Strategies for Safe Multi-Armed Bandits with Logarithmic Regret and Risk

Tianrui Chen, Aditya Gangrade, Venkatesh Saligrama

Safe Bandits

Extend Stochastic Multi-Armed Bandits to handle safety constraints.

Safe Bandits

Extend Stochastic Multi-Armed Bandits to handle safety constraints.

- K arms.
- Drugs & dosages.

Safe Bandits

Extend Stochastic Multi-Armed Bandits to handle safety constraints.

- K arms.
- Play $A_t \in [1 : K]$.

- Drugs & dosages.
- Treat patient t.

Safe Bandits

Extend Stochastic Multi-Armed Bandits to handle safety constraints.

- K arms.
- Play $A_t \in [1:K]$.
- Incur
  - Reward $R_t : \mathbb{E}[R_t | A_t = k] = \mu^k$
  - Safety Risk $S_t : \mathbb{E}[S_t | A_t = k] = \nu^k$

- Drugs & dosages.
- Treat patient t.
- Observe
  - Efficacy
  - Side Effects

Safe Bandits

Extend Stochastic Multi-Armed Bandits to handle safety constraints.

- K arms.
- Play $A_t \in [1:K]$.
- Incur
  - Reward $R_t : \mathbb{E}[R_t | A_t = k] = \mu^k$
  - Safety Risk $S_t : \mathbb{E}[S_t | A_t = k] = \nu^k$

- Drugs & dosages.
- Treat patient t.
- Observe
  - Efficacy
  - Side Effects

Tolerated risk level α; k is safe if $\nu^k \leq \alpha$

Safe Bandits

Extend Stochastic Multi-Armed Bandits to handle safety constraints.

- K arms.
- Play $A_t \in [1 : K]$.
- Incur
  - Reward $R_t : \mathbb{E}[R_t | A_t = k] = \mu^k$
  - Safety Risk $S_t : \mathbb{E}[S_t | A_t = k] = \nu^k$

- Drugs & dosages.
- Treat patient t.
- Observe
  - Efficacy
  - Side Effects

Tolerated risk level α; k is safe if $\nu^k \leq \alpha$

Optimal reward: $\mu^* = \max_k \mu^k \text{ s.t. } \nu^k \leq \alpha$.

Safe Bandits

Extend Stochastic Multi-Armed Bandits to handle safety constraints.

- K arms.
- Play $A_t \in [1 : K]$.
- Incur
  - Reward $R_t : \mathbb{E}[R_t | A_t = k] = \mu^k$
  - Safety Risk $S_t : \mathbb{E}[S_t | A_t = k] = \nu^k$

- Drugs & dosages.
- Treat patient t.
- Observe
  - Efficacy
  - Side Effects

Tolerated risk level α; k is safe if $\nu^k \leq \alpha$

Optimal reward: $\mu^* = \max_k \mu^k \text{ s.t. } \nu^k \leq \alpha$.

Goal: maximise reward while playing safely.

Context

- Safety criterion needs round-wise enforcement.

Context

- Safety criterion needs round-wise enforcement.
- Prior work : aggregate control - $\sum \nu^{A_t} \leq \alpha T(1 + o(1))$

Context

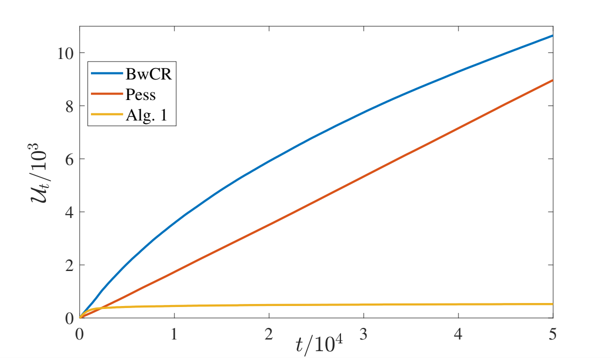
- Safety criterion needs round-wise enforcement.
- Prior work : aggregate control - $\sum \nu^{A_t} \leq \alpha T(1 + o(1))$
- Inappropriate for safety constraints:
    - Switching b/w (unsafe & very effective) and (safe & ineffective) arms.

Context

- Safety criterion needs round-wise enforcement.
- Prior work : aggregate control - $\sum \nu^{A_t} \leq \alpha T(1 + o(1))$
- Inappropriate for safety constraints:
    - Switching b/w (unsafe & very effective) and (safe & ineffective) arms.

Context

- Safety criterion needs round-wise enforcement.
- Prior work : aggregate control - $\sum \nu^{A_t} \leq \alpha T(1 + o(1))$
- Inappropriate for safety constraints:
  - Switching b/w (unsafe & very effective) and (safe & ineffective) arms.



Regret

Regret that captures safety scenarios.

$$\mathcal{R}_T := \sum_{t \leq T} \max\left(\mu^* - \mu^{A_t}, \nu^{A_t} - \alpha\right)$$

Regret

Regret that captures safety scenarios.

$$\mathcal{R}_T := \sum_{t \leq T} \max\left(\mu^* - \mu^{A_t}, \nu^{A_t} - \alpha\right)$$

- Round-wise safety enforcement in a smooth sense.
- Good schemes must be both effective and safe.

Regret

Regret that captures safety scenarios.

$$\mathcal{R}_T := \sum_{t \le T} \max\left(\mu^* - \mu^{A_t}, \nu^{A_t} - \alpha\right)$$

- Round-wise safety enforcement in a smooth sense.
- Good schemes must be both effective and safe.

Also control # of unsafe rounds

$$\mathcal{U}_T = \sum_{t \le T} \mathbb{1}\{\nu^{A_t} > \alpha\}.$$

Schemes

Doubly-Optimistic index-based schemes:

Schemes

Doubly-Optimistic index-based schemes:

- For each arm k,
  - Reward indices $\rho_t^k \geq \mu^k$

Schemes

Doubly-Optimistic index-based schemes:

- For each arm k,
  - Reward indices $\rho_t^k \geq \mu^k$
  - Safety indices $\sigma_t^k \leq \nu^k$

Schemes

Doubly-Optimistic index-based schemes:

- For each arm k,
  - Reward indices $\rho_t^k \geq \mu^k$
  - Safety indices $\sigma_t^k \leq \nu^k$
- Permissible set

$$\Pi_t := \{k : \sigma_t^k \leq \alpha\}.$$

Schemes

Doubly-Optimistic index-based schemes:

- For each arm k,
  - Reward indices $\rho_t^k \geq \mu^k$
  - Safety indices $\sigma_t^k \leq \nu^k$
- Permissible set

$$\Pi_t := \{k : \sigma_t^k \leq \alpha\}.$$

- Play optimistically from Π_t :

$$A_t \in \arg\max_{k \in \Pi_t} \rho_t^k.$$

Schemes

Doubly-Optimistic index-based schemes:

- For each arm k,
  - Reward indices $\rho_t^k \geq \mu^k$
  - Safety indices $\sigma_t^k \leq \nu^k$
- Permissible set

$$\Pi_t := \{k : \sigma_t^k \leq \alpha\}.$$

- Play optimistically from Π_t :

$$A_t \in \arg\max_{k \in \Pi_t} \rho_t^k.$$

Reasoning:

- Optimal arm k^*.

Schemes

Doubly-Optimistic index-based schemes:

- For each arm k,
  - Reward indices $\rho_t^k \geq \mu^k$
  - Safety indices $\sigma_t^k \leq \nu^k$
- Permissible set

$$\Pi_t := \{k : \sigma_t^k \leq \alpha\}.$$

- Play optimistically from Π_t :

$$A_t \in \arg\max_{k \in \Pi_t} \rho_t^k.$$

Reasoning:

- Optimal arm k^*.
- If $k^* \in \Pi_t$, and ρ_t is good,

$$\#t : \mu^{A_t} < \mu^* \text{ is small.}$$

Schemes

Doubly-Optimistic index-based schemes:

- For each arm k,
  - Reward indices $\rho_t^k \geq \mu^k$
  - Safety indices $\sigma_t^k \leq \nu^k$
- Permissible set

$$\Pi_t := \{k : \sigma_t^k \leq \alpha\}.$$

- Play optimistically from Π_t :

$$A_t \in \arg\max_{k \in \Pi_t} \rho_t^k.$$

Reasoning:

- Optimal arm k^*.
- If $k^* \in \Pi_t$, and ρ_t is good,

$$\#t : \mu^{A_t} < \mu^* \text{ is small.}$$

Critical use of optimistic σ_t^k.

- If σ_t is good,

$$\#t : \nu^{A_t} > \alpha \text{ is small.}$$

Schemes

Doubly-Optimistic index-based schemes:

- For each arm k,
  - Reward indices $\rho_t^k \geq \mu^k$
  - Safety indices $\sigma_t^k \leq \nu^k$
- Permissible set

$$\Pi_t := \{k : \sigma_t^k \leq \alpha\}.$$

- Play optimistically from Π_t :

$$A_t \in \arg\max_{k \in \Pi_t} \rho_t^k.$$

Reasoning:

- Optimal arm k^*.
- If $k^* \in \Pi_t$, and ρ_t is good,

$$\#t : \mu^{A_t} < \mu^* \text{ is small.}$$

  Critical use of optimistic σ_t^k.
- If σ_t is good,

$$\#t : \nu^{A_t} > \alpha \text{ is small.}$$

Both frequentist and Bayesian ways to design ρ_t^k, σ_t^k.

Theoretical Results

Theorem

For schemes with both Bayesian and Frequentist indices

$$\mathcal{R}_T \leq (1 + o(1)) \sum_{k \neq k^*} \frac{\log T}{2 \max(\Delta^k, \Gamma^k)},$$

$$\mathcal{U}_T \leq (1 + o(1)) \sum_{k : \Gamma^k > 0} \frac{\log T}{2 \max(\Delta^k, \Gamma^k)^2}.$$

Theoretical Results

Theorem

For schemes with both Bayesian and Frequentist indices

$$\mathcal{R}_T \leq (1 + o(1)) \sum_{k \neq k^*} \frac{\log T}{2 \max(\Delta^k, \Gamma^k)},$$

$$\mathcal{U}_T \leq (1 + o(1)) \sum_{k : \Gamma^k > 0} \frac{\log T}{2 \max(\Delta^k, \Gamma^k)^2}.$$

- Efficacy Gap: $\Delta^k := \max(\mu^* - \mu^k, 0)$.
- Safety Gap: $\Gamma^k := \max(\nu^k - \alpha, 0)$.

Theoretical Results

Theorem

For schemes with both Bayesian and Frequentist indices

$$\mathcal{R}_T \leq (1 + o(1)) \sum_{k \neq k^*} \frac{\log T}{2 \max(\Delta^k, \Gamma^k)},$$

$$\mathcal{U}_T \leq (1 + o(1)) \sum_{k : \Gamma^k > 0} \frac{\log T}{2 \max(\Delta^k, \Gamma^k)^2}.$$

- Efficacy Gap: $\Delta^k := \max(\mu^* - \mu^k, 0)$.
- Safety Gap: $\Gamma^k := \max(\nu^k - \alpha, 0)$.

Tight Lower Bounds;
Tight gap-free $\tilde{O}(\sqrt{KT})$ bounds.