types. Optimal policies for some natural generalisations can be defined and analysed.
- Can have two competing systems influencing in opposite directions.
  - Here the users also have a choice of the RS that they choose and one needs to define the such a matrix P.
  - Equilibrium, not surprisingly, outcome depends on the matrix B (polarised population or a uniform population), and on P.
  - Simulation results show interesting behaviour!

Extensions and way forward

- Generalises to N arms and N user types. Optimal policies for some natural generalisations can be defined and analysed.
- Can have two competing systems influencing in opposite directions.
    - Here the users also have a choice of the RS that they choose and one needs to define the such a matrix P.
    - Equilibrium, not surprisingly, outcome depends on the matrix B (polarised population or a uniform population), and on P.
    - Simulation results show interesting behavour!

Extensions and way forward

- Generalises to N arms and N user types. Optimal policies for some natural generalisations can be defined and analysed.
- Can have two competing systems influencing in opposite directions.
  - Here the users also have a choice of the RS that they choose and one needs to define the such a matrix P.
  - Equilibrium, not surprisingly, outcome depends on the matrix B (polarised population or a uniform population), and on P.
  - Simulation results show interesting behaviour!

Concluding remarks

- This strand of work is motivated by the belief that algorithms that learn population preferences also influence the prefences, either in a transient or in a permanent manner, perhaps benignly.
- The effect of this influence is not captured well in the models that I am famililar with.
- In other work:
  - Model the user state-of-mind with respect to an arm as a two-state Markov chain; gives rise to a restless bandit model and could use the Whittle index based policies when the parameters are known.
  - Could also consider a deterministic model for the preference evolution and design optimal bandit algorithms.
- **Holy grail:** A general model to describe the interaction between the learning algorithm and the '*state of mind* of the population' that the learning does not account for.

Concluding remarks

- This strand of work is motivated by the belief that algorithms that learn population preferences also influence the prefences, either in a transient or in a permanent manner, perhaps benignly.
- The effect of this influence is not captured well in the models that I am famililar with.
- In other work:
  - Modeled the user state-of-mind with respect to an arm as a two-state Markov chain; gives rise to a restless bandit model and could derive Whittle index based policies when the parameters are known.
  - Could also consider a deterministic model for the preference evolution and design optimal bandit algorithms.
- **Holy grail:** A general model to describe the interaction between the learning algorithm and the '*state of mind* of the population' that the learning does not account for.

Concluding remarks

- This strand of work is motivated by the belief that algorithms that learn population preferences also influence the prefences, either in a transient or in a permanent manner, perhaps benignly.
- The effect of this influence is not captured well in the models that I am famililar with.
- In other work:
  - Modeled the user state-of-mind with respect to an arm as a two-state Markov chain, gives rise to a restless bandit model and could determine Whittle index based policies when the parameters are known;
  - Could also consider a deterministic model for the preference evolution and design optimal bandit algorithms.
- **Holy grail:** A general model to describe the interaction between the learning algorithm and the '*state of mind* of the population' that the learning does not account for.

Concluding remarks

- This strand of work is motivated by the belief that algorithms that learn population preferences also influence the prefences, either in a transient or in a permanent manner, perhaps benignly.
- The effect of this influence is not captured well in the models that I am famililar with.
- In other work:
  - Modeled the user **state-of-mind** with respect to an arm as a two-state Markov chain.
- **Holy grail:** A general model to describe the interaction between the learning algorithm and the ‘*state of mind* of the population’ that the learning does not account for.

Concluding remarks

- This strand of work is motivated by the belief that algorithms that learn population preferences also influence the prefences, either in a transient or in a permanent manner, perhaps benignly.
- The effect of this influence is not captured well in the models that I am famililar with.
- In other work:
  - Modeled the user **state-of-mind** with respect to an arm as a two-state Markov chain. gives rise to a restless bandit model and could determine Whittle index based policies when the parameters are known;
  - Could also consider a deterministic model for the preference evolution and design optimal bandit algorithms
- **Holy grail:** A general model to describe the interaction between the learning algorithm and the '*state of mind* of the population' that the learning does not account for.

Concluding remarks

- This strand of work is motivated by the belief that algorithms that learn population preferences also influence the prefences, either in a transient or in a permanent manner, perhaps benignly.
- The effect of this influence is not captured well in the models that I am famililar with.
- In other work:
  - Modeled the user **state-of-mind** with respect to an arm as a two-state Markov chain. gives rise to a restless bandit model and could determine Whittle index based policies when the parameters are known;
  - Could also consider a deterministic model for the preference evolution and design optimal bandit algorithms
- **Holy grail:** A general model to describe the interaction between the learning algorithm and the '*state of mind* of the population' that the learning does not account for.

Concluding remarks

- This strand of work is motivated by the belief that algorithms that learn population preferences also influence the prefences, either in a transient or in a permanent manner, perhaps benignly.
- The effect of this influence is not captured well in the models that I am famililar with.
- In other work:
    - Modeled the user **state-of-mind** with respect to an arm as a two-state Markov chain. gives rise to a restless bandit model and could determine Whittle index based policies when the parameters are known;
    - Could also consider a deterministic model for the preference evolution and design optimal bandit algorithms
- **Holy grail:** A general model to describe the interaction between the learning algorithm and the '*state of mind* of the population' that the learning does not account for.

Concluding remarks

- This strand of work is motivated by the belief that algorithms that learn population preferences also influence the prefences, either in a transient or in a permanent manner, perhaps benignly.
- The effect of this influence is not captured well in the models that I am famililar with.
- In other work:
  - Modeled the user **state-of-mind** with respect to an arm as a two-state Markov chain. gives rise to a restless bandit model and could determine Whittle index based policies when the parameters are known;
  - Could also consider a deterministic model for the preference evolution and design optimal bandit algorithms
- **Holy grail:** A general model to describe the interaction between the learning algorithm and the '*state of mind* of the population' that the learning does not account for.

Concluding remarks

- This strand of work is motivated by the belief that algorithms that learn population preferences also influence the prefences, either in a transient or in a permanent manner, perhaps benignly.
- The effect of this influence is not captured well in the models that I am famililar with.
- In other work:
  - Modeled the user **state-of-mind** with respect to an arm as a two-state Markov chain. gives rise to a restless bandit model and could determine Whittle index based policies when the parameters are known;
  - Could also consider a deterministic model for the preference evolution and design optimal bandit algorithms
- **Holy grail:** A general model to describe the interaction between the learning algorithm and the '*state of mind* of the population' that the learning does not account for.