NONPARAMETRIC STOCHASTIC BANDITS

András Antos

BMGE, MIT, Intelligent Data Analysis, Nov 19, 2013

OUTLINE

OUTLINE

OUTLINE

OUTLINE

OUTLINE

OUTLINE

OUTLINE

OUTLINE

OUTLINE

BANDITS

- Y_{kt} – payoff of arm k when chosen the tth time, $1 \leq k \leq K$
- For k fixed, Y_{kt} is an i.i.d. sequence
- $\mu_k = \mathbb{E}[Y_{kt}]$
- $\mu^* = \max_k \mu_k$
- For $k \neq k'$, Y_{kt} and $Y_{k't'}$ are independent
- $J_{\text{bad}} = \{k \,|\, \mu_k < \mu^*\}$, set of "bad" arms
- $J_{\text{good}} = \{k \,|\, \mu_k = \mu^*\}$, set of "good" arms
- $I_t^{\mathcal{A}}$ – choice of arm at time t by some allocation rule $\mathcal{A}$
- $T_{kt}^{\mathcal{A}} = \sum_{s=1}^{t} \mathbb{I}_{\{I_s^{\mathcal{A}}=k\}}$ (# of choosing k)

We shall drop $\mathcal{A}$ from $I_t^{\mathcal{A}}$, $T_{kt}^{\mathcal{A}}$ when $\mathcal{A}$ is unambigous $\rightarrow I_t$, T_{kt}

BANDITS

- Y_{kt} – payoff of arm k when chosen the tth time, $1 \leq k \leq K$
- For k fixed, Y_{kt} is an i.i.d. sequence
- $\mu_k = \mathbb{E}[Y_{kt}]$
- $\mu^* = \max_k \mu_k$
- For $k \neq k'$, Y_{kt} and $Y_{k't'}$ are independent
- $J_{\mathrm{bad}} = \{k \,|\, \mu_k < \mu^*\}$, set of "bad" arms
- $J_{\mathrm{good}} = \{k \,|\, \mu_k = \mu^*\}$, set of "good" arms
- $I_t^{\mathcal{A}}$ – choice of arm at time t by some allocation rule $\mathcal{A}$
- $T_{kt}^{\mathcal{A}} = \sum_{s=1}^{t} \mathbb{I}_{\{I_s^{\mathcal{A}}=k\}}$ (# of choosing k)

We shall drop $\mathcal{A}$ from $I_t^{\mathcal{A}}$, $T_{kt}^{\mathcal{A}}$ when $\mathcal{A}$ is unambigous $\rightarrow I_t$, T_{kt}

BANDITS

- Y_{kt} – payoff of arm k when chosen the tth time, $1 \leq k \leq K$
- For k fixed, Y_{kt} is an i.i.d. sequence
- $\mu_k = \mathbb{E}[Y_{kt}]$
- $\mu^* = \max_k \mu_k$
- For $k \neq k'$, Y_{kt} and $Y_{k't'}$ are independent
- $J_{\mathrm{bad}} = \{k \,|\, \mu_k < \mu^*\}$, set of "bad" arms
- $J_{\mathrm{good}} = \{k \,|\, \mu_k = \mu^*\}$, set of "good" arms
- $I_t^{\mathcal{A}}$ – choice of arm at time t by some allocation rule $\mathcal{A}$
- $T_{kt}^{\mathcal{A}} = \sum_{s=1}^{t} \mathbb{I}_{\{I_s^{\mathcal{A}}=k\}}$ (# of choosing k)

We shall drop $\mathcal{A}$ from $I_t^{\mathcal{A}}$, $T_{kt}^{\mathcal{A}}$ when $\mathcal{A}$ is unambigous $\rightarrow I_t$, T_{kt}

BANDITS

- Y_{kt} – payoff of arm k when chosen the tth time, $1 \leq k \leq K$
- For k fixed, Y_{kt} is an i.i.d. sequence
- $\mu_k = \mathbb{E}\left[Y_{kt}\right]$
- $\mu^* = \max_k \mu_k$
- For $k \neq k'$, Y_{kt} and $Y_{k't'}$ are independent
- $J_{\mathrm{bad}} = \{k \,|\, \mu_k < \mu^*\}$, set of "bad" arms
- $J_{\mathrm{good}} = \{k \,|\, \mu_k = \mu^*\}$, set of "good" arms
- $I_t^{\mathcal{A}}$ – choice of arm at time t by some allocation rule $\mathcal{A}$
- $T_{kt}^{\mathcal{A}} = \sum_{s=1}^{t} \mathbb{I}_{\{I_s^{\mathcal{A}}=k\}}$ (# of choosing k)

We shall drop $\mathcal{A}$ from $I_t^{\mathcal{A}}$, $T_{kt}^{\mathcal{A}}$ when $\mathcal{A}$ is unambigous $\rightarrow I_t$, T_{kt}

BANDITS

- Y_{kt} – payoff of arm k when chosen the tth time, $1 \le k \le K$
- For k fixed, Y_{kt} is an i.i.d. sequence
- $\mu_k = \mathbb{E}[Y_{kt}]$
- $\mu^* = \max_k \mu_k$
- For $k \neq k'$, Y_{kt} and $Y_{k't'}$ are independent
- $J_{\text{bad}} = \{k \,|\, \mu_k < \mu^*\}$, set of “bad” arms
- $J_{\text{good}} = \{k \,|\, \mu_k = \mu^*\}$, set of “good” arms
- $I_t^{\mathcal{A}}$ – choice of arm at time t by some allocation rule $\mathcal{A}$
- $T_{kt}^{\mathcal{A}} = \sum_{s=1}^{t} \mathbb{I}_{\{I_s^{\mathcal{A}}=k\}}$ (# of choosing k)

We shall drop $\mathcal{A}$ from $I_t^{\mathcal{A}}$, $T_{kt}^{\mathcal{A}}$ when $\mathcal{A}$ is unambigous $\rightarrow I_t$, T_{kt}

Bandits

- Y_{kt} – payoff of arm k when chosen the tth time, $1 \leq k \leq K$
- For k fixed, Y_{kt} is an i.i.d. sequence
- $\mu_k = \mathbb{E}[Y_{kt}]$
- $\mu^* = \max_k \mu_k$
- For $k \neq k'$, Y_{kt} and $Y_{k't'}$ are independent
- $J_{\text{bad}} = \{k \,|\, \mu_k < \mu^*\}$, set of "bad" arms
- $J_{\text{good}} = \{k \,|\, \mu_k = \mu^*\}$, set of "good" arms
- $I_t^{\mathcal{A}}$ – choice of arm at time t by some allocation rule $\mathcal{A}$
- $T_{kt}^{\mathcal{A}} = \sum_{s=1}^{t} \mathbb{I}_{\{I_s^{\mathcal{A}}=k\}}$ (# of choosing k)

We shall drop $\mathcal{A}$ from $I_t^{\mathcal{A}}$, $T_{kt}^{\mathcal{A}}$ when $\mathcal{A}$ is unambigous $\rightarrow I_t$, T_{kt}

BANDITS

- Y_{kt} – payoff of arm k when chosen the tth time, $1 \leq k \leq K$
- For k fixed, Y_{kt} is an i.i.d. sequence
- $\mu_k = \mathbb{E}[Y_{kt}]$
- $\mu^* = \max_k \mu_k$
- For $k \neq k'$, Y_{kt} and $Y_{k't'}$ are independent
- $J_{\text{bad}} = \{k \,|\, \mu_k < \mu^*\}$, set of "bad" arms
- $J_{\text{good}} = \{k \,|\, \mu_k = \mu^*\}$, set of "good" arms
- $I_t^{\mathcal{A}}$ – choice of arm at time t by some allocation rule $\mathcal{A}$
- $T_{kt}^{\mathcal{A}} = \sum_{s=1}^{t} \mathbb{I}_{\{I_s^{\mathcal{A}}=k\}}$ (# of choosing k)

We shall drop $\mathcal{A}$ from $I_t^{\mathcal{A}}$, $T_{kt}^{\mathcal{A}}$ when $\mathcal{A}$ is unambigous $\rightarrow I_t$, T_{kt}

BANDITS

- Y_{kt} – payoff of arm k when chosen the tth time, $1 \leq k \leq K$
- For k fixed, Y_{kt} is an i.i.d. sequence
- $\mu_k = \mathbb{E}[Y_{kt}]$
- $\mu^* = \max_k \mu_k$
- For $k \neq k'$, Y_{kt} and $Y_{k't'}$ are independent
- $J_{\text{bad}} = \{k \,|\, \mu_k < \mu^*\}$, set of "bad" arms
- $J_{\text{good}} = \{k \,|\, \mu_k = \mu^*\}$, set of "good" arms
- $I_t^{\mathcal{A}}$ – choice of arm at time t by some allocation rule $\mathcal{A}$
- $T_{kt}^{\mathcal{A}} = \sum_{s=1}^{t} \mathbb{I}_{\{I_s^{\mathcal{A}}=k\}}$ (# of choosing k)

We shall drop $\mathcal{A}$ from $I_t^{\mathcal{A}}$, $T_{kt}^{\mathcal{A}}$ when $\mathcal{A}$ is unambigous $\to I_t$, T_{kt}

BANDITS

- Y_{kt} – payoff of arm k when chosen the tth time, $1 \leq k \leq K$
- For k fixed, Y_{kt} is an i.i.d. sequence
- $\mu_k = \mathbb{E}[Y_{kt}]$
- $\mu^* = \max_k \mu_k$
- For $k \neq k'$, Y_{kt} and $Y_{k't'}$ are independent
- $J_{\text{bad}} = \{k \,|\, \mu_k < \mu^*\}$, set of "bad" arms
- $J_{\text{good}} = \{k \,|\, \mu_k = \mu^*\}$, set of "good" arms
- $I_t^{\mathcal{A}}$ – choice of arm at time t by some allocation rule $\mathcal{A}$
- $T_{kt}^{\mathcal{A}} = \sum_{s=1}^{t} \mathbb{I}_{\{I_s^{\mathcal{A}}=k\}}$ (# of choosing k)

We shall drop $\mathcal{A}$ from $I_t^{\mathcal{A}}$, $T_{kt}^{\mathcal{A}}$ when $\mathcal{A}$ is unambigous $\to I_t$, T_{kt}

BANDITS

- Y_{kt} – payoff of arm k when chosen the tth time, $1 \leq k \leq K$
- For k fixed, Y_{kt} is an i.i.d. sequence
- $\mu_k = \mathbb{E}[Y_{kt}]$
- $\mu^* = \max_k \mu_k$
- For $k \neq k'$, Y_{kt} and $Y_{k't'}$ are independent
- $J_{\mathrm{bad}} = \{k \,|\, \mu_k < \mu^*\}$, set of "bad" arms
- $J_{\mathrm{good}} = \{k \,|\, \mu_k = \mu^*\}$, set of "good" arms
- $I_t^{\mathcal{A}}$ – choice of arm at time t by some allocation rule $\mathcal{A}$
- $T_{kt}^{\mathcal{A}} = \sum_{s=1}^{t} \mathbb{I}_{\{I_s^{\mathcal{A}}=k\}}$ (# of choosing k)

We shall drop $\mathcal{A}$ from $I_t^{\mathcal{A}}$, $T_{kt}^{\mathcal{A}}$ when $\mathcal{A}$ is unambigous $\rightarrow I_t$, T_{kt}

Outline

REGRET

- $T_{I_t,t}$ = # of pulls of arm I_t till time t
- $Y_{I_t,T_{I_t,t}}$ =payoff in the t-step
- Payoff/-Loss in n steps is

$$L_n(\mathcal{A}) = \sum_{t=1}^{n} Y_{I_t,T_{I_t,t}}$$

- Expected regret in n steps:

$$R_n(\mathcal{A}) \stackrel{\text{def}}{=} \sup_{\mathcal{A}'} \mathbb{E}\left[L_n(\mathcal{A}')\right] - \mathbb{E}\left[L_n(\mathcal{A})\right].$$

- Goal: Minimize regret!
- Constraint: Distributions of the payoffs are unknown.

This is stochastic bandit.
There is non-stochastic: $\{Y_{kt}\}_{t\geq 1}$ is not i.i.d. random, but *any individual* sequence, $\mathbb{E}[]$ is replaced by sup over them.
Variation of special case of *prediction with expert advice*.

REGRET

- $T_{I_t,t}$ = # of pulls of arm I_t till time t
- $Y_{I_t,T_{I_t,t}}$ =payoff in the t-step
- Payoff/-Loss in n steps is

$$L_n(\mathcal{A}) = \sum_{t=1}^{n} Y_{I_t,T_{I_t,t}}$$

- Expected regret in n steps:

$$R_n(\mathcal{A}) \stackrel{\text{def}}{=} \sup_{\mathcal{A}'} \mathbb{E}\left[L_n(\mathcal{A}')\right] - \mathbb{E}\left[L_n(\mathcal{A})\right].$$

- Goal: Minimize regret!
- Constraint: Distributions of the payoffs are unknown.

This is stochastic bandit.
There is non-stochastic: $\{Y_{kt}\}_{t\geq 1}$ is not i.i.d. random, but *any individual* sequence, $\mathbb{E}\left[\right]$ is replaced by sup over them.
Variation of special case of *prediction with expert advice*.

REGRET

- $T_{I_t,t}$ = # of pulls of arm I_t till time t
- $Y_{I_t,T_{I_t,t}}$ =payoff in the t-step
- Payoff/-Loss in n steps is

$$L_n(\mathcal{A}) = \sum_{t=1}^{n} Y_{I_t,T_{I_t,t}}$$

- Expected regret in n steps:

$$R_n(\mathcal{A}) \stackrel{\text{def}}{=} \sup_{\mathcal{A}'} \mathbb{E}\left[L_n(\mathcal{A}')\right] - \mathbb{E}\left[L_n(\mathcal{A})\right].$$

- Goal: Minimize regret!
- Constraint: Distributions of the payoffs are unknown.

This is stochastic bandit.

There is non-stochastic: $\{Y_{kt}\}_{t\geq 1}$ is not i.i.d. random, but *any individual* sequence, $\mathbb{E}\,[]$ is replaced by sup over them.

Variation of special case of *prediction with expert advice*.

REGRET

- $T_{I_t,t}$ = # of pulls of arm I_t till time t
- $Y_{I_t,T_{I_t,t}}$ =payoff in the t-step
- Payoff/-Loss in n steps is

$$L_n(\mathcal{A}) = \sum_{t=1}^{n} Y_{I_t,T_{I_t,t}}$$

- Expected regret in n steps:

$$R_n(\mathcal{A}) \stackrel{\text{def}}{=} \sup_{\mathcal{A}'} \mathbb{E}\left[L_n(\mathcal{A}')\right] - \mathbb{E}\left[L_n(\mathcal{A})\right].$$

- Goal: Minimize regret!
- Constraint: Distributions of the payoffs are unknown.

This is stochastic bandit.

There is non-stochastic: $\{Y_{kt}\}_{t\geq 1}$ is not i.i.d. random, but *any individual* sequence, $\mathbb{E}\,[]$ is replaced by sup over them.

Variation of special case of *prediction with expert advice*.

REGRET

- $T_{I_t,t}$ = # of pulls of arm I_t till time t
- $Y_{I_t,T_{I_t,t}}$ =payoff in the t-step
- Payoff/-Loss in n steps is

$$L_n(\mathcal{A}) = \sum_{t=1}^{n} Y_{I_t,T_{I_t,t}}$$

- Expected regret in n steps:

$$R_n(\mathcal{A}) \stackrel{\text{def}}{=} \sup_{\mathcal{A}'} \mathbb{E}\left[L_n(\mathcal{A}')\right] - \mathbb{E}\left[L_n(\mathcal{A})\right].$$

- Goal: Minimize regret!
- Constraint: Distributions of the payoffs are unknown.

This is stochastic bandit.

There is non-stochastic: $\{Y_{kt}\}_{t\geq 1}$ is not i.i.d. random, but *any individual* sequence, $\mathbb{E}\left[\right]$ is replaced by sup over them.

Variation of special case of *prediction with expert advice*.

REGRET

- $T_{I_t,t}$ = # of pulls of arm I_t till time t
- $Y_{I_t,T_{I_t,t}}$ =payoff in the t-step
- Payoff/-Loss in n steps is

$$L_n(\mathcal{A}) = \sum_{t=1}^{n} Y_{I_t,T_{I_t,t}}$$

- Expected regret in n steps:

$$R_n(\mathcal{A}) \stackrel{\text{def}}{=} \sup_{\mathcal{A}'} \mathbb{E}\left[L_n(\mathcal{A}')\right] - \mathbb{E}\left[L_n(\mathcal{A})\right].$$

- Goal: Minimize regret!
- Constraint: Distributions of the payoffs are unknown.

This is stochastic bandit.
There is non-stochastic: $\{Y_{kt}\}_{t\geq 1}$ is not i.i.d. random, but *any individual* sequence, $\mathbb{E}\left[\right]$ is replaced by sup over them.
Variation of special case of *prediction with expert advice*.

REGRET

- $T_{I_t,t}$ = # of pulls of arm I_t till time t
- $Y_{I_t,T_{I_t,t}}$ =payoff in the t-step
- Payoff/-Loss in n steps is

$$L_n(\mathcal{A}) = \sum_{t=1}^{n} Y_{I_t,T_{I_t,t}}$$

- Expected regret in n steps:

$$R_n(\mathcal{A}) \stackrel{\text{def}}{=} \sup_{\mathcal{A}'} \mathbb{E}\left[L_n(\mathcal{A}')\right] - \mathbb{E}\left[L_n(\mathcal{A})\right].$$

- Goal: Minimize regret!
- Constraint: Distributions of the payoffs are unknown.

This is stochastic bandit.
There is non-stochastic: $\{Y_{kt}\}_{t\geq 1}$ is not i.i.d. random, but *any individual* sequence, $\mathbb{E}\,[]$ is replaced by sup over them.
Variation of special case of *prediction with expert advice*.

AN EQUIVALENT FORM OF THE REGRET

- Exercise #1: Expected payoff

$$\mathbb{E}\left[L_n(\mathcal{A})\right] = \sum_{k=1}^{K} \mu_k \mathbb{E}\left[T_{kn}\right] \leq n\mu^*$$

Hint: Use Wald's identity. T_{kn} is stopping time w.r.t. ...

- Exercise #2:

$$\sup_{\mathcal{A}'} \mathbb{E}\left[R_n(\mathcal{A}')\right] = n\mu^*$$

- Let $\Delta_k = \mu^* - \mu_k$. Hence:

$$R_n(\mathcal{A}) = n\mu^* - \sum_{k=1}^{K} \mu_k \mathbb{E}\left[T_{kn}\right] = \sum_{k \in J_{\text{bad}}} \Delta_k \mathbb{E}\left[T_{kn}\right].$$

An equivalent form of the regret

- Exercise #1: Expected payoff

$$\mathbb{E}\left[L_n(\mathcal{A})\right] = \sum_{k=1}^{K} \mu_k \mathbb{E}\left[T_{kn}\right] \leq n\mu^*$$

Hint: Use Wald's identity. T_{kn} is stopping time w.r.t. …

- Exercise #2:

$$\sup_{\mathcal{A}'} \mathbb{E}\left[R_n(\mathcal{A}')\right] = n\mu^*$$

- Let $\Delta_k = \mu^* - \mu_k$. Hence:

$$R_n(\mathcal{A}) = n\mu^* - \sum_{k=1}^{K} \mu_k \mathbb{E}\left[T_{kn}\right] = \sum_{k \in J_{\mathrm{bad}}} \Delta_k \mathbb{E}\left[T_{kn}\right].$$

AN EQUIVALENT FORM OF THE REGRET

- Exercise #1: Expected payoff

$$\mathbb{E}\left[L_n(\mathcal{A})\right] = \sum_{k=1}^{K} \mu_k \mathbb{E}\left[T_{kn}\right] \leq n\mu^*$$

Hint: Use Wald's identity. T_{kn} is stopping time w.r.t. ...

- Exercise #2:

$$\sup_{\mathcal{A}'} \mathbb{E}\left[R_n(\mathcal{A}')\right] = n\mu^*$$

- Let $\Delta_k = \mu^* - \mu_k$. Hence:

$$R_n(\mathcal{A}) = n\mu^* - \sum_{k=1}^{K} \mu_k \mathbb{E}\left[T_{kn}\right] = \sum_{k \in J_{\text{bad}}} \Delta_k \mathbb{E}\left[T_{kn}\right].$$

An equivalent form of the regret

- Exercise #1: Expected payoff

$$\mathbb{E}\left[L_n(\mathcal{A})\right] = \sum_{k=1}^{K} \mu_k \mathbb{E}\left[T_{kn}\right] \leq n\mu^*$$

Hint: Use Wald's identity. T_{kn} is stopping time w.r.t. ...

- Exercise #2:

$$\sup_{\mathcal{A}'} \mathbb{E}\left[R_n(\mathcal{A}')\right] = n\mu^*$$

- Let $\Delta_k = \mu^* - \mu_k$. Hence:

$$R_n(\mathcal{A}) = n\mu^* - \sum_{k=1}^{K} \mu_k \mathbb{E}\left[T_{kn}\right] = \sum_{k \in J_{\text{bad}}} \Delta_k \mathbb{E}\left[T_{kn}\right].$$

AN EQUIVALENT FORM OF THE REGRET

- Exercise #1: Expected payoff

$$\mathbb{E}\left[L_n(\mathcal{A})\right] = \sum_{k=1}^{K} \mu_k \mathbb{E}\left[T_{kn}\right] \leq n\mu^*$$

  Hint: Use Wald's identity. T_{kn} is stopping time w.r.t. ...

- Exercise #2:

$$\sup_{\mathcal{A}'} \mathbb{E}\left[R_n(\mathcal{A}')\right] = n\mu^*$$

- Let $\Delta_k = \mu^* - \mu_k$. Hence:

$$R_n(\mathcal{A}) = n\mu^* - \sum_{k=1}^{K} \mu_k \mathbb{E}\left[T_{kn}\right] = \sum_{k \in J_{\text{bad}}} \Delta_k \mathbb{E}\left[T_{kn}\right].$$

Wald's Identities

A r.v. T is a stopping time w.r.t. a sequence $\{Y_t\}$ of r.v.'s, if for each t, $\mathbb{I}_{\{T \leq t\}}$ depends only on $Y_1, \ldots, Y_t$.

Lemma (Wald's Identities — special case)

Let $\{Y_t\}$ be an i.i.d. sequence of r.v.'s, T be a stopping time w.r.t. $\{Y_t\}$, and $\mathbb{E}[T] < \infty$. If $\mathbb{E}[|Y_1|] < \infty$ then

$$\mathbb{E}\left[\sum_{t=1}^{T} Y_t\right] = \mathbb{E}[Y_1]\,\mathbb{E}[T].$$

If $\mathbb{E}\left[Y_1^2\right] < \infty$ then

$$\mathbb{E}\left[\left(\sum_{t=1}^{T} Y_t - T\mathbb{E}[Y_1]\right)^2\right] = \mathrm{Var}[Y_1]\,\mathbb{E}[T].$$

Wald's Identities — general

T is a stopping time w.r.t. a filtration $\{\mathcal{F}_t\}$, if for each t, $\{T \leq t\} \in \mathcal{F}_t$.

Lemma (Wald's Identities)

Let $\{\mathcal{F}_t\}$ be a filtration and $\{Y_t\}$ be $\mathcal{F}_t$-adapted i.i.d. sequence of r.v.'s. Assume that $\mathcal{F}_t$ and $\sigma(\{Y_s : s \geq t+1\})$ are independent, T is a stopping time w.r.t. $\mathcal{F}_t$, and $\mathbb{E}[T] < \infty$. If $\mathbb{E}[|Y_1|] < \infty$ then

$$\mathbb{E}\left[\sum_{t=1}^{T} Y_t\right] = \mathbb{E}[Y_1]\,\mathbb{E}[T].$$

If $\mathbb{E}\left[Y_1^2\right] < \infty$ then

$$\mathbb{E}\left[\left(\sum_{t=1}^{T} Y_t - T\,\mathbb{E}[Y_1]\right)^2\right] = \mathrm{Var}[Y_1]\,\mathbb{E}[T].$$

OUTLINE

ϵ-greedy policies

- Notation:

$$\overline{Y}_{kt} = \frac{1}{t}\sum_{t'=1}^{t} Y_{kt'}$$

- Assumption: $0 \leq Y_{kt} \leq 1$ (in the rest of the talk!)

ϵ-greedy

1. Initialization: Choose all arms $1, \ldots, K$ once.
2. At time t choose arm with the maximal payoff with probability $1 - \epsilon_t$, otherwise an arm uniformly.

ϵ-GREEDY POLICIES

- Notation:

$$\overline{Y}_{kt} = \frac{1}{t}\sum_{t'=1}^{t} Y_{kt'}$$

- Assumption: $0 \leq Y_{kt} \leq 1$ (in the rest of the talk!)

ϵ-GREEDY

1. Initialization: Choose all arms $1, \ldots, K$ once.
2. At time t choose arm with the maximal payoff with probability $1 - \epsilon_t$, otherwise an arm uniformly.

ϵ-GREEDY POLICIES

- Notation:

$$\overline{Y}_{kt} = \frac{1}{t} \sum_{t'=1}^{t} Y_{kt'}$$

- Assumption: $0 \leq Y_{kt} \leq 1$ (in the rest of the talk!)

ϵ-GREEDY

1. Initialization: Choose all arms $1, \ldots, K$ once.
2. At time t choose arm with the maximal payoff with probability $1 - \epsilon_t$, otherwise an arm uniformly

ϵ-GREEDY POLICIES

- Notation:

$$\overline{Y}_{kt} = \frac{1}{t} \sum_{t'=1}^{t} Y_{kt'}$$

- Assumption: $0 \leq Y_{kt} \leq 1$ (in the rest of the talk!)

ϵ-GREEDY

1. Initialization: Choose all arms $1, \ldots, K$ once.
2. At time t choose arm with the maximal payoff with probability $1 - \epsilon_t$, otherwise an arm uniformly

ϵ-GREEDY POLICIES

- Notation:

$$\overline{Y}_{kt} = \frac{1}{t}\sum_{t'=1}^{t} Y_{kt'}$$

- Assumption: $0 \leq Y_{kt} \leq 1$ (in the rest of the talk!)

ϵ-GREEDY

1. Initialization: Choose all arms $1, \dots, K$ once.
2. At time t choose arm with the maximal payoff with probability $1 - \epsilon_t$, otherwise an arm uniformly

PERFORMANCE – FIRST STEPS

- ϵ-greedy choice:

$$\mathbb{P}\left(I_t = \mathrm{argmax}_k \overline{Y}_{kt} \middle| \{\overline{Y}_{kt}\}_{1\leq k\leq K}\right) = 1 - \epsilon_t.$$

- $\epsilon_t = 0$: always choose maximum. Why is this bad?
  Exercise #3: Give a lower bound on the regret for Bernoulli bandits
- $\epsilon_t = 1$ clearly not good
- Fix $\epsilon_t = \epsilon$: regret still linear. Exercise #4: Give a lower bound on the regret for $0 < \epsilon < 1$

Exploration-exploitation tradeoff!

PERFORMANCE – FIRST STEPS

- ϵ-greedy choice:

$$\mathbb{P}\left(I_t = \text{argmax}_k \overline{Y}_{kt} \middle| \{\overline{Y}_{kt}\}_{1 \le k \le K}\right) = 1 - \epsilon_t.$$

- $\epsilon_t = 0$: always choose maximum. Why is this bad?

  Exercise #3: Give a lower bound on the regret for Bernoulli bandits

- $\epsilon_t = 1$ clearly not good
- Fix $\epsilon_t = \epsilon$: regret still linear. Exercise #4: Give a lower bound on the regret for $0 < \epsilon < 1$

Exploration-exploitation tradeoff!

PERFORMANCE – FIRST STEPS

- ϵ-greedy choice:

$$\mathbb{P}\left(I_t = \mathrm{argmax}_k \overline{Y}_{kt} \middle| \{\overline{Y}_{kt}\}_{1\leq k\leq K}\right) = 1 - \epsilon_t.$$

- $\epsilon_t = 0$: always choose maximum. Why is this bad?

  Exercise #3: Give a lower bound on the regret for Bernoulli bandits
- $\epsilon_t = 1$ clearly not good
- Fix $\epsilon_t = \epsilon$: regret still linear. Exercise #4: Give a lower bound on the regret for $0 < \epsilon < 1$

Exploration-exploitation tradeoff!

PERFORMANCE – FIRST STEPS

- ϵ-greedy choice:

$$\mathbb{P}\left(I_t = \mathrm{argmax}_k \overline{Y}_{kt} \middle| \{\overline{Y}_{kt}\}_{1 \leq k \leq K}\right) = 1 - \epsilon_t.$$

- $\epsilon_t = 0$: always choose maximum. Why is this bad?
  Exercise #3: Give a lower bound on the regret for Bernoulli bandits
- $\epsilon_t = 1$ clearly not good
- Fix $\epsilon_t = \epsilon$: regret still linear. Exercise #4: Give a lower bound on the regret for $0 < \epsilon < 1$

Exploration-exploitation tradeoff!

PERFORMANCE – FIRST STEPS

- ϵ-greedy choice:

$$\mathbb{P}\left(I_t = \text{argmax}_k \overline{Y}_{kt} \middle| \{\overline{Y}_{kt}\}_{1\leq k\leq K}\right) = 1 - \epsilon_t.$$

- $\epsilon_t = 0$: always choose maximum. Why is this bad?
  Exercise #3: Give a lower bound on the regret for Bernoulli bandits
- $\epsilon_t = 1$ clearly not good
- Fix $\epsilon_t = \epsilon$: regret still linear. Exercise #4: Give a lower bound on the regret for $0 < \epsilon < 1$

Exploration-exploitation tradeoff!

PERFORMANCE – FIRST STEPS

- ϵ-greedy choice:

$$\mathbb{P}\left(I_t = \mathrm{argmax}_k \overline{Y}_{kt} \middle| \{\overline{Y}_{kt}\}_{1 \le k \le K}\right) = 1 - \epsilon_t.$$

- $\epsilon_t = 0$: always choose maximum. Why is this bad?
  Exercise #3: Give a lower bound on the regret for Bernoulli bandits
- $\epsilon_t = 1$ clearly not good
- Fix $\epsilon_t = \epsilon$: regret still linear. Exercise #4: Give a lower bound on the regret for $0 < \epsilon < 1$

Exploration-exploitation tradeoff!

LOGARITHMIC REGRET

IDEA!

In order to achieve logarithmic (cumulative) regret, the probability of not selecting the best looking arm in step t should be $\approx 1/t$, since $\sum_{t=1}^{n} 1/t \approx \ln n$!

THEOREM (INSTANTANEOUS REGRET BOUND [AUER ET AL., 2002])

Let $\Delta_{\min} = \min_{j \in J_{\text{bad}}} \Delta_j$. Let $\epsilon_t = \min(1, \frac{5K}{\Delta_{\min}^2 t})$ time dependent. If $n \geq 5K/\Delta_{\min}$ then

$$\mathbb{P}\left(I_n \notin J_{\text{good}}\right) = O\left(\frac{1}{\Delta_{\min}^2 n}\right) \quad \text{and} \quad R_n(\mathcal{A}_\epsilon) \leq O\left(\frac{1}{\Delta_{\min}^2}\right) \ln n.$$

Logarithmic regret

Idea!

In order to achieve logarithmic (cumulative) regret, the probability of not selecting the best looking arm in step t should be $\approx 1/t$, since $\sum_{t=1}^{n} 1/t \approx \ln n$!

Theorem (Instantaneous regret bound [Auer et al., 2002])

Let $\Delta_{\min} = \min_{j \in J_{\text{bad}}} \Delta_j$. *Let* $\epsilon_t = \min(1, \frac{5K}{\Delta_{\min}^2 t})$ *time dependent. If* $n \geq 5K/\Delta_{\min}$ *then*

$$\mathbb{P}\left(I_n \notin J_{\text{good}}\right) = O\left(\frac{1}{\Delta_{\min}^2 n}\right) \quad \textit{and} \quad R_n(\mathcal{A}_\epsilon) \leq O\left(\frac{1}{\Delta_{\min}^2}\right) \ln n.$$

PROOF

- Two sources of error:
  1. Randomization (fine, by design!)
  2. Not picking an optimal arm when we wanted to; assuming single optimal arm with index k^*, with $I_t = \operatorname{argmax}_i \overline{Y}_{it}$:

$$\mathbb{P}(I_t \neq k^*) = \mathbb{P}\left(\overline{Y}_{I_t,t} > \overline{Y}_{k^*,t}\right) = \ldots$$

- We need to compare the probability that one average is larger than another one
- How to do this? Solution: Law of large numbers: Averages are close to their expected values: $\overline{Y}_{k,t} \approx \mu_k < \mu^* \approx \overline{Y}_{k^*,t}$
- But how close?? $\Rightarrow$ Concentration inequalities!

PROOF

- Two sources of error:
  1. Randomization (fine, by design!)
  2. Not picking an optimal arm when we wanted to; assuming single optimal arm with index k^*, with $I_t = \operatorname{argmax}_i \overline{Y}_{it}$:

$$\mathbb{P}(I_t \neq k^*) = \mathbb{P}\left(\overline{Y}_{I_t,t} > \overline{Y}_{k^*,t}\right) = \dots$$

- We need to compare the probability that one average is larger than another one
- How to do this? Solution: Law of large numbers: Averages are close to their expected values: $\overline{Y}_{k,t} \approx \mu_k < \mu^* \approx \overline{Y}_{k^*,t}$
- But how close?? $\Rightarrow$ Concentration inequalities!

PROOF

- Two sources of error:
  1. Randomization (fine, by design!)
  2. Not picking an optimal arm when we wanted to; assuming single optimal arm with index k^*, with $I_t = \mathrm{argmax}_i \overline{Y}_{it}$:

$$\mathbb{P}(I_t \neq k^*) = \mathbb{P}\left(\overline{Y}_{I_t,t} > \overline{Y}_{k^*,t}\right) = \dots$$

- We need to compare the probability that one average is larger than another one
- How to do this? Solution: Law of large numbers: Averages are close to their expected values: $\overline{Y}_{k,t} \approx \mu_k < \mu^* \approx \overline{Y}_{k^*,t}$
- But how close?? $\Rightarrow$ Concentration inequalities!

PROOF

- Two sources of error:
  1. Randomization (fine, by design!)
  2. Not picking an optimal arm when we wanted to; assuming single optimal arm with index k^*, with $I_t = \mathrm{argmax}_i \overline{Y}_{it}$:

$$\mathbb{P}(I_t \neq k^*) = \mathbb{P}\left(\overline{Y}_{I_t,t} > \overline{Y}_{k^*,t}\right) = \dots$$

- We need to compare the probability that one average is larger than another one
- How to do this? Solution: Law of large numbers: Averages are close to their expected values: $\overline{Y}_{k,t} \approx \mu_k < \mu^* \approx \overline{Y}_{k^*,t}$
- But how close?? $\Rightarrow$ Concentration inequalities!

PROOF

- Two sources of error:
  1. Randomization (fine, by design!)
  2. Not picking an optimal arm when we wanted to; assuming single optimal arm with index k^*, with $I_t = \mathrm{argmax}_i \overline{Y}_{it}$:

$$\mathbb{P}\left(I_t \neq k^*\right) = \mathbb{P}\left(\overline{Y}_{I_t,t} > \overline{Y}_{k^*,t}\right) = \dots$$

- We need to compare the probability that one average is larger than another one
- How to do this? Solution: Law of large numbers: Averages are close to their expected values: $\overline{Y}_{k,t} \approx \mu_k < \mu^* \approx \overline{Y}_{k^*,t}$
- But how close?? $\Rightarrow$ Concentration inequalities!

PROOF

- Two sources of error:
  1. Randomization (fine, by design!)
  2. Not picking an optimal arm when we wanted to; assuming single optimal arm with index k^*, with $I_t = \mathrm{argmax}_i \overline{Y}_{it}$:

$$\mathbb{P}\left(I_t \neq k^*\right) = \mathbb{P}\left(\overline{Y}_{I_t,t} > \overline{Y}_{k^*,t}\right) = \ldots$$

- We need to compare the probability that one average is larger than another one
- How to do this? Solution: Law of large numbers: Averages are close to their expected values: $\overline{Y}_{k,t} \approx \mu_k < \mu^* \approx \overline{Y}_{k^*,t}$
- But how close?? $\Rightarrow$ Concentration inequalities!

OUTLINE

SOME INEQUALITIES

Mild assumptions on X! (no parametric forms)

- Markov: $X \geq 0$ then $\mathbb{P}(X \geq \epsilon) \leq \mathbb{E}[X]/\epsilon$
- Now, for any $\phi : \mathbb{R}^+ \to \mathbb{R}^+$ strictly increasing,

$$\mathbb{P}(X \geq \epsilon) = \mathbb{P}(\phi(X) \geq \phi(\epsilon)) \leq \mathbb{E}[\phi(X)]/\phi(\epsilon).$$

- Chebyshev: Choose $\phi(\epsilon) = \epsilon^2$! Let X be such that $\mathrm{Var}[X] < \infty$. Then using for $|X - \mathbb{E}[X]|$:

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \epsilon) \leq \frac{\mathrm{Var}[X]}{\epsilon^2}.$$

How tight is Chebyshev's inequality??

SOME INEQUALITIES

Mild assumptions on X! (no parametric forms)

- Markov: $X \geq 0$ then $\mathbb{P}(X \geq \epsilon) \leq \mathbb{E}[X]/\epsilon$
- Now, for any $\phi : \mathbb{R}^+ \to \mathbb{R}^+$ strictly increasing,

$$\mathbb{P}(X \geq \epsilon) = \mathbb{P}(\phi(X) \geq \phi(\epsilon)) \leq \mathbb{E}[\phi(X)]/\phi(\epsilon).$$

- Chebyshev: Choose $\phi(\epsilon) = \epsilon^2$! Let X be such that $\mathrm{Var}[X] < \infty$. Then using for $|X - \mathbb{E}[X]|$:

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \epsilon) \leq \frac{\mathrm{Var}[X]}{\epsilon^2}.$$

How tight is Chebyshev's inequality??

SOME INEQUALITIES

Mild assumptions on X! (no parametric forms)

- Markov: $X \geq 0$ then $\mathbb{P}(X \geq \epsilon) \leq \mathbb{E}[X] / \epsilon$
- Now, for any $\phi : \mathbb{R}^+ \to \mathbb{R}^+$ strictly increasing,

$$\mathbb{P}(X \geq \epsilon) = \mathbb{P}(\phi(X) \geq \phi(\epsilon)) \leq \mathbb{E}[\phi(X)] / \phi(\epsilon).$$

- Chebyshev: Choose $\phi(\epsilon) = \epsilon^2$! Let X be such that $\mathrm{Var}[X] < \infty$. Then using for $|X - \mathbb{E}[X]|$:

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \epsilon) \leq \frac{\mathrm{Var}[X]}{\epsilon^2}.$$

How tight is Chebyshev's inequality??

SOME INEQUALITIES

Mild assumptions on X! (no parametric forms)

- Markov: $X \geq 0$ then $\mathbb{P}(X \geq \epsilon) \leq \mathbb{E}[X]/\epsilon$
- Now, for any $\phi : \mathbb{R}^+ \to \mathbb{R}^+$ strictly increasing,

$$\mathbb{P}(X \geq \epsilon) = \mathbb{P}(\phi(X) \geq \phi(\epsilon)) \leq \mathbb{E}[\phi(X)]/\phi(\epsilon).$$

- Chebyshev: Choose $\phi(\epsilon) = \epsilon^2$! Let X be such that $\mathrm{Var}[X] < \infty$. Then using for $|X - \mathbb{E}[X]|$:

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \epsilon) \leq \frac{\mathrm{Var}[X]}{\epsilon^2}.$$

How tight is Chebyshev's inequality??

CHEBYSHEV'S INEQUALITY FOR AVERAGE

$$\mathbb{P}\left(\left|\overline{Y}_n - \mathbb{E}\left[Y_1\right]\right| \geq \epsilon\right) \leq \frac{\mathrm{Var}\left[Y_1\right]}{n\epsilon^2}$$

INTUITION: CENTRAL LIMIT THEOREM

$$\mathbb{P}\left(\overline{Y}_n - \mathbb{E}\left[Y_1\right] \geq \epsilon\right) = \mathbb{P}\left(\frac{\sqrt{n}}{\sigma}\left(\overline{Y}_n - \mathbb{E}\left[Y_1\right]\right) \geq \frac{\sqrt{n}}{\sigma}\epsilon\right)$$

$$\to 1 - \Phi\left(\frac{\sqrt{n}}{\sigma}\epsilon\right) \approx e^{-n\epsilon^2/(2\sigma^2)}\frac{\sigma}{\sqrt{n}\epsilon} \approx e^{-n\epsilon^2/(2\sigma^2)}.$$

exponential ⇒ much sharper could be!

Chebyshev's inequality for average

$$\mathbb{P}\left(|\overline{Y}_n - \mathbb{E}[Y_1]| \geq \epsilon\right) \leq \frac{\mathrm{Var}[Y_1]}{n\epsilon^2}$$

Intuition: Central limit theorem

$$\begin{aligned} \mathbb{P}\left(\overline{Y}_n - \mathbb{E}[Y_1] \geq \epsilon\right) &= \mathbb{P}\left(\frac{\sqrt{n}}{\sigma}\left(\overline{Y}_n - \mathbb{E}[Y_1]\right) \geq \frac{\sqrt{n}}{\sigma}\epsilon\right) \\ \to 1 - \Phi\left(\frac{\sqrt{n}}{\sigma}\epsilon\right) &\approx e^{-n\epsilon^2/(2\sigma^2)}\frac{\sigma}{\sqrt{n}\epsilon} \approx e^{-n\epsilon^2/(2\sigma^2)}. \end{aligned}$$

exponential $\Rightarrow$ much sharper could be!

SHARPENING THE BOUNDS

- For $\phi \geq 0$ strictly increasing:

$$\mathbb{P}(X \geq \epsilon) = \mathbb{P}(\phi(X) \geq \phi(\epsilon)) \leq \mathbb{E}[\phi(X)]/\phi(\epsilon).$$

- Higher moments: $\phi(\epsilon) = \epsilon^q$, $q \geq 2$:

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \epsilon) \leq \mathbb{E}[|X - \mathbb{E}[X]|^q]/\epsilon^q.$$

– improvement! (though requires $\mathbb{E}[|X - \mathbb{E}[X]|^q] < \infty$)

- Exponential ϕ?

SHARPENING THE BOUNDS

- For $\phi \geq 0$ strictly increasing:

$$\mathbb{P}(X \geq \epsilon) = \mathbb{P}(\phi(X) \geq \phi(\epsilon)) \leq \mathbb{E}[\phi(X)] / \phi(\epsilon).$$

- Higher moments: $\phi(\epsilon) = \epsilon^q$, $q \geq 2$:

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \epsilon) \leq \mathbb{E}\left[|X - \mathbb{E}[X]|^q\right] / \epsilon^q.$$

  – improvement! (though requires $\mathbb{E}[|X - \mathbb{E}[X]|^q] < \infty$)
- Exponential ϕ?

Sharpening the bounds

- For $\phi \geq 0$ strictly increasing:

$$\mathbb{P}(X \geq \epsilon) = \mathbb{P}(\phi(X) \geq \phi(\epsilon)) \leq \mathbb{E}[\phi(X)] / \phi(\epsilon).$$

- Higher moments: $\phi(\epsilon) = \epsilon^q$, $q \geq 2$:

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \epsilon) \leq \mathbb{E}\left[|X - \mathbb{E}[X]|^q\right] / \epsilon^q.$$

  – improvement! (though requires $\mathbb{E}[|X - \mathbb{E}[X]|^q] < \infty$)

- Exponential ϕ?

SHARPENING THE BOUNDS/2

- Chernoff's method: $\phi(x) = e^{sx}$, $s > 0$;

$$\mathbb{P}(X \geq t) \leq \mathbb{E}\left[e^{sX}\right] e^{-st}$$

  and optimize for s!

- Apply to $n\overline{Y}_n = S_n = \sum_{t=1}^{n} Y_t$:

$$\begin{aligned}\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) &\leq e^{-st}\mathbb{E}\left[e^{s(S_n - n\mathbb{E}[Y_1])}\right] \\ &= e^{-st}\Pi_{t=1}^{n}\mathbb{E}\left[e^{s(Y_t - \mathbb{E}[Y_1])}\right]\end{aligned}$$

- Hoeffding: $\mathbb{E}[X] = 0$, $a \leq X \leq b$ then $\mathbb{E}\left[e^{sX}\right] \leq e^{s^2(b-a)^2/8}$

SHARPENING THE BOUNDS/2

- Chernoff's method: $\phi(x) = e^{sx}$, $s > 0$;

$$\mathbb{P}(X \geq t) \leq \mathbb{E}\left[e^{sX}\right] e^{-st}$$

  and optimize for s!
- Apply to $n\overline{Y}_n = S_n = \sum_{t=1}^{n} Y_t$:

$$\begin{aligned} \mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) &\leq e^{-st}\mathbb{E}\left[e^{s(S_n - n\mathbb{E}[Y_1])}\right] \\ &= e^{-st}\Pi_{t=1}^{n}\mathbb{E}\left[e^{s(Y_t - \mathbb{E}[Y_1])}\right] \end{aligned}$$

- Hoeffding: $\mathbb{E}[X] = 0$, $a \leq X \leq b$ then $\mathbb{E}\left[e^{sX}\right] \leq e^{s^2(b-a)^2/8}$

SHARPENING THE BOUNDS/2

- Chernoff's method: $\phi(x) = e^{sx}$, $s > 0$;

$$\mathbb{P}\left(X \geq t\right) \leq \mathbb{E}\left[e^{sX}\right] e^{-st}$$

  and optimize for s!
- Apply to $n\overline{Y}_n = S_n = \sum_{t=1}^{n} Y_t$:

$$\begin{aligned} \mathbb{P}\left(S_n - \mathbb{E}\left[S_n\right] \geq t\right) &\leq e^{-st}\mathbb{E}\left[e^{s(S_n - n\mathbb{E}[Y_1])}\right] \\ &= e^{-st}\Pi_{t=1}^{n}\mathbb{E}\left[e^{s(Y_t - \mathbb{E}[Y_1])}\right] \end{aligned}$$

- Hoeffding: $\mathbb{E}\left[X\right] = 0$, $a \leq X \leq b$ then $\mathbb{E}\left[e^{sX}\right] \leq e^{s^2(b-a)^2/8}$

Theorem (Hoeffding's inequality)

For $Y_i \in [0,1]$ *i.i.d.,* $\mu = \mathbb{E}[Y_1]$, $\overline{Y}_n = \sum_{t=1}^{n} Y_t/n$,

$$\mathbb{P}\left(\overline{Y}_n \geq \mu + \epsilon\right) \leq e^{-2n\epsilon^2}$$

$$\mathbb{P}\left(\overline{Y}_n \leq \mu - \epsilon\right) \leq e^{-2n\epsilon^2}$$

Useful variations

Let the error probability be δ: $\Rightarrow e^{-2n\epsilon^2} = \delta$, $n = n(\epsilon, \delta) = ?$ (sample complexity), $\epsilon = \epsilon(n, \delta) = ?$

$$n = \frac{\ln(1/\delta)}{2\epsilon^2}, \quad \epsilon = \sqrt{\frac{\ln(1/\delta)}{2n}}$$

So with probability $\geq 1 - \delta$

$$\overline{Y}_n - \mu \leq \sqrt{\frac{\ln(1/\delta)}{2n}}$$

THEOREM (HOEFFDING'S INEQUALITY)

For $Y_i \in [0,1]$ *i.i.d.*, $\mu = \mathbb{E}[Y_1]$, $\overline{Y}_n = \sum_{t=1}^{n} Y_t/n$,

$$\mathbb{P}\left(\overline{Y}_n \geq \mu + \epsilon\right) \leq e^{-2n\epsilon^2}$$

$$\mathbb{P}\left(\overline{Y}_n \leq \mu - \epsilon\right) \leq e^{-2n\epsilon^2}$$

USEFUL VARIATIONS

Let the error probability be δ: $\Rightarrow e^{-2n\epsilon^2} = \delta$, $n = n(\epsilon, \delta) =?$ (sample complexity), $\epsilon = \epsilon(n, \delta) =?$

$$n = \frac{\ln(1/\delta)}{2\epsilon^2}, \quad \epsilon = \sqrt{\frac{\ln(1/\delta)}{2n}}.$$

So with probability $\geq 1 - \delta$

$$\overline{Y}_n - \mu < \sqrt{\frac{\ln(1/\delta)}{2n}}$$

THEOREM (HOEFFDING'S INEQUALITY)

For $Y_i \in [0,1]$ i.i.d., $\mu = \mathbb{E}[Y_1]$, $\overline{Y}_n = \sum_{t=1}^{n} Y_t/n$,

$$\mathbb{P}\left(\overline{Y}_n \geq \mu + \epsilon\right) \leq e^{-2n\epsilon^2}$$
$$\mathbb{P}\left(\overline{Y}_n \leq \mu - \epsilon\right) \leq e^{-2n\epsilon^2}$$

USEFUL VARIATIONS

Let the error probability be δ: $\Rightarrow e^{-2n\epsilon^2} = \delta$, $n = n(\epsilon, \delta) =?$ (sample complexity), $\epsilon = \epsilon(n, \delta) =?$

$$n = \frac{\ln(1/\delta)}{2\epsilon^2}, \quad \epsilon = \sqrt{\frac{\ln(1/\delta)}{2n}}.$$

So with probability $\geq 1 - \delta$

$$\overline{Y}_n - \mu < \sqrt{\frac{\ln(1/\delta)}{2n}}$$ universal: holds for any n and δ!

Theorem (Hoeffding's inequality)

For $Y_i \in [0, 1]$ i.i.d., $\mu = \mathbb{E}[Y_1]$, $\overline{Y}_n = \sum_{t=1}^{n} Y_t / n$,

$$\mathbb{P}\left(\overline{Y}_n \geq \mu + \epsilon\right) \leq e^{-2n\epsilon^2}$$

$$\mathbb{P}\left(\overline{Y}_n \leq \mu - \epsilon\right) \leq e^{-2n\epsilon^2}$$

Useful variations

Let the error probability be δ: $\Rightarrow e^{-2n\epsilon^2} = \delta$, $n = n(\epsilon, \delta) =?$ (sample complexity), $\epsilon = \epsilon(n, \delta) =?$

$$n = \frac{\ln(1/\delta)}{2\epsilon^2}, \quad \epsilon = \sqrt{\frac{\ln(1/\delta)}{2n}}.$$

So with probability $\geq 1 - \delta$

$$\overline{Y}_n - \mu < \sqrt{\frac{\ln(1/\delta)}{2n}}$$ universal: holds for any n and δ!

THEOREM (HOEFFDING'S INEQUALITY)

For $Y_i \in [0,1]$ *i.i.d.*, $\mu = \mathbb{E}[Y_1]$, $\overline{Y}_n = \sum_{t=1}^{n} Y_t/n$,

$$\mathbb{P}\left(\overline{Y}_n \geq \mu + \epsilon\right) \leq e^{-2n\epsilon^2}$$
$$\mathbb{P}\left(\overline{Y}_n \leq \mu - \epsilon\right) \leq e^{-2n\epsilon^2}$$

USEFUL VARIATIONS

Let the error probability be δ: $\Rightarrow e^{-2n\epsilon^2} = \delta$, $n = n(\epsilon, \delta) = ?$ (sample complexity), $\epsilon = \epsilon(n, \delta) = ?$

$$n = \frac{\ln(1/\delta)}{2\epsilon^2}, \quad \epsilon = \sqrt{\frac{\ln(1/\delta)}{2n}}.$$

So with probability $\geq 1 - \delta$

$$\overline{Y}_n - \mu < \sqrt{\frac{\ln(1/\delta)}{2n}}$$ universal: holds for any n and δ!

THEOREM (HOEFFDING'S INEQUALITY)

For $Y_i \in [0,1]$ *i.i.d.*, $\mu = \mathbb{E}[Y_1]$, $\overline{Y}_n = \sum_{t=1}^{n} Y_t/n$,

$$\mathbb{P}\left(\overline{Y}_n \geq \mu + \epsilon\right) \leq e^{-2n\epsilon^2}$$
$$\mathbb{P}\left(\overline{Y}_n \leq \mu - \epsilon\right) \leq e^{-2n\epsilon^2}$$

USEFUL VARIATIONS

Let the error probability be δ: $\Rightarrow e^{-2n\epsilon^2} = \delta$, $n = n(\epsilon, \delta) =?$ (sample complexity), $\epsilon = \epsilon(n, \delta) =?$

$$n = \frac{\ln(1/\delta)}{2\epsilon^2}, \quad \epsilon = \sqrt{\frac{\ln(1/\delta)}{2n}}.$$

So with probability $\geq 1 - \delta$

$$\overline{Y}_n - \mu < \sqrt{\frac{\ln(1/\delta)}{2n}}$$ universal: holds for any n and δ!

OUTLINE

UPPER CONFIDENCE BOUNDS: IDEA

We want to pull arms both with high $\overline{Y}_{kt}$ and/or with high uncertainty. Idea: Let's aggregate (add) $\overline{Y}_{kt}$ and its uncertainty! I.e., bias the estimates directly by the uncertainties, then do greedy (without ϵ_t)!

UCB1

1. Initialization: Use all arms once
2. Step $t > K$: Use arm with highest index

$$\overline{Y}_{kt} + c_{t,T_{kt}}.$$

$c_{t,T_{kt}}$ = uncertainity of $\overline{Y}_{kt}$

OPTIMISM IN THE FACE OF UNCERTAINTY

Estimate payoffs in an optimistic way (taking into account uncertainty), choose the arm with the best biased estimate.

How to select $c_{t,T_{kt}}$?

UPPER CONFIDENCE BOUNDS: IDEA

We want to pull arms both with high $\overline{Y}_{kt}$ and/or with high uncertainty. Idea: Let's aggregate (add) $\overline{Y}_{kt}$ and its uncertainty! I.e., bias the estimates directly by the uncertainties, then do greedy (without ϵ_t)!

UCB1

1. Initialization: Use all arms once
2. Step $t > K$: Use arm with highest index

$$\overline{Y}_{kt} + c_{t,T_{kt}}.$$

$c_{t,T_{kt}}$ = uncertainity of $\overline{Y}_{kt}$

OPTIMISM IN THE FACE OF UNCERTAINTY

Estimate payoffs in an optimistic way (taking into account uncertainty), choose the arm with the best biased estimate.

How to select $c_{t,T_{kt}}$?

UPPER CONFIDENCE BOUNDS: IDEA

We want to pull arms both with high $\overline{Y}_{kt}$ and/or with high uncertainty. Idea: Let's aggregate (add) $\overline{Y}_{kt}$ and its uncertainty! I.e., bias the estimates directly by the uncertainties, then do greedy (without ϵ_t)!

UCB1

1. Initialization: Use all arms once
2. Step $t > K$: Use arm with highest index

$$\overline{Y}_{kt} + c_{t,T_{kt}}.$$

$c_{t,T_{kt}}$ = uncertainity of $\overline{Y}_{kt}$

OPTIMISM IN THE FACE OF UNCERTAINTY

Estimate payoffs in an optimistic way (taking into account uncertainty), choose the arm with the best biased estimate.

How to select $c_{t,T_{kt}}$?

UPPER CONFIDENCE BOUNDS: IDEA

We want to pull arms both with high $\overline{Y}_{kt}$ and/or with high uncertainty. Idea: Let's aggregate (add) $\overline{Y}_{kt}$ and its uncertainty! I.e., bias the estimates directly by the uncertainties, then do greedy (without ϵ_t)!

UCB1

1. Initialization: Use all arms once
2. Step $t > K$: Use arm with highest index

$$\overline{Y}_{kt} + c_{t,T_{kt}}.$$

$c_{t,T_{kt}} =$ uncertainity of $\overline{Y}_{kt}$

OPTIMISM IN THE FACE OF UNCERTAINTY

Estimate payoffs in an optimistic way (taking into account uncertainty), choose the arm with the best biased estimate.

How to select $c_{t,T_{kt}}$?

HOW TO SELECT $c_{t,T_{kt}}$?

- Central limit theorem: $\sqrt{T_{kt}}(\overline{Y}_{kt} - \mu_k) \sim \mathcal{N}_{0,\sigma_k}$, so w. high prob. $\overline{Y}_{kt} - \mu_k \in [-2\sigma_k/\sqrt{T_{kt}}, 2\sigma_k/\sqrt{T_{kt}}]$
- Hoeffding (for $Y \in [0,1]$): w. probability $\geq 1 - 2\delta$, $\overline{Y}_{kt} - \mu_k \in [-\sqrt{\ln(1/\delta)/2T_{kt}}, \sqrt{\ln(1/\delta)/2T_{kt}}]$
- Confidence Bounds — measures the uncertainty of $\overline{Y}_{kt}$
- Let it match the confidence radius: $c_{t,T_{kt}} \sim \sqrt{\ln(1/\delta)/T_{kt}}$
- $\overline{Y}_{kt} + c_{t,T_{kt}}$ = Upper Confidence Bound

  Note: Fixed $c_{t,T}$ for fixed T are not enough for infinite exploration (sticks to wrong arm with probability> 0 if the first samples for the optimal arm are bad). We need $\lim_{t\to\infty} c_{t,T} \to \infty$ for T fixed, i.e., $\delta = \delta_t \to 0$ as $t \to \infty$!
- $\delta_t \sim t^{-p} \to 0$, Hoeffding: Let $c_{t,T_{kt}} \sim \sqrt{p \ln t / T_{kt}}$;

HOW TO SELECT $c_{t,T_{kt}}$?

- Central limit theorem: $\sqrt{T_{kt}}(\overline{Y}_{kt} - \mu_k) \sim \mathcal{N}_{0,\sigma_k}$, so w. high prob. $\overline{Y}_{kt} - \mu_k \in [-2\sigma_k/\sqrt{T_{kt}}, 2\sigma_k/\sqrt{T_{kt}}]$
- Hoeffding (for $Y \in [0,1]$): w. probability $\geq 1 - 2\delta$, $\overline{Y}_{kt} - \mu_k \in [-\sqrt{\ln(1/\delta)/2T_{kt}}, \sqrt{\ln(1/\delta)/2T_{kt}}]$
- Confidence Bounds — measures the uncertainty of $\overline{Y}_{kt}$
- Let it match the confidence radius: $c_{t,T_{kt}} \sim \sqrt{\ln(1/\delta)/T_{kt}}$
- $\overline{Y}_{kt} + c_{t,T_{kt}}$ = Upper Confidence Bound

  Note: Fixed $c_{t,T}$ for fixed T are not enough for infinite exploration (sticks to wrong arm with probability > 0 if the first samples for the optimal arm are bad). We need $\lim_{t\to\infty} c_{t,T} \to \infty$ for T fixed, i.e., $\delta = \delta_t \to 0$ as $t \to \infty$!
- $\delta_t \sim t^{-p} \to 0$, Hoeffding: Let $c_{t,T_{kt}} \sim \sqrt{p \ln t / T_{kt}}$;
  - satisfies $\lim_{t\to\infty} c_{t,T} \to \infty$ and
  - the total probability of any confidence intervals failing is small for $p > 2$:

HOW TO SELECT $c_{t,T_{kt}}$?

- Central limit theorem: $\sqrt{T_{kt}}(\overline{Y}_{kt} - \mu_k) \sim \mathcal{N}_{0,\sigma_k}$, so w. high prob. $\overline{Y}_{kt} - \mu_k \in [-2\sigma_k/\sqrt{T_{kt}}, 2\sigma_k/\sqrt{T_{kt}}]$
- Hoeffding (for $Y \in [0,1]$): w. probability $\geq 1 - 2\delta$, $\overline{Y}_{kt} - \mu_k \in [-\sqrt{\ln(1/\delta)/2T_{kt}}, \sqrt{\ln(1/\delta)/2T_{kt}}]$
- Confidence Bounds — measures the uncertainty of $\overline{Y}_{kt}$
- Let it match the confidence radius: $c_{t,T_{kt}} \sim \sqrt{\ln(1/\delta)/T_{kt}}$
- $\overline{Y}_{kt} + c_{t,T_{kt}}$ = Upper Confidence Bound

  Note: Fixed $c_{t,T}$ for fixed T are not enough for infinite exploration (sticks to wrong arm with probability> 0 if the first samples for the optimal arm are bad). We need $\lim_{t\to\infty} c_{t,T} \to \infty$ for T fixed, i.e., $\delta = \delta_t \to 0$ as $t \to \infty$!
- $\delta_t \sim t^{-p} \to 0$, Hoeffding: Let $c_{t,T_{kt}} \sim \sqrt{p \ln t/T_{kt}}$;
  - satisfies $\lim_{t\to\infty} c_{t,T} \to \infty$ and
  - the total probability of any confidence intervals failing is small for $p > 2$: $K\sum_{t=K+1}^{n} t^{-p} \leq K\int_K^{\infty} t^{-p}dt = \frac{1}{(p-1)K^{p-2}}$

HOW TO SELECT $c_{t,T_{kt}}$?

- Central limit theorem: $\sqrt{T_{kt}}(\overline{Y}_{kt} - \mu_k) \sim \mathcal{N}_{0,\sigma_k}$, so w. high prob. $\overline{Y}_{kt} - \mu_k \in [-2\sigma_k/\sqrt{T_{kt}}, 2\sigma_k/\sqrt{T_{kt}}]$
- Hoeffding (for $Y \in [0,1]$): w. probability $\geq 1 - 2\delta$, $\overline{Y}_{kt} - \mu_k \in [-\sqrt{\ln(1/\delta)/2T_{kt}}, \sqrt{\ln(1/\delta)/2T_{kt}}]$
- Confidence Bounds — measures the uncertainty of $\overline{Y}_{kt}$
- Let it match the confidence radius: $c_{t,T_{kt}} \sim \sqrt{\ln(1/\delta)/T_{kt}}$
- $\overline{Y}_{kt} + c_{t,T_{kt}}$ = Upper Confidence Bound

  Note: Fixed $c_{t,T}$ for fixed T are not enough for infinite exploration (sticks to wrong arm with probability> 0 if the first samples for the optimal arm are bad). We need $\lim_{t\to\infty} c_{t,T} \to \infty$ for T fixed, i.e., $\delta = \delta_t \to 0$ as $t \to \infty$!
- $\delta_t \sim t^{-p} \to 0$, Hoeffding: Let $c_{t,T_{kt}} \sim \sqrt{p \ln t / T_{kt}}$;
  - satisfies $\lim_{t\to\infty} c_{t,T} \to \infty$ and
  - the total probability of *any* confidence intervals failing is small for $p > 2$:

HOW TO SELECT $c_{t,T_{kt}}$?

- Central limit theorem: $\sqrt{T_{kt}}(\overline{Y}_{kt} - \mu_k) \sim \mathcal{N}_{0,\sigma_k}$, so w. high prob. $\overline{Y}_{kt} - \mu_k \in [-2\sigma_k/\sqrt{T_{kt}}, 2\sigma_k/\sqrt{T_{kt}}]$
- Hoeffding (for $Y \in [0,1]$): w. probability $\geq 1 - 2\delta$, $\overline{Y}_{kt} - \mu_k \in [-\sqrt{\ln(1/\delta)/2T_{kt}}, \sqrt{\ln(1/\delta)/2T_{kt}}]$
- Confidence Bounds — measures the uncertainty of $\overline{Y}_{kt}$
- Let it match the confidence radius: $c_{t,T_{kt}} \sim \sqrt{\ln(1/\delta)/T_{kt}}$
- $\overline{Y}_{kt} + c_{t,T_{kt}} =$ Upper Confidence Bound

  Note: Fixed $c_{t,T}$ for fixed T are not enough for infinite exploration (sticks to wrong arm with probability > 0 if the first samples for the optimal arm are bad). We need $\lim_{t\to\infty} c_{t,T} \to \infty$ for T fixed, i.e., $\delta = \delta_t \to 0$ as $t \to \infty$!
- $\delta_t \sim t^{-p} \to 0$, Hoeffding: Let $c_{t,T_{kt}} \sim \sqrt{p \ln t/T_{kt}}$;

HOW TO SELECT $c_{t,T_{kt}}$?

- Central limit theorem: $\sqrt{T_{kt}}(\overline{Y}_{kt} - \mu_k) \sim \mathcal{N}_{0,\sigma_k}$, so w. high prob. $\overline{Y}_{kt} - \mu_k \in [-2\sigma_k/\sqrt{T_{kt}}, 2\sigma_k/\sqrt{T_{kt}}]$
- Hoeffding (for $Y \in [0,1]$): w. probability $\geq 1 - 2\delta$, $\overline{Y}_{kt} - \mu_k \in [-\sqrt{\ln(1/\delta)/2T_{kt}}, \sqrt{\ln(1/\delta)/2T_{kt}}]$
- Confidence Bounds — measures the uncertainty of $\overline{Y}_{kt}$
- Let it match the confidence radius: $c_{t,T_{kt}} \sim \sqrt{\ln(1/\delta)/T_{kt}}$
- $\overline{Y}_{kt} + c_{t,T_{kt}} =$ Upper Confidence Bound

  Note: Fixed $c_{t,T}$ for fixed T are not enough for infinite exploration (sticks to wrong arm with probability > 0 if the first samples for the optimal arm are bad). We need $\lim_{t\to\infty} c_{t,T} \to \infty$ for T fixed, i.e., $\delta = \delta_t \to 0$ as $t \to \infty$!
- $\delta_t \sim t^{-p} \to 0$, Hoeffding: Let $c_{t,T_{kt}} \sim \sqrt{p \ln t / T_{kt}}$;
  - satisfies $\lim_{t\to\infty} c_{t,T} \to \infty$ *and*
  - the total probability of any confidence intervals failing is

How to select $c_{t,T_{kt}}$?

- Central limit theorem: $\sqrt{T_{kt}}(\overline{Y}_{kt} - \mu_k) \sim \mathcal{N}_{0,\sigma_k}$, so w. high prob. $\overline{Y}_{kt} - \mu_k \in [-2\sigma_k/\sqrt{T_{kt}}, 2\sigma_k/\sqrt{T_{kt}}]$
- Hoeffding (for $Y \in [0,1]$): w. probability $\geq 1 - 2\delta$, $\overline{Y}_{kt} - \mu_k \in [-\sqrt{\ln(1/\delta)/2T_{kt}}, \sqrt{\ln(1/\delta)/2T_{kt}}]$
- Confidence Bounds — measures the uncertainty of $\overline{Y}_{kt}$
- Let it match the confidence radius: $c_{t,T_{kt}} \sim \sqrt{\ln(1/\delta)/T_{kt}}$
- $\overline{Y}_{kt} + c_{t,T_{kt}} =$ Upper Confidence Bound

  Note: Fixed $c_{t,T}$ for fixed T are not enough for infinite exploration (sticks to wrong arm with probability > 0 if the first samples for the optimal arm are bad). We need $\lim_{t\to\infty} c_{t,T} \to \infty$ for T fixed, i.e., $\delta = \delta_t \to 0$ as $t \to \infty$!
- $\delta_t \sim t^{-p} \to 0$, Hoeffding: Let $c_{t,T_{kt}} \sim \sqrt{p \ln t / T_{kt}}$;
  - satisfies $\lim_{t\to\infty} c_{t,T} \to \infty$ *and*
  - the total probability of *any* confidence intervals failing is small for $p > 2$: $K \sum_{t=K+1}^{n} t^{-p} \leq K \int_K^\infty t^{-p} dt = \frac{1}{(p-1)K^{p-2}}$.

HOW TO SELECT $c_{t,T_{kt}}$?

- Central limit theorem: $\sqrt{T_{kt}}(\overline{Y}_{kt} - \mu_k) \sim \mathcal{N}_{0,\sigma_k}$, so w. high prob. $\overline{Y}_{kt} - \mu_k \in [-2\sigma_k/\sqrt{T_{kt}}, 2\sigma_k/\sqrt{T_{kt}}]$
- Hoeffding (for $Y \in [0,1]$): w. probability $\geq 1 - 2\delta$, $\overline{Y}_{kt} - \mu_k \in [-\sqrt{\ln(1/\delta)/2T_{kt}}, \sqrt{\ln(1/\delta)/2T_{kt}}]$
- Confidence Bounds — measures the uncertainty of $\overline{Y}_{kt}$
- Let it match the confidence radius: $c_{t,T_{kt}} \sim \sqrt{\ln(1/\delta)/T_{kt}}$
- $\overline{Y}_{kt} + c_{t,T_{kt}}$ = **U**pper **C**onfidence **B**ound

  Note: Fixed $c_{t,T}$ for fixed T are not enough for infinite exploration (sticks to wrong arm with probability> 0 if the first samples for the optimal arm are bad). We need $\lim_{t\to\infty} c_{t,T} \to \infty$ for T fixed, i.e., $\delta = \delta_t \to 0$ as $t \to \infty$!
- $\delta_t \sim t^{-p} \to 0$, Hoeffding: Let $c_{t,T_{kt}} \sim \sqrt{p \ln t/T_{kt}}$;
  - satisfies $\lim_{t\to\infty} c_{t,T} \to \infty$ *and*
  - the total probability of *any* confidence intervals failing is small for $p > 2$: $K\sum_{t=K+1}^{n} t^{-p} \leq K\int_K^{\infty} t^{-p}dt = \frac{1}{(p-1)K^{p-2}}$.

HOW TO SELECT $c_{t,T_{kt}}$?

- Central limit theorem: $\sqrt{T_{kt}}(\overline{Y}_{kt} - \mu_k) \sim \mathcal{N}_{0,\sigma_k}$, so w. high prob. $\overline{Y}_{kt} - \mu_k \in [-2\sigma_k/\sqrt{T_{kt}}, 2\sigma_k/\sqrt{T_{kt}}]$
- Hoeffding (for $Y \in [0,1]$): w. probability $\geq 1 - 2\delta$, $\overline{Y}_{kt} - \mu_k \in [-\sqrt{\ln(1/\delta)/2T_{kt}}, \sqrt{\ln(1/\delta)/2T_{kt}}]$
- Confidence Bounds — measures the uncertainty of $\overline{Y}_{kt}$
- Let it match the confidence radius: $c_{t,T_{kt}} \sim \sqrt{\ln(1/\delta)/T_{kt}}$
- $\overline{Y}_{kt} + c_{t,T_{kt}} =$ **U**pper **C**onfidence **B**ound

  Note: Fixed $c_{t,T}$ for fixed T are not enough for infinite exploration (sticks to wrong arm with probability> 0 if the first samples for the optimal arm are bad). We need $\lim_{t\to\infty} c_{t,T} \to \infty$ for T fixed, i.e., $\delta = \delta_t \to 0$ as $t \to \infty$!
- $\delta_t \sim t^{-p} \to 0$, Hoeffding: Let $c_{t,T_{kt}} \sim \sqrt{p \ln t/T_{kt}}$;
  - satisfies $\lim_{t\to\infty} c_{t,T} \to \infty$ *and*
  - the total probability of *any* confidence intervals failing is small for $p > 2$: $K\sum_{t=K+1}^{n} t^{-p} \leq K\int_K^\infty t^{-p}dt = \frac{1}{(p-1)K^{p-2}}$.

OUTLINE

UCB1 Regret Theorem

[Agrawal, 1995] Asymptotic results: large-deviation theory

[Auer et al., 2002] Avoid asymptotics, use Hoeffding's ineq.

Theorem (UCB1 Regret)

Let $0 \leq Y_{it} \leq 1$. Then the regret of UCB1 when used with $c_{t,T} = \sqrt{\frac{p \ln t}{2T}}$ and $p > 2$ satisfies

$$R_n(\mathcal{A}_{\mathrm{UCB1}}) \leq 2p \left(\sum_{i \in J_{\mathrm{bad}}} \frac{1}{\Delta_i} \right) \ln n + \left(3 + \frac{2}{p-2} \right) \sum_{i=1}^{K} \Delta_i .$$

- Slightly better than [Auer et al., 2002]: tradeoff in p explicit
- Coefficient $\sum_{i \in J_{\mathrm{bad}}} \frac{1}{\Delta_i}$ is large, if many small $\Delta_i > 0$, i.e., hard to distinguish the best arms.

UCB1 Regret Theorem

[Agrawal, 1995] Asymptotic results: large-deviation theory

[Auer et al., 2002] Avoid asymptotics, use Hoeffding's ineq.

Theorem (UCB1 Regret)

Let $0 \leq Y_{it} \leq 1$. Then the regret of UCB1 when used with $c_{t,T} = \sqrt{\frac{p \ln t}{2T}}$ and $p > 2$ satisfies

$$R_n(\mathcal{A}_{\mathrm{UCB1}}) \leq 2p \left(\sum_{i \in J_{\mathrm{bad}}} \frac{1}{\Delta_i} \right) \ln n + \left(3 + \frac{2}{p-2} \right) \sum_{i=1}^{K} \Delta_i .$$

- Slightly better than [Auer et al., 2002]: tradeoff in p explicit
- Coefficient $\sum_{i \in J_{\mathrm{bad}}} \frac{1}{\Delta_i}$ is large, if many small $\Delta_i > 0$, i.e., hard to distinguish the best arms.

UCB1 Regret Theorem

[Agrawal, 1995] Asymptotic results: large-deviation theory

[Auer et al., 2002] Avoid asymptotics, use Hoeffding's ineq.

Theorem (UCB1 Regret)

Let $0 \leq Y_{it} \leq 1$. Then the regret of UCB1 when used with $c_{t,T} = \sqrt{\frac{p \ln t}{2T}}$ and $p > 2$ satisfies

$$R_n(\mathcal{A}_{\mathrm{UCB1}}) \leq 2p \left(\sum_{i \in J_{\mathrm{bad}}} \frac{1}{\Delta_i} \right) \ln n + \left(3 + \frac{2}{p-2} \right) \sum_{i=1}^{K} \Delta_i.$$

- Slightly better than [Auer et al., 2002]: tradeoff in p explicit
- Coefficient $\sum_{i \in J_{\mathrm{bad}}} \frac{1}{\Delta_i}$ is large, if many small $\Delta_i > 0$, i.e., hard to distinguish the best arms.

UCB1 Regret Theorem

[Agrawal, 1995] Asymptotic results: large-deviation theory

[Auer et al., 2002] Avoid asymptotics, use Hoeffding's ineq.

Theorem (UCB1 Regret)

Let $0 \leq Y_{it} \leq 1$. Then the regret of UCB1 when used with $c_{t,T} = \sqrt{\frac{p \ln t}{2T}}$ and $p > 2$ satisfies

$$R_n(\mathcal{A}_{\mathrm{UCB1}}) \leq 2p \left(\sum_{i \in J_{\mathrm{bad}}} \frac{1}{\Delta_i} \right) \ln n + \left(3 + \frac{2}{p-2} \right) \sum_{i=1}^{K} \Delta_i .$$

- Slightly better than [Auer et al., 2002]: tradeoff in p explicit
- Coefficient $\sum_{i \in J_{\mathrm{bad}}} \frac{1}{\Delta_i}$ is large, if many small $\Delta_i > 0$, i.e., hard to distinguish the best arms.

UCB1 Regret Theorem

[Agrawal, 1995] Asymptotic results: large-deviation theory

[Auer et al., 2002] Avoid asymptotics, use Hoeffding's ineq.

Theorem (UCB1 Regret)

Let $0 \leq Y_{it} \leq 1$. *Then the regret of UCB1 when used with* $c_{t,T} = \sqrt{\frac{p \ln t}{2T}}$ *and* $p > 2$ *satisfies*

$$R_n(\mathcal{A}_{\mathrm{UCB1}}) \leq 2p \left(\sum_{i \in J_{\mathrm{bad}}} \frac{1}{\Delta_i} \right) \ln n + \left(3 + \frac{2}{p-2} \right) \sum_{i=1}^{K} \Delta_i .$$

- Slightly better than [Auer et al., 2002]: tradeoff in p explicit
- Coefficient $\sum_{i \in J_{\mathrm{bad}}} \frac{1}{\Delta_i}$ is large, if many small $\Delta_i > 0$, i.e., hard to distinguish the best arms.

HEURISTIC ANALYSIS

Recall: $R_n(\mathcal{A}) = \sum_{i \in J_{\text{bad}}} \Delta_i \mathbb{E}\left[T_{in}\right]$, hence we bound $\mathbb{E}\left[T_{in}\right]$ for bad i arms.

FACT 1

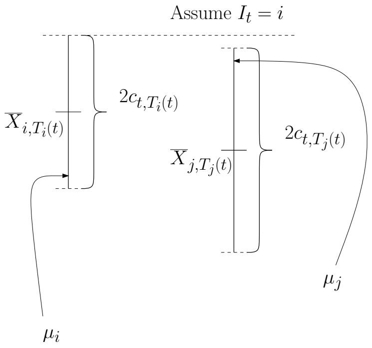
If confidence intervals do not fail and $I_t = i$ then

$$\mu^* - \mu_i = \max_j \mu_j - \mu_i \leq 2c_{t,T_{it}},$$

hence $c_{t,T_{it}} \geq \Delta_i/2$.

PROOF BY FIGURE!

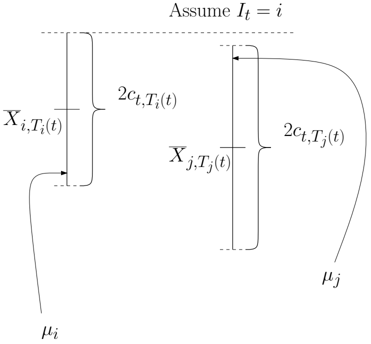
GOAL: ASSUMING $I_t = i$, PROVE $c_{t,T_{it}} \geq \Delta_i/2$!



(Actually, the conclusion holds even if we only have $\mu_i \geq \overline{Y}_{i,T_{it}} - c_{t,T_{it}}$ and $\mu_j \leq \overline{Y}_{j,T_{jt}} + c_{t,T_{jt}}$.)

Proof by figure!

Goal: Assuming $I_t = i$, prove $c_{t,T_{it}} \geq \Delta_i/2$!



(Actually, the conclusion holds even if we only have $\mu_i \geq \overline{Y}_{i,T_{it}} - c_{t,T_{it}}$ and $\mu_j \leq \overline{Y}_{j,T_{jt}} + c_{t,T_{jt}}$.)

HEURISTIC ANALYSIS/2

- By Fact 1, with high prob. if $I_t = i$ then $c_{t,T_{it}} \geq \Delta_i/2$, i.e.,

$$\frac{\Delta_i^2}{4} \leq c_{t,T_{it}}^2 \sim \frac{p \ln t}{2T_{it}}, \qquad \text{hence} \qquad T_{it} \leq\sim \frac{2p \ln t}{\Delta_i^2}.$$

Thus, using $t \leq n$, for a bad arm $\mathbb{E}\left[T_{in}\right] \leq\sim 2p \ln n/\Delta_i^2$, and

$$R_n = \sum_i \Delta_i \mathbb{E}\left[T_{in}\right] \leq\sim \sum_{i \in J_{\text{bad}}} \frac{1}{\Delta_i} \cdot O(\ln n).$$

Outline

EXTENSIONS

- UCT $\equiv$ UCB applied to searching in Trees [Kocsis and Szepesvári, 2006];
  - Improved trajectory-tree building in MDPs
  - searching in games
    Go: used by Mogo*, Valkyria_UCT* (was #1 on CGOS 9×9, had ELŐ points > 2000 for the first time!)
- Budgeted learning: some costs instead of time steps
- Best arm identification
- UCB applied to MDPs: [Auer and Ortner, 2007, Tewari and Bartlett, 2008, Auer et al., 2009, Bartlett and Tewari, 2009]
- Bandit problem is special case of MDP/RL with one state. More states?
- Continuous state spaces?
- Continuous action spaces?

EXTENSIONS

- UCT $\equiv$ UCB applied to searching in Trees [Kocsis and Szepesvári, 2006];
  - Improved trajectory-tree building in MDPs
  - searching in games
    Go: used by Mogo*, Valkyria_UCT* (was #1 on CGOS 9×9, had ELŐ points > 2000 for the first time!)
- Budgeted learning: some costs instead of time steps
- Best arm identification
- UCB applied to MDPs: [Auer and Ortner, 2007, Tewari and Bartlett, 2008, Auer et al., 2009, Bartlett and Tewari, 2009]
- Bandit problem is special case of MDP/RL with one state. More states?
- Continuous state spaces?
- Continuous action spaces?

EXTENSIONS

- UCT $\equiv$ UCB applied to searching in Trees [Kocsis and Szepesvári, 2006];
  - Improved trajectory-tree building in MDPs
  - searching in games
    Go: used by Mogo*, Valkyria_UCT* (was #1 on CGOS 9×9, had ELŐ points > 2000 for the first time!)
- Budgeted learning: some costs instead of time steps
- Best arm identification
- UCB applied to MDPs: [Auer and Ortner, 2007, Tewari and Bartlett, 2008, Auer et al., 2009, Bartlett and Tewari, 2009]
- Bandit problem is special case of MDP/RL with one state. More states?
- Continuous state spaces?
- Continuous action spaces?

EXTENSIONS

- UCT $\equiv$ UCB applied to searching in Trees [Kocsis and Szepesvári, 2006];
  - Improved trajectory-tree building in MDPs
  - searching in games
    Go: used by Mogo*, Valkyria_UCT* (was #1 on CGOS 9×9, had ELŐ points > 2000 for the first time!)
- Budgeted learning: some costs instead of time steps
- Best arm identification
- UCB applied to MDPs: [Auer and Ortner, 2007, Tewari and Bartlett, 2008, Auer et al., 2009, Bartlett and Tewari, 2009]
- Bandit problem is special case of MDP/RL with one state. More states?
- Continuous state spaces?
- Continuous action spaces?

EXTENSIONS

- UCT $\equiv$ UCB applied to searching in Trees [Kocsis and Szepesvári, 2006];
  - Improved trajectory-tree building in MDPs
  - searching in games
    Go: used by Mogo*, Valkyria_UCT* (was #1 on CGOS 9×9, had ELŐ points > 2000 for the first time!)
- Budgeted learning: some costs instead of time steps
- Best arm identification
- UCB applied to MDPs: [Auer and Ortner, 2007, Tewari and Bartlett, 2008, Auer et al., 2009, Bartlett and Tewari, 2009]
- Bandit problem is special case of MDP/RL with one state. More states?
- Continuous state spaces?
- Continuous action spaces?

EXTENSIONS

- UCT $\equiv$ UCB applied to searching in Trees [Kocsis and Szepesvári, 2006];
  - Improved trajectory-tree building in MDPs
  - searching in games
    Go: used by Mogo*, Valkyria_UCT* (was #1 on CGOS 9×9, had ELŐ points > 2000 for the first time!)
- Budgeted learning: some costs instead of time steps
- Best arm identification
- UCB applied to MDPs: [Auer and Ortner, 2007, Tewari and Bartlett, 2008, Auer et al., 2009, Bartlett and Tewari, 2009]
- Bandit problem is special case of MDP/RL with one state. More states?
- Continuous state spaces?
- Continuous action spaces?

EXTENSIONS

- UCT $\equiv$ UCB applied to searching in Trees [Kocsis and Szepesvári, 2006];
  - Improved trajectory-tree building in MDPs
  - searching in games
    Go: used by Mogo*, Valkyria_UCT* (was #1 on CGOS 9×9, had ELŐ points > 2000 for the first time!)
- Budgeted learning: some costs instead of time steps
- Best arm identification
- UCB applied to MDPs: [Auer and Ortner, 2007, Tewari and Bartlett, 2008, Auer et al., 2009, Bartlett and Tewari, 2009]
- Bandit problem is special case of MDP/RL with one state. More states?
- Continuous state spaces?
- Continuous action spaces?

EXTENSIONS

- UCT $\equiv$ UCB applied to searching in Trees [Kocsis and Szepesvári, 2006];
  - Improved trajectory-tree building in MDPs
  - searching in games
    Go: used by Mogo*, Valkyria_UCT* (was #1 on CGOS 9×9, had ELŐ points > 2000 for the first time!)
- Budgeted learning: some costs instead of time steps
- Best arm identification
- UCB applied to MDPs: [Auer and Ortner, 2007, Tewari and Bartlett, 2008, Auer et al., 2009, Bartlett and Tewari, 2009]
- Bandit problem is special case of MDP/RL with one state. More states?
- Continuous state spaces?
- Continuous action spaces?

EXTENSIONS

- UCT $\equiv$ UCB applied to searching in Trees [Kocsis and Szepesvári, 2006];
  - Improved trajectory-tree building in MDPs
  - searching in games
    Go: used by Mogo*, Valkyria_UCT* (was #1 on CGOS 9×9, had ELŐ points > 2000 for the first time!)
- Budgeted learning: some costs instead of time steps
- Best arm identification
- UCB applied to MDPs: [Auer and Ortner, 2007, Tewari and Bartlett, 2008, Auer et al., 2009, Bartlett and Tewari, 2009]
- Bandit problem is special case of MDP/RL with one state. More states?
- Continuous state spaces?
- Continuous action spaces?

APPLICATION OF BANDIT MODELS

- Gambling :-)
- UCT
- Adaptive routing for minimizing delays in networks (arm = route, payoff = − delay)
- Online ad serving (showing relevant ads; arm = ad type shown, payoff = click)
- Clinical trials investigating effects of experimental treatments (arm = treatment, payoff = healing; legal, ethic issues, interference)
- Managing competing research projects in a large organization (science found., pharmacy; arm = project given resource, payoff = results (i.i.d.?))
- Tuning parameter setting for a program given a deadline
- Choosing a partner during limited number of dates

APPLICATION OF BANDIT MODELS

- Gambling :-)
- UCT
- Adaptive routing for minimizing delays in networks (arm = route, payoff = – delay)
- Online ad serving (showing relevant ads; arm = ad type shown, payoff = click)
- Clinical trials investigating effects of experimental treatments (arm = treatment, payoff = healing; legal, ethic issues, interference)
- Managing competing research projects in a large organization (science found., pharmacy; arm = project given resource, payoff = results (i.i.d.?))
- Tuning parameter setting for a program given a deadline
- Choosing a partner during limited number of dates

APPLICATION OF BANDIT MODELS

- Gambling :-)
- UCT
- Adaptive routing for minimizing delays in networks (arm = route, payoff = $-$ delay)
- Online ad serving (showing relevant ads; arm = ad type shown, payoff = click)
- Clinical trials investigating effects of experimental treatments (arm = treatment, payoff = healing; legal, ethic issues, interference)
- Managing competing research projects in a large organization (science found., pharmacy; arm = project given resource, payoff = results (i.i.d.?))
- Tuning parameter setting for a program given a deadline
- Choosing a partner during limited number of dates

APPLICATION OF BANDIT MODELS

- Gambling :-)
- UCT
- Adaptive routing for minimizing delays in networks (arm = route, payoff = $-$ delay)
- Online ad serving (showing relevant ads; arm = ad type shown, payoff = click)
- Clinical trials investigating effects of experimental treatments (arm = treatment, payoff = healing; legal, ethic issues, interference)
- Managing competing research projects in a large organization (science found., pharmacy; arm = project given resource, payoff = results (i.i.d.?))
- Tuning parameter setting for a program given a deadline
- Choosing a partner during limited number of dates

APPLICATION OF BANDIT MODELS

- Gambling :-)
- UCT
- Adaptive routing for minimizing delays in networks (arm = route, payoff = $-$ delay)
- Online ad serving (showing relevant ads; arm = ad type shown, payoff = click)
- Clinical trials investigating effects of experimental treatments (arm = treatment, payoff = healing; legal, ethic issues, interference)
- Managing competing research projects in a large organization (science found., pharmacy; arm = project given resource, payoff = results (i.i.d.?))
- Tuning parameter setting for a program given a deadline
- Choosing a partner during limited number of dates

APPLICATION OF BANDIT MODELS

- Gambling :-)
- UCT
- Adaptive routing for minimizing delays in networks (arm = route, payoff = $-$ delay)
- Online ad serving (showing relevant ads; arm = ad type shown, payoff = click)
- Clinical trials investigating effects of experimental treatments (arm = treatment, payoff = healing; legal, ethic issues, interference)
- Managing competing research projects in a large organization (science found., pharmacy; arm = project given resource, payoff = results (i.i.d.?))
- Tuning parameter setting for a program given a deadline
- Choosing a partner during limited number of dates

APPLICATION OF BANDIT MODELS

- Gambling :-)
- UCT
- Adaptive routing for minimizing delays in networks (arm = route, payoff = $-$ delay)
- Online ad serving (showing relevant ads; arm = ad type shown, payoff = click)
- Clinical trials investigating effects of experimental treatments (arm = treatment, payoff = healing; legal, ethic issues, interference)
- Managing competing research projects in a large organization (science found., pharmacy; arm = project given resource, payoff = results (i.i.d.?))
- Tuning parameter setting for a program given a deadline
- Choosing a partner during limited number of dates

APPLICATION OF BANDIT MODELS

- Gambling :-)
- UCT
- Adaptive routing for minimizing delays in networks (arm = route, payoff = $-$ delay)
- Online ad serving (showing relevant ads; arm = ad type shown, payoff = click)
- Clinical trials investigating effects of experimental treatments (arm = treatment, payoff = healing; legal, ethic issues, interference)
- Managing competing research projects in a large organization (science found., pharmacy; arm = project given resource, payoff = results (i.i.d.?))
- Tuning parameter setting for a program given a deadline
- Choosing a partner during limited number of dates

OUTLINE

FOR FURTHER READING

Agrawal, R. (1995).
Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem.
Advances in Applied Probability, 27:1054–1078.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002).
Finite time analysis of the multiarmed bandit problem.
Machine Learning, 47(2-3):235–256.

Auer, P., Jaksch, T., and Ortner, R. (2009).
Near-optimal regret bounds for reinforcement learning.
In *NIPS-21*, pages 89–96.

Auer, P. and Ortner, R. (2007).
Logarithmic online regret bounds for undiscounted reinforcement learning.
In *NIPS-19*, pages 49–56.

Bartlett, P. L. and Tewari, A. (2009).
REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs.
In *UAI 2009*.

Kocsis, L. and Szepesvári, Cs. (2006).
Bandit based Monte-Carlo planning.
In *Proceedings of the 17th European conference on Machine Learning*, pages 282–293.

Tewari, A. and Bartlett, P. (2008).
Optimistic linear programming gives logarithmic regret for irreducible mdps.
In *NIPS-20*, pages 1505–1512.