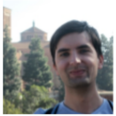
Old Dog Learns New Tricks: Randomized UCB for Bandit Problems

Sharan Vaswani
Mila, U. Montreal

Abbas Mehrabian
McGill University

Audrey Durand
Mila, U. Laval

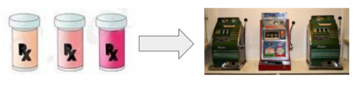
Branislav Kveton
Google Research

AISTATS 2020

Motivating example: clinical trials

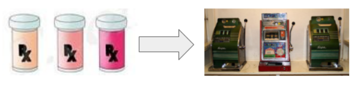
- Do not have complete information about the effectiveness or side-effects of the drugs.
- **Aim**: Infer the "best" drug by running a sequence of trials.

Motivating example: clinical trials



- Do not have complete information about the effectiveness or side-effects of the drugs.
- **Aim**: Infer the "best" drug by running a sequence of trials.
- **Abstraction to Multi-armed Bandits**: Each drug choice is mapped to an arm and the drug's effectiveness is mapped to the arm's reward.

Motivating example: clinical trials



- Do not have complete information about the effectiveness or side-effects of the drugs.
- **Aim**: Infer the "best" drug by running a sequence of trials.
- **Abstraction to Multi-armed Bandits**: Each drug choice is mapped to an arm and the drug's effectiveness is mapped to the arm's reward.
- Administering a drug is an action that is equivalent to pulling the corresponding arm. The trial goes on for T rounds.

Bandits 101: problem setup

Initialize the expected rewards according to some prior knowledge.

for $t = 1 \to T$ **do**

  SELECT: Use a bandit algorithm to decide which arm to pull.

  ACT and OBSERVE: Pull the selected arm and observe the reward.

  UPDATE: Update the estimated reward for the arm(s).

end

Bandits 101: problem setup

Initialize the expected rewards according to some prior knowledge.

for $t = 1 \rightarrow T$ **do**

  SELECT: Use a bandit algorithm to decide which arm to pull.

  ACT and OBSERVE: Pull the selected arm and observe the reward.

  UPDATE: Update the estimated reward for the arm(s).

end

- **Stochastic bandits**: Reward for each arm is sampled i.i.d from its underlying distribution.

Bandits 101: problem setup

Initialize the expected rewards according to some prior knowledge.

for $t = 1 \to T$ **do**

  SELECT: Use a bandit algorithm to decide which arm to pull.

  ACT and OBSERVE: Pull the selected arm and observe the reward.

  UPDATE: Update the estimated reward for the arm(s).

end

- **Stochastic bandits**: Reward for each arm is sampled i.i.d from its underlying distribution.
- **Objective**: Minimize the expected cumulative regret $R(T)$:

$$R(T) = \sum_{t=1}^{T} \left(\mathbf{E}[\text{Reward for best arm}] - \mathbf{E}[\text{Reward for arm pulled in round } t] \right)$$

Bandits 101: problem setup

Initialize the expected rewards according to some prior knowledge.

for $t = 1 \rightarrow T$ **do**

  SELECT: Use a bandit algorithm to decide which arm to pull.

  ACT and OBSERVE: Pull the selected arm and observe the reward.

  UPDATE: Update the estimated reward for the arm(s).

end

- **Stochastic bandits**: Reward for each arm is sampled i.i.d from its underlying distribution.
- **Objective**: Minimize the expected cumulative regret $R(T)$:

$$R(T) = \sum_{t=1}^{T} \Big(\mathbf{E}[\text{Reward for best arm}] - \mathbf{E}[\text{Reward for arm pulled in round } t] \Big)$$

- Minimizing $R(T)$ boils down to a exploration-exploitation trade-off.

Bandits 101: structured bandits

- In problems with a large number of arms, learning about each arm separately is inefficient.
  $\implies$ use a shared parameterization for the arms.
- **Structured bandits**: Each arm i has a feature vector x_i and
  there exists an unknown vector θ^* such that $\mathbf{E}[\text{reward for arm } i] = g(x_i, \theta^*)$.
- **Linear bandits**: $g(x_i, \theta^*) = \langle x_i, \theta^* \rangle$.
- **Generalized linear bandits**: g is a strictly increasing, differentiable link function.
  E.g. $g(x, \theta^*) = 1/(1 + \exp(-\langle x_i, \theta^* \rangle))$ for logistic bandits.

Bandits 101: algorithms

- **Optimism in the Face of Uncertainty** (OFU): Uses closed-form high-probability confidence sets.
  - Theoretically optimal. Does not depend on the exact distribution of rewards.
  - Poor empirical performance on typical problem instances.
- **Thompson Sampling** (TS): Randomized strategy that samples from a posterior distribution.
  - Good empirical performance on typical problem instances.
  - Depends on the reward distributions. Computationally expensive in the absence of closed-form posteriors. Theoretically sub-optimal in the (generalized) linear bandit setting.

Bandits 101: algorithms

- **Optimism in the Face of Uncertainty** (OFU): Uses closed-form high-probability confidence sets.
    - Theoretically optimal. Does not depend on the exact distribution of rewards.
    - Poor empirical performance on typical problem instances.
- **Thompson Sampling** (TS): Randomized strategy that samples from a posterior distribution.
    - Good empirical performance on typical problem instances.
    - Depends on the reward distributions. Computationally expensive in the absence of closed-form posteriors. Theoretically sub-optimal in the (generalized) linear bandit setting.

Can we obtain the best of OFU and TS?

The RandUCB meta-algorithm

Theoretical study

RandUCB Meta-algorithm

- **Generic OFU algorithm**: If $\widehat{\mu}_i(t)$ is the mean reward for arm i at round t, $\mathcal{C}_i(t)$ is the corresponding confidence set, pick the arm with the largest upper confidence bound.

$$i_t = \arg\max_{i \in [K]} \left\{ \widehat{\mu}_i(t) + \beta \, \mathcal{C}_i(t) \right\} .$$

Here, β is deterministic and chosen to trade off exploration and exploitation optimally.

`RandUCB` Meta-algorithm

- **Generic OFU algorithm**: If $\widehat{\mu}_i(t)$ is the mean reward for arm i at round t, $\mathcal{C}_i(t)$ is the corresponding confidence set, pick the arm with the largest upper confidence bound.

$$i_t = \arg\max_{i \in [K]} \left\{ \widehat{\mu}_i(t) + \beta \, \mathcal{C}_i(t) \right\}.$$

Here, β is deterministic and chosen to trade off exploration and exploitation optimally.

- `RandUCB`: Replace deterministic β by a random variable Z_t:

$$i_t = \arg\max_{i \in [K]} \left\{ \widehat{\mu}_i(t) + Z_t \, \mathcal{C}_i(t) \right\}.$$

$Z_1, \ldots, Z_T$ are i.i.d. samples from the sampling distribution.

`RandUCB` Meta-algorithm

- **Generic OFU algorithm**: If $\widehat{\mu}_i(t)$ is the mean reward for arm i at round t, $\mathcal{C}_i(t)$ is the corresponding confidence set, pick the arm with the largest upper confidence bound.

$$i_t = \arg\max_{i \in [K]} \left\{ \widehat{\mu}_i(t) + \beta \, \mathcal{C}_i(t) \right\} .$$

  Here, β is deterministic and chosen to trade off exploration and exploitation optimally.

- `RandUCB`: Replace deterministic β by a random variable Z_t:

$$i_t = \arg\max_{i \in [K]} \left\{ \widehat{\mu}_i(t) + Z_t \, \mathcal{C}_i(t) \right\} .$$

  $Z_1, \ldots, Z_T$ are i.i.d. samples from the sampling distribution.

- **Uncoupled** `RandUCB`:

$$i_t = \arg\max_{i \in [K]} \left\{ \widehat{\mu}_i(t) + Z_{i,t} \, \mathcal{C}_i(t) \right\} .$$

- **General sampling distribution**: Discrete distribution on the interval $[L, U]$, supported on M equally-spaced points, $\alpha_1 = L, \dots, \alpha_M = U$. Define $p_m := \mathbf{P}(Z = \alpha_m)$.

- **General sampling distribution**: Discrete distribution on the interval $[L, U]$, supported on M equally-spaced points, $\alpha_1 = L, \ldots, \alpha_M = U$. Define $p_m := \mathbf{P}(Z = \alpha_m)$.
- **Default sampling distribution**: Gaussian distribution truncated in the $[0, U]$ interval with tunable hyper-parameters $\varepsilon, \sigma > 0$ such that $p_M = \varepsilon$ and

$$\text{For } 1 \leq m \leq M-1, \quad p_m \propto \exp(-\alpha_m^2/2\sigma^2).$$

- **General sampling distribution**: Discrete distribution on the interval $[L, U]$, supported on M equally-spaced points, $\alpha_1 = L, \ldots, \alpha_M = U$. Define $p_m := \mathbf{P}(Z = \alpha_m)$.
- **Default sampling distribution**: Gaussian distribution truncated in the $[0, U]$ interval with tunable hyper-parameters $\varepsilon, \sigma > 0$ such that $p_M = \varepsilon$ and

$$\text{For } 1 \leq m \leq M-1, \quad p_m \propto \exp(-\alpha_m^2/2\sigma^2).$$

- **Default choice across bandit problems**: Coupled `RandUCB` with $U = O(\beta)$, $M = 10$, $\varepsilon = 10^{-8}$, $\sigma = 0.25$.

RandUCB for multi-armed bandits

- Let $Y_i(t)$ be the sum of rewards obtained for arm i until round t and $s_i(t)$ be the number of pulls for arm i until round t.
  Mean $\widehat{\mu}_i(t) = Y_i(t)/s_i(t)$ and confidence interval $\mathcal{C}_i(t) = \sqrt{1/s_i(t)}$.

RandUCB for multi-armed bandits

- Let $Y_i(t)$ be the sum of rewards obtained for arm i until round t and $s_i(t)$ be the number of pulls for arm i until round t.
  Mean $\widehat{\mu}_i(t) = Y_i(t)/s_i(t)$ and confidence interval $\mathcal{C}_i(t) = \sqrt{1/s_i(t)}$.
- **OFU algorithm for MAB**: Pull each arm once, and for $t > K$, pull arm

$$i_t = \arg\max_i \left\{ \widehat{\mu}_i(t) + \beta \sqrt{\frac{1}{s_i(t)}} \right\}.$$

$\texttt{RandUCB}$ for multi-armed bandits

- Let $Y_i(t)$ be the sum of rewards obtained for arm i until round t and $s_i(t)$ be the number of pulls for arm i until round t.
  Mean $\widehat{\mu}_i(t) = Y_i(t)/s_i(t)$ and confidence interval $\mathcal{C}_i(t) = \sqrt{1/s_i(t)}$.
- **OFU algorithm for MAB**: Pull each arm once, and for $t > K$, pull arm

$$i_t = \arg\max_i \left\{ \widehat{\mu}_i(t) + \beta \sqrt{\frac{1}{s_i(t)}} \right\}.$$

- UCB1 [Auer, Cesa-Bianchi and Fischer 2002]: $\beta = \sqrt{2\ln(T)}$

`RandUCB` for multi-armed bandits

- Let $Y_i(t)$ be the sum of rewards obtained for arm i until round t and $s_i(t)$ be the number of pulls for arm i until round t.
  Mean $\widehat{\mu}_i(t) = Y_i(t)/s_i(t)$ and confidence interval $\mathcal{C}_i(t) = \sqrt{1/s_i(t)}$.
- **OFU algorithm for MAB**: Pull each arm once, and for $t > K$, pull arm

$$i_t = \arg\max_i \left\{ \widehat{\mu}_i(t) + \beta \sqrt{\frac{1}{s_i(t)}} \right\}.$$

- UCB1 [Auer, Cesa-Bianchi and Fischer 2002]: $\beta = \sqrt{2\ln(T)}$
- `RandUCB`: $L = 0, U = 2\sqrt{\ln(T)}$.
- We can also construct optimistic Thompson sampling and adaptive ε-greedy algorithms.

Regret of RandUCB for multi-armed bandits

Theorem 1 (Instance-dependent regret of uncoupled RandUCB for MAB)

If $\Delta_i = \mu_1 - \mu_i$ is the gap for arm i, and Z takes M different values $0 \leq \alpha_1 \leq \cdots \leq \alpha_M$ with probabilities $p_1, p_2, \ldots, p_M$, the regret $R(T)$ of uncoupled RandUCB *can be bounded as:*

$$O\left(\sum_{\Delta_i > 0} \Delta_i^{-1}\right) \times \left(\frac{M}{p_M} + Te^{-2\alpha_M^2} + \alpha_M^2\right).$$

Regret of `RandUCB` for multi-armed bandits

Theorem 1 (Instance-dependent regret of uncoupled `RandUCB` for MAB)

If $\Delta_i = \mu_1 - \mu_i$ is the gap for arm i, and Z takes M different values $0 \leq \alpha_1 \leq \cdots \leq \alpha_M$ with probabilities $p_1, p_2, \ldots, p_M$, the regret $R(T)$ of uncoupled `RandUCB` *can be bounded as:*

$$O\left(\sum_{\Delta_i > 0} \Delta_i^{-1}\right) \times \left(\frac{M}{p_M} + Te^{-2\alpha_M^2} + \alpha_M^2\right).$$

- Using $U = \alpha_M = 2\sqrt{\ln T}$ results in the problem-dependent $O\left(\ln T \times \left(\sum \Delta_i^{-1}\right)\right)$ regret.

Regret of RandUCB for multi-armed bandits

Theorem 1 (Instance-dependent regret of uncoupled RandUCB for MAB)

If $\Delta_i = \mu_1 - \mu_i$ is the gap for arm i, and Z takes M different values $0 \leq \alpha_1 \leq \cdots \leq \alpha_M$ with probabilities $p_1, p_2, \ldots, p_M$, the regret $R(T)$ of uncoupled RandUCB can be bounded as:

$$O\left(\sum_{\Delta_i > 0} \Delta_i^{-1}\right) \times \left(\frac{M}{p_M} + Te^{-2\alpha_M^2} + \alpha_M^2\right).$$

- Using $U = \alpha_M = 2\sqrt{\ln T}$ results in the problem-dependent $O\left(\ln T \times \left(\sum \Delta_i^{-1}\right)\right)$ regret.
- Standard reduction implies a problem-independent $\widetilde{O}(\sqrt{KT})$ regret matching that of UCB1 and Thompson sampling [Agrawal and Goyal, 2012].

Regret of `RandUCB` for multi-armed bandits

Theorem 1 (Instance-dependent regret of uncoupled `RandUCB` for MAB)

If $\Delta_i = \mu_1 - \mu_i$ is the gap for arm i, and Z takes M different values $0 \leq \alpha_1 \leq \cdots \leq \alpha_M$ with probabilities $p_1, p_2, \ldots, p_M$, the regret $R(T)$ of uncoupled `RandUCB` can be bounded as:

$$O\left(\sum_{\Delta_i > 0} \Delta_i^{-1}\right) \times \left(\frac{M}{p_M} + Te^{-2\alpha_M^2} + \alpha_M^2\right).$$

- Using $U = \alpha_M = 2\sqrt{\ln T}$ results in the problem-dependent $O\left(\ln T \times \left(\sum \Delta_i^{-1}\right)\right)$ regret.
- Standard reduction implies a problem-independent $\widetilde{O}(\sqrt{KT})$ regret matching that of UCB1 and Thompson sampling [Agrawal and Goyal, 2012].
- We also show the same problem-independent regret for the default coupled variant of `RandUCB`.

- Let $X_t = x_{i_t}$ and $M_t := \lambda I_d + \sum_{\ell=1}^{t-1} X_\ell X_\ell^\mathsf{T}$. $\widehat{\theta}_t := M_t^{-1} \sum_{\ell=1}^{t-1} Y_\ell X_\ell$. Mean $\widehat{\mu}_i(t) = \langle \widehat{\theta}_t, x_i \rangle$ and confidence width $\mathcal{C}_i(t) = \|x_i\|_{M_t^{-1}}$.

`RandUCB` for linear bandits

- Let $X_t = x_{i_t}$ and $M_t := \lambda I_d + \sum_{\ell=1}^{t-1} X_\ell X_\ell^\mathsf{T}$. $\widehat{\theta}_t := M_t^{-1} \sum_{\ell=1}^{t-1} Y_\ell X_\ell$. Mean $\widehat{\mu}_i(t) = \langle \widehat{\theta}_t, x_i \rangle$ and confidence width $\mathcal{C}_i(t) = \|x_i\|_{M_t^{-1}}$.
- **OFU algorithm for linear bandit**: Pull arm:

$$i_t = \underset{i \in [K]}{\arg\max} \left\{ \langle \widehat{\theta}_t, x_i \rangle + \beta \|x_i\|_{M_t^{-1}} \right\}.$$

`RandUCB` for linear bandits

- Let $X_t = x_{i_t}$ and $M_t := \lambda I_d + \sum_{\ell=1}^{t-1} X_\ell X_\ell^\mathsf{T}$. $\widehat{\theta}_t := M_t^{-1} \sum_{\ell=1}^{t-1} Y_\ell X_\ell$. Mean $\widehat{\mu}_i(t) = \langle \widehat{\theta}_t, x_i \rangle$ and confidence width $\mathcal{C}_i(t) = \|x_i\|_{M_t^{-1}}$.
- **OFU algorithm for linear bandit**: Pull arm:

$$i_t = \arg\max_{i \in [K]} \left\{ \langle \widehat{\theta}_t, x_i \rangle + \beta \|x_i\|_{M_t^{-1}} \right\}.$$

- OFU [Abbasi-Yadkori, Pál and Szepesvári 2011]: $\beta = \sqrt{\lambda} + \frac{1}{2}\sqrt{\ln(T^2 \lambda^{-d} \det(M_t))}$.

RandUCB for linear bandits

- Let $X_t = x_{i_t}$ and $M_t := \lambda I_d + \sum_{\ell=1}^{t-1} X_\ell X_\ell^\mathsf{T}$. $\widehat{\theta}_t := M_t^{-1} \sum_{\ell=1}^{t-1} Y_\ell X_\ell$. Mean $\widehat{\mu}_i(t) = \langle \widehat{\theta}_t, x_i \rangle$ and confidence width $\mathcal{C}_i(t) = \|x_i\|_{M_t^{-1}}$.
- **OFU algorithm for linear bandit**: Pull arm:

$$i_t = \arg\max_{i \in [K]} \left\{ \langle \widehat{\theta}_t, x_i \rangle + \beta \|x_i\|_{M_t^{-1}} \right\}.$$

- OFU [Abbasi-Yadkori, Pál and Szepesvári 2011]: $\beta = \sqrt{\lambda} + \frac{1}{2}\sqrt{\ln(T^2 \lambda^{-d} \det(M_t))}$.
- RandUCB: $L = 0,\ U = 3\left[\sqrt{\lambda} + \frac{1}{2}\sqrt{d \ln\left(T + T^2/d\lambda\right)}\right]$.

Regret of RandUCB for linear bandits

Theorem 2

Let $c_1 = \sqrt{\lambda} + \frac{1}{2}\sqrt{d \ln\left(T + T^2/d\lambda\right)}$ *and* $c_3 := 2d \ln\left(1 + \frac{T}{d\lambda}\right)$*. For any* $c_2 > c_1$*, the regret of* RandUCB *for linear bandits is bounded by*

$$(c_1 + c_2)\left(1 + \frac{2}{\mathbf{P}\left(Z > c_1\right) - \mathbf{P}\left(|Z| > c_2\right)}\right) \times \sqrt{c_3 T} + T\,\mathbf{P}\left(|Z| > c_2\right) + 1.$$

Regret of `RandUCB` for linear bandits

Theorem 2

Let $c_1 = \sqrt{\lambda} + \frac{1}{2}\sqrt{d \ln\left(T + T^2/d\lambda\right)}$ *and* $c_3 := 2d \ln\left(1 + \frac{T}{d\lambda}\right)$. *For any* $c_2 > c_1$, *the regret of* `RandUCB` *for linear bandits is bounded by*

$$(c_1 + c_2)\left(1 + \frac{2}{\mathbf{P}\left(Z > c_1\right) - \mathbf{P}\left(|Z| > c_2\right)}\right) \times \sqrt{c_3 T} + T\,\mathbf{P}\left(|Z| > c_2\right) + 1.$$

- Setting $U = 3c_1 < c_2$ ensures $\mathbf{P}\left(Z > c_1\right)$ is a positive constant and $\mathbf{P}\left(|Z| > c_2\right) = 0$, resulting in $\widetilde{O}(d\sqrt{T})$ regret bound.

Regret of `RandUCB` for linear bandits

Theorem 2

Let $c_1 = \sqrt{\lambda} + \frac{1}{2}\sqrt{d \ln\left(T + T^2/d\lambda\right)}$ *and* $c_3 := 2d \ln\left(1 + \frac{T}{d\lambda}\right)$*. For any* $c_2 > c_1$*, the regret of* `RandUCB` *for linear bandits is bounded by*

$$(c_1 + c_2)\left(1 + \frac{2}{\mathbf{P}\left(Z > c_1\right) - \mathbf{P}\left(|Z| > c_2\right)}\right) \times \sqrt{c_3 T} + T\,\mathbf{P}\left(|Z| > c_2\right) + 1.$$

- Setting $U = 3c_1 < c_2$ ensures $\mathbf{P}\left(Z > c_1\right)$ is a positive constant and $\mathbf{P}\left(|Z| > c_2\right) = 0$, resulting in $\widetilde{O}(d\sqrt{T})$ regret bound.
- Regret bound does not depend on K and holds for infinite arms.

Regret of `RandUCB` for linear bandits

Theorem 2

Let $c_1 = \sqrt{\lambda} + \frac{1}{2}\sqrt{d \ln\left(T + T^2/d\lambda\right)}$ *and* $c_3 := 2d \ln\left(1 + \frac{T}{d\lambda}\right)$*. For any* $c_2 > c_1$*, the regret of* `RandUCB` *for linear bandits is bounded by*

$$(c_1 + c_2)\left(1 + \frac{2}{\mathbf{P}\left(Z > c_1\right) - \mathbf{P}\left(|Z| > c_2\right)}\right) \times \sqrt{c_3 T} + T\,\mathbf{P}\left(|Z| > c_2\right) + 1.$$

- Setting $U = 3c_1 < c_2$ ensures $\mathbf{P}\left(Z > c_1\right)$ is a positive constant and $\mathbf{P}\left(|Z| > c_2\right) = 0$, resulting in $\widetilde{O}(d\sqrt{T})$ regret bound.
- Regret bound does not depend on K and holds for infinite arms.
- Matches the bound of OFU in [Abbasi-Yadkori et al., 2011] and is better than the $O(d^{3/2}\sqrt{T})$ bound for TS [Agrawal and Goyal, 2013].

Regret of `RandUCB` for linear bandits

Theorem 2

Let $c_1 = \sqrt{\lambda} + \frac{1}{2}\sqrt{d \ln\left(T + T^2/d\lambda\right)}$ *and* $c_3 := 2d \ln\left(1 + \frac{T}{d\lambda}\right)$. *For any* $c_2 > c_1$*, the regret of* `RandUCB` *for linear bandits is bounded by*

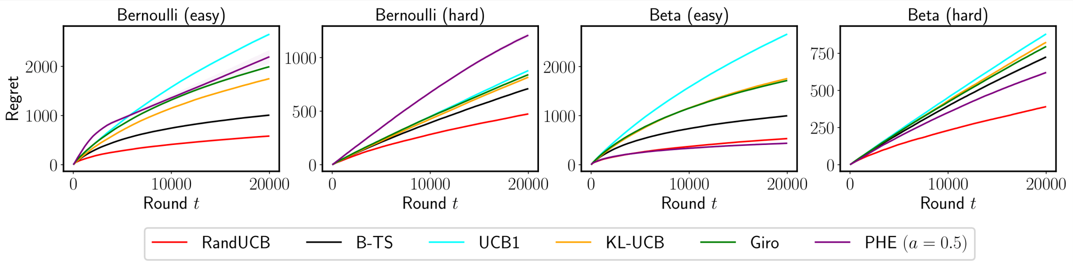
$$(c_1 + c_2)\left(1 + \frac{2}{\mathbf{P}\left(Z > c_1\right) - \mathbf{P}\left(|Z| > c_2\right)}\right) \times \sqrt{c_3 T} + T\,\mathbf{P}\left(|Z| > c_2\right) + 1.$$

- Setting $U = 3c_1 < c_2$ ensures $\mathbf{P}\left(Z > c_1\right)$ is a positive constant and $\mathbf{P}\left(|Z| > c_2\right) = 0$, resulting in $\widetilde{O}(d\sqrt{T})$ regret bound.
- Regret bound does not depend on K and holds for infinite arms.
- Matches the bound of OFU in [Abbasi-Yadkori et al., 2011] and is better than the $O(d^{3/2}\sqrt{T})$ bound for TS [Agrawal and Goyal, 2013].
- We prove a similar $\widetilde{O}(d\sqrt{T})$ bound for generalized linear bandits.

The `RandUCB` meta-algorithm

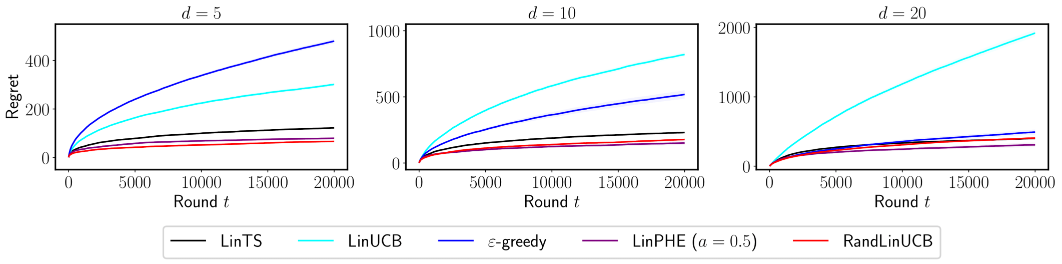
Empirical study

Experiments - multi-armed bandit



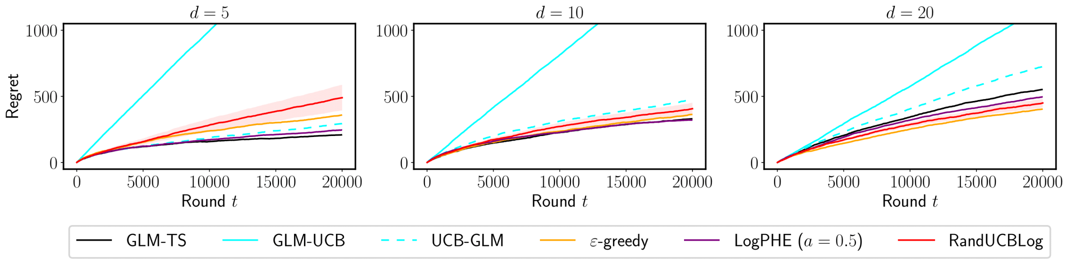
- B-TS: Thompson Sampling with a beta posterior
- KL-UCB [Garivier and Cappé, 2011]: UCB with tighter confidence intervals.
- Randomized exploration baselines: Giro [Kveton et al., 2019c], PHE [Kveton et al., 2019b]

Experiments - linear bandit



- Lin-TS: Thompson Sampling with a Gaussian posterior
- ε-greedy [Langford and Zhang, 2008]
- Randomized exploration baseline: LinPHE [Kveton et al., 2019a]

Experiments - logistic bandit



- GLM-TS [Kveton et al., 2019d]: TS with a Laplace approximation to the posterior.
- GLM-UCB [Filippi et al., 2010] and UCB-GLM [Li et al., 2017]
- ε-greedy [Langford and Zhang, 2008]
- Randomized exploration baseline: LogPHE [Kveton et al., 2019d]

Proposed RandUCB, a generic meta-algorithm achieving the theoretical performance of UCB and the practical performance of Thompson sampling.

Paper: `https://arxiv.org/abs/1910.04928`
Code: `https://github.com/vaswanis/randucb`

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *NIPS*, 2011.

Shipra Agrawal and Navin Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *COLT*, 2012.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *ICML*, 2013.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *NIPS*, 2010.

Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *COLT*, 2011.

Branislav Kveton, Csaba Szepesvári, Mohammad Ghavamzadeh, and Craig Boutilier. Perturbed-history exploration in stochastic linear bandits. *UAI*, 2019a.

Branislav Kveton, Csaba Szepesvári, Mohammad Ghavamzadeh, and Craig Boutilier. Perturbed-history exploration in stochastic multi-armed bandits. In *IJCAI-19*, 2019b.

Branislav Kveton, Csaba Szepesvári, Sharan Vaswani, Zheng Wen, Tor Lattimore, and Mohammad Ghavamzadeh. Garbage in, reward out: Bootstrapping exploration in multi-armed bandits. In *ICML*, 2019c.

Branislav Kveton, Manzil Zaheer, Csaba Szepesvári, Lihong Li, Mohammad Ghavamzadeh, and Craig Boutilier. Randomized exploration in generalized linear bandits. *arXiv:1906.08947*, 2019d.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *NIPS*, 2008.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *ICML*, 2017.