Advanced Multi-Armed Bandit Algorithms

James A. Grant

February 19, 2025
Lancaster AI

1. Recap on MAB

Multi-armed Bandits (MABs) are a simple* family of models in reinforcement learning.

Simplest case ('Stochastic K-armed bandit'):

- Actions: $k \in \{1, \ldots, K\} := [K]$, time steps: $t \in \{1, 2, \ldots\}$.
- Each action k associated with distribution ν_k.
- Learner chooses an action $a_t \in [K]$ at each round t.
- Learner observes a reward $X_{a_t,t} \sim \nu_{a_t}$.

1. Recap on MAB

Multi-armed Bandits (MABs) are a simple* family of models in reinforcement learning.

Simplest case ('Stochastic K-armed bandit'):

- Actions: $k \in \{1, \dots, K\} := [K]$, time steps: $t \in \{1, 2, \dots\}$.
- Each action k associated with distribution ν_k.
- Learner chooses an action $a_t \in [K]$ at each round t.
- Learner observes a reward $X_{a_t,t} \sim \nu_{a_t}$.

There are a number of plausible objectives within this framework (Lattimore and Szepesvári, 2020).

$$\text{Discounted Reward Maximisation:} \quad \max_{a_1, a_2, \dots} \mathbb{E}\left(\sum_{t=1}^{\infty} \gamma^t X_{a_t,t}\right)$$

where $\gamma \in (0, 1)$ (e.g. Gittins, 1979; Gittins et al., 2011).

1. Recap on MAB

Multi-armed Bandits (MABs) are a simple* family of models in reinforcement learning.

Simplest case ('Stochastic K-armed bandit'):

- Actions: $k \in \{1, \dots, K\} := [K]$, time steps: $t \in \{1, 2, \dots\}$.
- Each action k associated with distribution ν_k.
- Learner chooses an action $a_t \in [K]$ at each round t.
- Learner observes a reward $X_{a_t,t} \sim \nu_{a_t}$.

There are a number of plausible objectives within this framework (Lattimore and Szepesvári, 2020).

$$\text{Best Arm Identification:} \quad \max_{a_1, a_2, \dots a_T} \mathbb{P}\left(\max_{k \in [K]} \frac{\sum_{t=1}^{T} X_{k,t} \mathbb{I}\{a_t = k\}}{\sum_{t=1}^{T} \mathbb{I}\{a_t = k\}} = \max_{k \in [K]} \mathbb{E}(X_{k,t}) \right)$$

for some budget $T \in \mathbb{N}$ (e.g. Bubeck et al., 2009; Audibert and Bubeck, 2010).

1. Recap on MAB

Multi-armed Bandits (MABs) are a simple* family of models in reinforcement learning.

Simplest case ('Stochastic K-armed bandit'):

- Actions: $k \in \{1, \ldots, K\} := [K]$, time steps: $t \in \{1, 2, \ldots\}$.
- Each action k associated with distribution ν_k.
- Learner chooses an action $a_t \in [K]$ at each round t.
- Learner observes a reward $X_{a_t,t} \sim \nu_{a_t}$.

There are a number of plausible objectives within this framework (Lattimore and Szepesvári, 2020).

$$\text{(Pseudo)regret Minimisation:} \quad \min_{a_1, a_2, \ldots a_T} \sum_{t=1}^{T} \max_{k \in [K]} \mathbb{E}_\nu(X_{k,t}) - \mathbb{E}_\nu(X_{a_t,k})$$

for some budget $T \in \mathbb{N}$ (e.g. Lai and Robbins, 1985; Auer et al., 2002) - **Today's Focus**

2. Regret Minimisation in Stochastic K-armed bandit

The regret,

$$\sum_{t=1}^{T} \max_{k\in[K]} \mathbb{E}_\nu(X_{k,t}) - \mathbb{E}_\nu(X_{a_t,k}) := \sum_{t=1}^{T} \mu^* - \mu_{a_t},$$

is inevitably increasing - an *oracle* would achieve zero regret, any other learner more due to uncertainty.

2. Regret Minimisation in Stochastic K-armed bandit

The regret,

$$\sum_{t=1}^{T} \max_{k\in[K]} \mathbb{E}_\nu(X_{k,t}) - \mathbb{E}_\nu(X_{a_t,k}) := \sum_{t=1}^{T} \mu^* - \mu_{a_t},$$

is inevitably increasing - an *oracle* would achieve zero regret, any other learner more due to uncertainty.

We seek *policies* whose regret is of an optimal order for large families of $\{\nu_1, ...\nu_K\}$.
A policy π maps from previously observed data to the action set $[K]$.

Optimality is measured with respect to lower bounds on the best possible regret.

2. Regret Minimisation in Stochastic K-armed bandit

We have two main families of lower bound: instance dependent and minimax.

- Instance Dependent (Lai and Robbins, 1985; Burnetas and Katehakis, 1996)

$$\lim_{T\to\infty} \frac{Reg(T)}{\log(T)} \leq \sum_{k\neq k^*} \frac{\mu^* - \mu_k}{\inf_{\nu'}\{D_{KL}(\nu_k \,||\, \nu') : \mathbb{E}'_{\nu}(X) > \mu^*\}}$$

- Minimax (e.g. Bubeck et al., 2013) (see also (LeCam, 1973))

$$Reg(T) = \Omega(\sqrt{KT})$$

2. Regret Minimisation in Stochastic K-armed Bandit

Analysis of the expected pseudo-regret often focusses on the construction of a high-probability good event:

- e.g. (informally) estimates of the mean rewards of each action remain within certain regions around the true parameter across all rounds.
- Outside the good event: potential for linear regret
- Inside the good event: only actions of a reasonable quality are played often
- As t increases, the conditions of the good event become stricter, meaning the regret per round decreases.
- An algorithm that achieves the good event with high-probability will do so by ensuring a balance of exploration and exploitation.

2. Regret Minimisation in Stochastic K-armed Bandit

Two principles have been especially popular: optimism and randomisation. Both

1. are data-driven ('adaptive'),
2. encourage exploration proportional to uncertainty,
3. converge to greedy decision making eventually.

2. Regret Minimisation in Stochastic K-armed Bandit

Two principles have been especially popular: optimism and randomisation. Both

1. are data-driven ('adaptive'),
2. encourage exploration proportional to uncertainty,
3. converge to greedy decision making eventually.

Example: Optimism for rewards in $[0, 1]$ - UCB1 (Auer et al., 2002)

- Choose each action once to initialise mean estimates $\hat{\mu}_k$
- In round $t = K + 1, K + 2, \ldots$ choose

$$a_t = \arg\max_{k \in [K]} \left[\hat{\mu}_k + \sqrt{\frac{2\log(t)}{N_k(t)}} \right]$$

where $N_k(t)$ is number of times played action k.

2. Regret Minimisation in Stochastic K-armed Bandit

Example: Optimism for rewards in $[0,1]$ - UCB1 (Auer et al., 2002)

- Choose each action once to initialise mean estimates $\hat{\mu}_k$
- In round $t = K+1, K+2, \ldots$ choose

$$a_t = \arg\max_{k \in [K]} \left[\hat{\mu}_k + \sqrt{\frac{2\log(t)}{N_k(t)}} \right]$$

where $N_k(t)$ is number of times played action k.

The regret of UCB1 is known to satisfy

$$Reg(T) \leq 8 \sum_{k \neq k^*} \frac{\log(T)}{\mu^* - \mu_k} + C$$

which is order-optimal, but not coefficient-wise.

Example: Randomisation for Bernoulli Rewards - Thompson Sampling (Thompson, 1933)

- Initialise with priors $p_{k,0}$ for each action's distribution.
- In round $t = 1, 2, ...$ draw a sample $\tilde{\mu}_{k,t}$ from current posterior belief $p_{k,t}$ for each action.
- Choose $a_t \in \arg\max_{k \in [K]} \tilde{\mu}_{k,t}$

Example: Randomisation for Bernoulli Rewards - Thompson Sampling (Thompson, 1933)

- Initialise with priors $p_{k,0}$ for each action's distribution.
- In round $t = 1, 2, \ldots$ draw a sample $\tilde{\mu}_{k,t}$ from current posterior belief $p_{k,t}$ for each action.
- Choose $a_t \in \arg\max_{k \in [K]} \tilde{\mu}_{k,t}$

Thompson Sampling is known to asymptotically achieve optimal instance-wise regret (Agrawal and Goyal, 2012; Kaufmann et al., 2012b)

3. Sharper Confidence Bounds

Since the confidence bounds drive the exploration, any slackness therein directly leads to increased regret.

UCB1 utilises Hoeffding's Inequality - which is generally useful for bounded observations, but not tight when additional assumptions hold, e.g. Bernoulli data.

KL-UCB (Garivier and Cappé, 2011; Maillard et al., 2011; Cappé et al., 2013) is based around sharper confidence sets (based on Chernoff bounds) for parametric bandits.

Bayes-UCB (Kaufmann et al., 2012a) uses quantiles of the posterior distribution in place of frequentist upper confidence limits.

3. Sharper Confidence Bounds

KL-UCB:

- Choose each arm once to initialise mean estimates $\hat{\mu}_k$
- In round $t = K + 1, K + 2, \ldots$ choose

$$a_t = \arg\max_{k \in [K]} \left[\max \left\{ \mu \in [0, 1] : D_{KL}(\hat{\mu}_{k,t} \mid\mid \mu) \leq \frac{\log(1 + t \log^2(t))}{N_k(t)} \right\} \right]$$

3. Sharper Confidence Bounds

KL-UCB:

- Choose each arm once to initialise mean estimates $\hat{\mu}_k$
- In round $t = K+1, K+2, \ldots$ choose

$$a_t = \arg\max_{k \in [K]} \left[\max \left\{ \mu \in [0,1] : D_{KL}(\hat{\mu}_{k,t} \mid\mid \mu) \leq \frac{\log(1 + t\log^2(t))}{N_k(t)} \right\} \right]$$

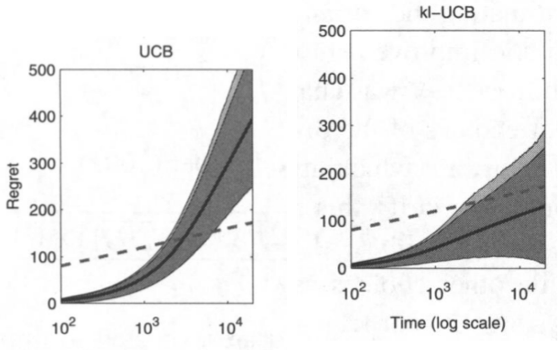
These tighter confidence sets yield an improved (over UCB1) regret bound (in terms of the coefficients):

$$Reg(T, \pi^{KL-UCB}) \leq \sum_{k \neq k^*} (\mu^* - \mu_k) \inf_{\epsilon_1, \epsilon_2} \left(\frac{\log(1 + t\log^2(t))}{D_{KL}(\mu_k + \epsilon_1 \mid\mid \mu^* - \epsilon_2)} + C(\epsilon_1, \epsilon_2) \right)$$

which is asymptotically optimal ($T \to \infty$).

3. Sharper Confidence Bounds: Comparison

This also realises an improved empirical performance. In a ten-armed Bernoulli bandit, we compare regret of UCB1 and KL-UCB on the log-scale. Figures taken from Cappé et al. (2013).



A note of caution: optimised algorithms such as KL-UCB can degrade in quality rapidly outside their assumptions (Fan and Glynn, 2024).

3. Sharper Confidence Bounds

Bayes-UCB:

- Initialise with priors $p_{0,k}$ on the parameters of each ν_k
- In round $t = 1, 2, ..., T$ choose

$$a_t = \arg\max_{k \in [K]} \left[Q_{p_{t-1,k}} \left(1 - (t \log(T))^c\right) \right]$$

  where $Q_p(t)$ is the t quantile of distribution p.
- Observe $X_{a_t,t}$ and update posteriors.

This also achieves an asymptotically optimal regret (Kaufmann et al., 2012a). So optimal policies are not unique.

4. Contextual Bandits

In the vanilla K-armed bandit, reward distributions ν_k are static. Without going to a fully general RL model, we can relax this assumption. Contextual bandits:

- Actions $k \in [K]$, time steps $t \in \{1, 2, \ldots\}$, contexts $\mathbf{x}_t \in \mathcal{X} \subset \mathbb{R}^d$.
- Each action k is associated with a distribution $\nu_k(\mathbf{x})$.
- Learner observes context $\mathbf{x}_t$ and chooses an action $a_t \in [K]$ in each round t.
- Learner observes a reward $X_{a_t,t} \sim \nu_{a_t}(\mathbf{x}_t)$, with expectation $\mu_{a_t}(\mathbf{x}_t) = \mathbb{E}(X_{a_t,t})$.

The same regret minimisation objective may be considered

$$\min_{a_1,a_2,\ldots,a_T} \sum_{t=1}^{T} \max_{k\in[K]} \mu_k(\mathbf{x}_t) - \mu_{a_t}(\mathbf{x}_t).$$

4. Contextual Bandits

The same regret minimisation objective may be considered

$$\min_{a_1, a_2, \dots, a_T} \sum_{t=1}^{T} \max_{k \in [K]} \mu_k(\mathbf{x}_t) - \mu_{a_t}(\mathbf{x}_t).$$

Particularly well studied in the parametric setting, with generalised linear models, e.g.

- Linear bandit: $X_{k,t} \sim N(\mathbf{x}_t^T \boldsymbol{\theta}_k, \sigma^2)$ (Auer, 2002; Li et al., 2010)
- Logistic bandit: $X_{k,t} \sim \text{Bern}\left((1 + \exp(-\mathbf{x}_t^T \boldsymbol{\theta}_k))^{-1}\right)$ (Filippi et al., 2010; Faury et al., 2020)

where $\boldsymbol{\theta}_k \in \Theta \subset \mathbb{R}^d$ are unknown, action specific parameters.

4. Contextual Bandits: Optimism

Regret minimisation again requires learning parameters of unknown distributions sufficiently well - this time $\boldsymbol{\theta}_k,\ k \in [K]$.

In linear bandits, an oracle policy would select $a_t \in \arg\max_{k \in [K]} \mathbf{x}_t^T \boldsymbol{\theta}_k$ in each round.

An optimistic approach can again work well, here for each action we compute:

$$UCB_{k,t} = \max_{\boldsymbol{\theta} \in \Theta_{k,t}} \mathbf{x}_t^T \boldsymbol{\theta},$$

where $\Theta_{k,t}$ is a high-probability confidence region for $\boldsymbol{\theta}_k$.

Much attention has focussed on deriving confidence sets which are tight and can be computed efficiently.

4. Contextual Bandits: Optimism

For instance, in logistic bandits, GLM-UCB (Filippi et al., 2010) forms indices based on a regularised parameter estimate $\hat{\boldsymbol{\theta}}_k$:

$$UCB_{k,t} = \sigma(\mathbf{x}_t^T \hat{\boldsymbol{\theta}}_k) + \rho(t)||\mathbf{x}_t||_{\Sigma_{k,t}{}^{-1}}$$

where we let σ denote the logistic link function, $\rho(t)$ controls the amount of exploration, and $\Sigma_{k,t}$ is a design matrix specific to action k.

4. Contextual Bandits: Optimism

For instance, in logistic bandits, GLM-UCB (Filippi et al., 2010) forms indices based on a regularised parameter estimate $\hat{\boldsymbol{\theta}}_k$:

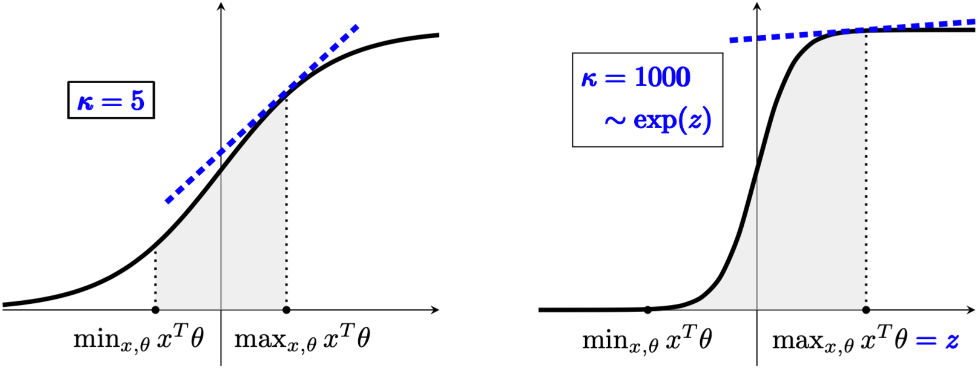
$$UCB_{k,t} = \sigma(\mathbf{x}_t^T \hat{\boldsymbol{\theta}}_k) + \rho(t)||\mathbf{x}_t||_{\Sigma_{k,t}^{-1}}$$

where we let σ denote the logistic link function, $\rho(t)$ controls the amount of exploration, and $\Sigma_{k,t}$ is a design matrix specific to action k.

This algorithm has a near optimal $O\left(\sqrt{T \log^{3/2}(T)}\right)$ bound on its minimax, however it has a linear dependence on a potentially large problem-specific parameter.

4. Contextual Bandits: Optimism

GLM-UCB has a regret which is linear in $\kappa = \sup_{\mathbf{x}\in\mathcal{X},\theta\in\Theta} 1/\sigma'(\mathbf{x}^T\boldsymbol{\theta})$.
Unfortunately, this can become quite large in certain problems. Figure from Faury et al. (2020).



4. Contextual Bandits: Randomisation

Thompson Sampling is also applicable, but potentially challenging, in the logistic bandit setting (Russo et al., 2018; Dong et al., 2019).

For each arm we have a prior/posterior over $\boldsymbol{\theta}_k$, and the policy should in round $t = 1, 2...$

- Draw a sample $\tilde{\boldsymbol{\theta}}_{k,t}$ from the posterior for each action
- Choose an action $a_t \in \arg\max_{k \in [K]} \sigma(\mathbf{x}_t^T \tilde{\theta}_{k,t})$.

4. Contextual Bandits: Randomisation

Thompson Sampling is also applicable, but potentially challenging, in the logistic bandit setting (Russo et al., 2018; Dong et al., 2019).

For each arm we have a prior/posterior over $\boldsymbol{\theta}_k$, and the policy should in round $t = 1, 2...$

- Draw a sample $\tilde{\boldsymbol{\theta}}_{k,t}$ from the posterior for each action
- Choose an action $a_t \in \arg\max_{k\in[K]} \sigma(\mathbf{x}_t^T \tilde{\theta}_{k,t})$.

Practically this is challenging, due to the intractability of the posterior in logistic models. Potential solutions: Laplace approximation (Russo et al., 2018), or a Gibbs sampler built on Polya-Gamma approximation (Dumitrascu et al., 2018).

Dependence on κ is also an issue and the subject of ongoing research (Gouverneur et al., 2024; Neu et al., 2022).

5. Further Variants and Extensions

The optimism and randomisation principles, and accompanying regret analysis have been extended substantially beyond K-armed and generalised linear bandits.

- Continuum-armed/Lipschitz Bandits
- Combinatorial Bandits
- Non-stationary Bandits
- Partial Monitoring
- Federated Bandits

Lattimore and Szepesvári (2020) is a great resource for more detail on foundational and theoretical aspects.

References I

Agrawal, S. and Goyal, N. (2012). Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings.

Audibert, J.-Y. and Bubeck, S. (2010). Best arm identification in multi-armed bandits. In *COLT-23th Conference on learning theory-2010*, pages 13–p.

Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*.

Bubeck, S., Munos, R., and Stoltz, G. (2009). Pure exploration in multi-armed bandits problems. In *Algorithmic Learning Theory: 20th International Conference, ALT 2009, Porto, Portugal, October 3-5, 2009. Proceedings 20*, pages 23–37. Springer.

Bubeck, S., Perchet, V., and Rigollet, P. (2013). Bounded regret in stochastic multi-armed bandits. In *Conference on Learning Theory*, pages 122–134. PMLR.

Burnetas, A. N. and Katehakis, M. N. (1996). Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142.

References II

Cappé, O., Garivier, A., Maillard, O.-A., Munos, R., and Stoltz, G. (2013). Kullback-leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, pages 1516–1541.

Dong, S., Ma, T., and Van Roy, B. (2019). On the performance of thompson sampling on logistic bandits. In *Conference on Learning Theory*, pages 1158–1160. PMLR.

Dumitrascu, B., Feng, K., and Engelhardt, B. (2018). Pg-ts: Improved thompson sampling for logistic contextual bandits. *Advances in neural information processing systems*, 31.

Fan, L. and Glynn, P. W. (2024). The fragility of optimized bandit algorithms. *Operations Research*.

Faury, L., Abeille, M., Calauzènes, C., and Fercoq, O. (2020). Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pages 3052–3060. PMLR.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. (2010). Parametric bandits: The generalized linear case. *Advances in neural information processing systems*, 23.

Garivier, A. and Cappé, O. (2011). The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376. JMLR Workshop and Conference Proceedings.

Gittins, J., Glazebrook, K., and Weber, R. (2011). *Multi-armed bandit allocation indices*. John Wiley & Sons.

References III

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 41(2):148–164.

Gouverneur, A., Rodríguez-Gálvez, B., Oechtering, T. J., and Skoglund, M. (2024). An information-theoretic analysis of thompson sampling for logistic bandits. *arXiv preprint arXiv:2412.02861*.

Kaufmann, E., Cappé, O., and Garivier, A. (2012a). On bayesian upper confidence bounds for bandit problems. In *Artificial intelligence and statistics*, pages 592–600. PMLR.

Kaufmann, E., Korda, N., and Munos, R. (2012b). Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pages 199–213. Springer.

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

LeCam, L. (1973). Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pages 38–53.

References IV

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670.

Maillard, O.-A., Munos, R., and Stoltz, G. (2011). A finite-time analysis of multi-armed bandits problems with kullback-leibler divergences. In *Proceedings of the 24th annual Conference On Learning Theory*, pages 497–514. JMLR Workshop and Conference Proceedings.

Neu, G., Olkhovskaia, I., Papini, M., and Schwartz, L. (2022). Lifting the information ratio: An information-theoretic analysis of thompson sampling for contextual bandits. *Advances in Neural Information Processing Systems*, 35:9486–9498.

Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., Wen, Z., et al. (2018). A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294.