A Note on Linear Bandits and Confidence Sets

Yisong Yue

April 12, 2016

1 Basics

The basic stochastic multi-armed bandit problem [2] operates over a set of K arms, each with mean reward μ_a for $a \in \{1, \ldots, K\}$. The setting proceeds as a sequential learning and decision making problem as follows:

1. Init $t = 1$
2. Algorithm chooses an arm $a_t \in \{1, \ldots K\}$
3. World provides stochastic reward $\hat{r}_t$, with mean $E[\hat{r}_t] = \mu_{a_t}$ and indepedent noise
4. $t = t + 1$, repeat from Step 2 until $t = T$ (T possibly unknown).

For the UCB1 algorithm [2], the choice at each time t is:

$$a_t = \text{argmax}_{a \in \{1,\ldots,K\}} \hat{\mu}_{a,t} + \hat{c}_{a,t}, \tag{1}$$

where $\hat{\mu}_{a,t}$ is the empirical mean estimate of μ_a at time t given previous observations, and $\hat{c}_{a,t} \equiv c_t\sqrt{1/t}$ is an upper confidence bound (with c_t controlling the confidence level). One interpretation is that this choice is a good trade-off between exploitation ($\hat{\mu}_{a,t}$) and exploration ($\hat{c}_{a,t}$).

From this setup, we also know with some confidence level that, for each action $a \in \{1, \ldots, K\}$, the true mean reward is contained in:

$$\mu_a \in C_{a,t} \equiv [\hat{\mu}_{a,t} - \hat{c}_{a,t}, \hat{\mu}_{a,t} + \hat{c}_{a,t}], \tag{2}$$

since that is the definition of a confidence bound. One can thus rewrite (1) as:

$$a_t = \text{argmax}_{a \in \{1,\ldots,K\}} \max_{u \in C_t} u, \tag{3}$$

The interpretation here is that we will be as optimistic as possible about what each μ_a is, and then select the arm with the best optimistic estimate. Because of the equivalence between (1) and (3), either can be used as the starting point for extending to more complicated settings, depending on whichever is more convenient.

1.1 Regret Analysis

For the basic stochastic MAB setting, we measure regret via:

$$R_T = \sum_{t=1}^{T} \left[\mu_* - \mu_{a_t}\right], \tag{4}$$

where μ_* denotes the mean reward of the best arm. If we run the UCB1 algorithm, we can show that, with high probability:

$$\begin{aligned} R_T &= \sum_{t=1}^{T} \left[\mu_* - \mu_{a_t}\right] && (5) \\ &\leq \sum_{t=1}^{T} \left[\hat{\mu}_{a_t,t} + \hat{c}_{a_t,t} - \mu_{a_t}\right] && (6) \\ &\leq \sum_{t=1}^{T} \left[\hat{\mu}_{a_t,t} + \hat{c}_{a_t,t} - (\hat{\mu}_{a_t,t} - \hat{c}_{a_t,t})\right] && (7) \\ &= 2\sum_{t=1}^{T} \hat{c}_{a_t,t} && (8) \end{aligned}$$

(6) follows from the fact that a_t is chosen to have the highest upper confidence bound of any arm, so $\hat{\mu}_{a_t,t} + \hat{c}_{a_t,t} \geq \mu_*$ with high probability. (7) follows from the same fact applied to $\hat{\mu}_{a_t,t} + \hat{c}_{a_t,t} \geq \mu_{a_t}$ also holding with high probability. Thus we can see that the regret of the UCB1 algorithm is upper bounded by the sum of diameters of the confidence regions. If the confidence regions shrink fast enough, then the UCB1 algorithm will have sub-linear regret with high probability.[1] For the UCB1 algorithm with bounded rewards in $[0, 1]$, it was shown that for $c_{a,t} = \sqrt{2 \log t / n_{a,t}}$, where $n_{a,t}$ denotes the number of times arm a has been played up to time t, the regret scales as $\mathcal{O}(\frac{K}{\epsilon} \log T)$, where $\epsilon = \mu_* - \max_{k \neq a_*} \mu_k$ is the minimum distinguishability between the best arm and the rest. This same analysis will hold for more complicated settings as well, so long as one can define suitable notions of confidence regions.

Note that one can always use a looser confidence bound. For instance, twice the confidence region is still a high probability confidence region. But then the regret would also be twice as much. So the trick is to use confidence bounds that are as tight as possible – this has practical implications as well, because over-exploring by a constant factor can be very bad in practice. Oftentimes, in practice one uses really tight confidence regions that cannot be proven to actually be high probability confidence regions [3].

[1] Note that one would need a confidence bound that holds over all time steps. A naive approach is to use the union bound.

2 Linear Contextual Bandits

For clarity, we focus here on the simple disjoint linear contextual bandits setting from [3]. At each iteration t we are given K contexts $x_{t,a} \in \Re^D$ in the form of D-dimensional feature vectors, and that the expected reward of each action a is $\mu_{a,t} \equiv w_a^T x_{t,a}$, for unknown w_a. The setup is then:

1. Init $t = 1$
2. Receive contexts $x_{t,a} \in \Re^D$ for each $a \in \{1, \ldots, K\}$
3. Algorithm chooses $a_t \in \{1, \ldots, K\}$
4. Receive reward $\hat{r}_t$ with expected reward $E[\hat{r}_t] = \mu_{a_t,t} \equiv w_{a_t}^T x_{t,a_t}$, and independent noise.
5. $t = t + 1$, repeat from Step 2 until $t = T$ (T possibly unknown)

Regret. The regret definition is then modified to be:

$$R_T = \sum_{t=1}^{T} \left[w_{a_t^*}^T x_{t,*} - w_{a_t}^T x_{t,a_t} \right], \tag{9}$$

where $a_t^* = \text{argmax}_a\, w_a^T x_{t,a}$ is the best action at time t given features $x_{t,a}$, and $x_{t,*} = x_{t,a_t^*}$ is the feature representation of the best action at time t. Note that the best action now varies depending on the features/contexts provided at each time step. For example, each action can be a news article, and features $x_{t,a}$ can describe characteristics such as the age, gender, and location of a user [3]. This is still a stochastic setting because model doesn't change adversarially, and the reward is stochastic. In other words, if we knew each w_a exactly, we would not need to explore, just like in the UCB1 setting.

LinUCB. To develop the LinUCB algorithm, we begin by extending (3). Our goal is to use confidence regions $C_{a,t}$ that contain each w_a with sufficient confidence. Suppose that we use ellipsoid confidence regions:

$$C_{a,t} = \{v \mid \|v - \hat{w}_{a,t}\|_{\Sigma_{a,t}^{-1}} \leq c_t\}, \tag{10}$$

where $\Sigma_{a,t}$ is a symmetric positive definite matrix, and:

$$\|v - \hat{w}_{a,t}\|_{\Sigma_{a,t}^{-1}} = \sqrt{(v - \hat{w}_{a,t})^T \Sigma_{a,t}^{-1} (v - \hat{w}_{a,t})}. \tag{11}$$

For instance, if $\Sigma_{a,t}$ were the identity matrix, then (11) would simply be the standard 2-norm, and $C_{a,t}$ would be a (uniformly shaped) ball of radius c_t centered at $\hat{w}_{a,t}$. Typically, $\hat{w}_{a,t}$ is estimated via ridge regression:

$$\hat{w}_{a,t} \equiv \text{argmin}_v\, \lambda \|v\|^2 + \sum_{t' : a_{t'} = a} (\hat{r}_{t'} - v^T x_{t',a})^2. \tag{12}$$

The form of (10) is reminicent of Gaussian confidence regions. For instance, if one were Bayesian and modelled the posterior distribution of w_a as Gaussian with mean $\hat{w}_{a,t}$ and covariance $\Sigma_{a,t}$, then (10) corresponds exactly to a high confidence region that contains w_a with confidence determined by the parameter c_t.

More generally, one only needs to assume that that the measurements we receive at time t, $\hat{r}_t$, are independent random variables with mean $w_{a_t}^T x_{t,a_t}$ and sub-Gaussian tails. It was shown in Theorem 2 in [1] that this is a sufficient condition for using ellipsoid confidence regions.

Using ellipsoid confidence regions, then the UCB1-style approach would be to choose:

$$a_t = \text{argmax}_{a \in \{1,\ldots,K\}} \max_{v \in C_{a,t}} v^T x_{t,a}, \tag{13}$$

which can be shown to be equivalent to:

$$a_t = \text{argmax}_{a \in \{1,\ldots,K\}} \hat{w}_a^T x_{t,a} + c_t \sqrt{x_{t,a}^T \Sigma_{a,t}^{-1} x_{t,a}}. \tag{14}$$

(14) is essentially the LinUCB algorithm that generalizes the basic K-armed bandit setting to the feature-based linear setting.

3 General Linear Stochastic Bandits

The disjoint linear contextual bandits setting above is a special case of general linear stochastic bandits [1], and in fact the LinUCB algorithm can applied to this more general setting as well with rigorous regret bounds. In this more general setting, at each iteration, we are given a set of A_t arms, each of which has a feature vector x_a. There is a single global weight vector w that characterizes the mean payoffs of each arm $w^T x_a$. The setting proceeds as:

1. Init $t = 1$
2. Receive contexts $x_{t,a}$ for $a \in A_t$
3. Algorithm chooses $a_t \in A_t$
4. Receive reward $\hat{r}_t$ with expected reward $E[\hat{r}_t] = w^T x_{t,a_t}$, and independent noise.
5. $t = t + 1$, repeat from Step 2 until $t = T$ (T possibly unknown)

Note that the set A_t can change from iteration to iteration.

In this case, LinUCB maintains a single $\hat{w}_t$, and at each iteration chooses the action such that:

$$a_t = \text{argmax}_{a \in A_t} \hat{w}_t^T x_a + c_t \sqrt{x_a^T \Sigma_t^{-1} x_a}. \tag{15}$$

The global ellipsoid confidence region for w is then:

$$C_t = \{v \mid \|v - \hat{w}_t\|_{\Sigma_t^{-1}} \leq c_t\}. \tag{16}$$

Regret is measured relative to mean reward of the optimal actions from each A_t:

$$R_T = \sum_{t=1}^{T} \left[w^T x_{t,a_t^*} - w^T x_{t,a_t} \right], \tag{17}$$

where $a_t^* = \text{argmax}_{a \in A_t} w^T x_{t,a}$.

Mapping to Disjoint Linear Contextual Bandits. One can map the linear contextual bandits setting into the general setting by simply choosing the global vector w as the concatenation of the K vectors from the contextual bandits setting:

$$w = \begin{bmatrix} w_1 \\ \vdots \\ w_K \end{bmatrix}, \tag{18}$$

and the action set A_t as set of K actions from the contextual bandits setting. The feature vector for action $a \in A_t$ is written as:

$$x_{t,a} = \begin{bmatrix} \vdots \\ 0 \\ \tilde{x}_{t,a} \\ 0 \\ \vdots \end{bmatrix}, \tag{19}$$

where $\tilde{x}_{t,a}$ denotes the feature vector from the disjoint linear contextual bandits setting. It is straightforward to check that this is an equivalent model, but now everything is written as a single global model. Furthermore, the "hybrid model" in [3] can also be written as some global model as well. This means that we can apply the technical results from [1] to develop regret bounds for the linear contextual bandits setting.

Regret Analysis. The regret analysis of LinUCB hinges on three conceptual steps:

1. If we have valid confidence intervals, then the regret analysis from (5)-(8) applies, and the regret can be bounded by the size of the confidence intervals used at each time step.

2. Theorem 2 in [1] provides general conditions on when ellipsoid confidence intervals are valid, which we elaborate on below.

3. Lemma 11 in [1] provides a bound on the sum of confidence intervals $\sum_{t=1}^{T} \hat{c}_{a_t,t}$, when they are ellipsoid confidence intervals. That is used in to prove the regret bound stated as Theorem 3 in [1]. We elaborate on this below as well.

Ellipsoid Confidence Interals. If we assume that everything is Gaussian, as described briefly in Section 2, then it is easy to show that for:

$$\Sigma_t = \lambda I + \sum_{t'=1}^{t} x_{t,a_t} x_{t,a_t}^T, \tag{20}$$

then (16) forms a valid confidence region that contains the true w with confidence controlled by c_t. Theorem 2 in [1] provides a more general condition where one can use ellipsoid confidence regions. Specifically, if the true model is linear, and the randomness of each feedback is independent with R-sub-Gaussian tails, then:

$$C_t = \left\{ v \;\middle|\; \|v - \hat{w}_t\|_{\Sigma_t^{-1}} \leq R\sqrt{2\log\left(\frac{\det(\Sigma_t)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)} + \lambda^{1/2} S \right\}, \tag{21}$$

where $1 - \delta$ is the target confidence and $S \geq \|w\|$. Generalizing from (12) to the global model setting, we have that $\hat{w}_t$ is:

$$\hat{w}_t \equiv \text{argmin}_v \, \lambda \|v\|^2 + \sum_{t'=1}^{t-1} (\hat{r}_{t'} - v^T x_{t',a_{t'}})^2. \tag{22}$$

Deploying this confidence region requires knowing that the noise has R-sub-Gaussian tails and an upper bound S on $\|w\|$. It can be shown that (10) is at most:

$$\mathcal{O}(R\sqrt{DK \log(t/\delta)} + \lambda^{1/2} S), \tag{23}$$

where DK is the total dimensionality of w.

Regret Bound. Assume that the regularization parameter λ satisfies $\lambda \geq \max_x \|x\|^2$. Following (a simplification of) the proof of Theorem 3 in [1], starting from (8), we have:

$$\begin{aligned}
R_T &\leq 2 \sum_{t=1}^{T} \hat{c}_{a_t,t} && (24)\\
&= 2 \sum_{t=1}^{T} c_t \|x_{t,a_t}\|_{\Sigma_t^{-1}} && (25)\\
&\leq 2 \sqrt{\sum_{t=1}^{T} c_t^2 \|x_{t,a_t}\|_{\Sigma_t^{-1}}^2} && (26)\\
&\leq 2 \sqrt{c_T^2 \sum_{t=1}^{T} \|x_{t,a_t}\|_{\Sigma_t^{-1}}^2} && (27)\\
&= 2 c_T \sqrt{\sum_{t=1}^{T} \|x_{t,a_t}\|_{\Sigma_t^{-1}}^2} && (28)
\end{aligned}$$

(25) is a substitution of the definition of the confidence interval. (26) follows from Cauchy-Schwartz. (27) follows from c_t being monotincally increasing.

Lemma 11 in [1] proved that, for $\lambda \geq \max_x \|x\|^2$, we have:

$$\sum_{t=1}^{T} \|x_{t,a_t}\|^2_{\Sigma_t^{-1}} \leq 2 \log \det(\Sigma_T) \leq \mathcal{O}(DK \log(T)), \tag{29}$$

which combined with (28) and (23) yield a regret bound of:

$$R_T = \mathcal{O}\left(RS\lambda^{1/2} DK \log(T/\delta)\sqrt{T}\right), \tag{30}$$

where R is the width of the tail of the randomness of the feedback, $S \geq \|w\|$ is an upper bound on the norm of the true model, λ is the regularization parameter for estimating the mean and is required to be $\lambda \geq \max_x \|x\|^2$, DK is the total dimensional of the global model, and $1 - \delta$ is the target confidence.

Acknowledgments

Thanks to Pengchuan Zhang for helping proofread and fix typos and inconsistent notation.

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Neural Information Processing Systems (NIPS)*, 2011.

[2] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[3] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *International World Wide Web Conference (WWW)*, 2010.