CSC 696H:

LinUCB analysis: Confidence Set Construction & regret analysis

Lecturer: Chicheng Zhang *Scribe: Brandon Hall*

1 Review: Linear Contextual Bandits

1.1 Basic overview of LCB

For $t = 1, 2, \ldots, T$:

- Observe context $x_t \in \mathcal{X}$
- Take action $a_t \in \mathcal{A}$
- Receive reward $r_t = f^*(x_t, a_t) + \underbrace{\epsilon_t}_{\text{zero-mean 1-SG}}$

Goal: Maximize $\mathbb{E}\left[\sum_{t=1}^{T} f^*(x_t, a_t)\right] = \mathbb{E}\left[\sum_{t=1}^{T} r_t\right]$
Linearity: $\underbrace{f^*}_{f^*(x,a)=\langle\theta^*,\Phi(x,a)\rangle} \in \mathcal{F} = \{f(x,a) = \langle\theta, \Phi(x,a)\rangle : \|\theta\|_2 \le 1\}$

1.2 Performance Measure of LCB

The equation gives the performance measure:

$$\text{Reg}(T) = \mathbb{E}\left[\sum_{t=1}^{T} \underbrace{\max_{a\in\mathcal{A}} \left(f^*(x_t, a) - f^*(x_t, a_t)\right)}_{InstintaneousRegret}\right]$$

In this equation, the term $f^*(x_t, a) - f^*(x_t, a_t)$ represents the instantaneous regret at time t, which measures the difference between the reward of the best possible action a and the reward of the action taken a_t.

The entire summation, $\sum_{t=1}^{T} \max_{a\in\mathcal{A}} \underset{I}{\left(f^*(x_t, a) - f^*(x_t, a_t)\right)}$, is referred to as the pseudo-regret (PReg) over the time horizon T, capturing the cumulative regret incurred across all time steps.

1.3 The UCB Algorithm for Linear Contextual Bandits

For $t = 1, 2, \ldots, T$:

- Construct the confidence set $\mathbf{\Theta}_t$ for θ^*.
  - **Hope:** $\theta^* \in \mathbf{\Theta}_t$ - This implies that we believe the true parameter θ^* is contained within our constructed set, allowing us to make reliable inferences about the reward function based on this assumption.
  - For this rest of this iteration, assume that $\theta^* \in \Theta_t$ is the only information we know about the ground truth reward predictor θ^*. In other words, Θ_t is the set of all "plausible" values of θ^*.
- Observe the context x_t.

- For every action a, find the highest plausible reward. This is determined by calculating the upper confidence bound $\text{UCB}_t(x_t, a)$:

$$\max_{\theta \in \Theta_t} \langle \theta, \phi(x_t, a) \rangle = \text{UCB}_t(x_t, a)$$

- Take the action:

$$a_t = \arg\max_{a \in \mathcal{A}} \text{UCB}_t(x_t, a)$$

  This step selects the action a_t that maximizes the upper confidence bound, balancing between exploiting known information and exploring uncertain options.

2 Analysis of LinUCB

2.1 Main topics of this lecture

- **Q1:** How to compute the confidence set $\mathbf{\Theta}_t$?
- **Q2:** How can we analyze the regret of LinUCB?

2.2 How to compute the confidence set $\mathbf{\Theta}_t$?

Answer 1: We are going to do this in two steps:

$$\begin{Bmatrix} \begin{cases} 1. & \text{Best guess of } \theta^* \text{ using data} \to \hat{\theta}^t \\ 2. & \text{Quantify the closeness between } \hat{\theta}^t \text{ and } \theta^* \end{cases} \end{Bmatrix}$$

2.2.1 Answer to Q1: Computing the Confidence Set $\mathbf{\Theta}_t$

The estimate $\hat{\theta}^t(\gamma)$ is computed as follows:

$$\hat{\theta}^t(\gamma) = \arg\min_{\theta} \left\{ \sum_{s=1}^{t-1} \left(\langle \theta, \Phi(x_s, a_s) \rangle - r_s \right)^2 + \gamma \|\theta\|^2 \right\}$$

In this equation: - The first term, $\sum_{s=1}^{t-1} \left(\langle \theta, \Phi(x_s, a_s) \rangle - r_s \right)^2$, represents the squared loss incurred from the estimated rewards compared to the actual rewards observed. - The second term, $\gamma \|\theta\|^2$, acts as a regularization term, ensuring that the estimated parameters do not grow excessively large.

For convenience, we denote $\Phi(x_s, a_s)$ as Φ_s to simplify notation.

2.2.2 Answer to Q2: Quantify the closeness between $\hat{\theta}^t$ and θ^*

We cannot expect $\hat{\theta}^t(\gamma)$ to be coordinate-wise close to $\theta^* = \begin{pmatrix} \theta_1^* \\ \theta_2^* \end{pmatrix}$.

This limitation can be illustrated through the following example:
- Let the dimension:

$$d = 2$$

- Consider the feature vectors:

$$\Phi_1, \dots, \Phi_{t-2} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$\Phi_{t-1} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

- The observed rewards are:

$$r_1, \dots, r_{t-2}, r_{t-1}$$

In this scenario:

- All the rewards $r_1, \ldots, r_{t-2}$ only depend on the parameter θ_1^* and not θ_2^*. This is because the first t-1 feature vectors have a 0 in the second part of the vector. This causes them to give no information on θ_2^*.

- Similarly, the reward r_{t-1} is influenced solely by θ_2^*.

This relationship is illustrated in Figure 1, which depicts the dependence graph of these variables:

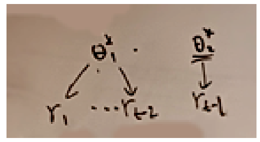

Figure 1: This diagram demonstrates how θ_1^* affects the rewards r_1 to r_{t-2}, while having no influence on r_{t-1}. Conversely, θ_2^* solely impacts r_{t-1}. This indicates that r_{t-1} does not have sufficient data to enable accurate estimation.

This disparity highlights the challenge that certain parameters may not have enough observed data for reliable estimation. Consequently, it becomes unrealistic to expect coordinate-wise closeness between $\hat{\theta}^t(\gamma)$ and θ^*.

Lemma 1 (Confidence Set)**.** *There exists an event E such that $P(E) \geq 1 - \frac{1}{T}$ and, on the event E, for all t:*

$$\theta^* \in \mathbf{\Theta}_t(\lambda) = \left\{ \theta : \|\theta - \hat{\theta}^t(\lambda)\|_{V_t-1(\lambda)} \leq B_t(\lambda) = \hat{O}\left(\sqrt{\lambda} + \sqrt{d}\right) \right\}$$

Specifically, if $\lambda = 1$:

$$\theta^* \in \mathbf{\Theta}_t(1) = \left\{ \theta : \|\theta - \hat{\theta}^t(1)\|_{V_t-1}(\lambda)^2 \leq B_t(1) = \hat{O}\left(\sqrt{d}\right) \right\}$$

Positive Semi-definite Matrices

A matrix $M \in \mathbb{R}^{n\times n}$ is said to be **positive semi-definite** (PSD) if for all vectors $\mathbf{v} \in \mathbb{R}^n$:

$$\mathbf{v}^T M \mathbf{v} \geq 0$$

This means that the quadratic form associated with M is non-negative for all vectors $\mathbf{v}$.
Example 1: Consider the identity matrix:

$$M = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

For any vector $\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$, we have:

$$\mathbf{v}^T M \mathbf{v} = \begin{pmatrix} v_1 & v_2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = v_1^2 + v_2^2 \geq 0$$

Since v_1^2 and v_2^2 are both non-negative, M is positive semi-definite.
A matrix M is **positive definite** (PD) if for all vectors $\mathbf{v} \neq \mathbf{0}$:

$$\mathbf{v}^T M \mathbf{v} > 0$$

Example 2: Consider the matrix:

$$M = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

For the vector $\mathbf{v} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$:

$$\mathbf{v}^T M \mathbf{v} = \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = 0$$

Here, $\mathbf{v}^T M \mathbf{v} = 0$, which shows that M is positive semi-definite, but not positive definite because there exists a vector $\mathbf{v}$ such that $\mathbf{v}^T M \mathbf{v} = 0$.

Mahalanobis Norm

The Mahalanobis norm, denoted as $\|\mathbf{x}\|_M$ for a positive definite matrix $M\geq\succ 0$, is defined as:

$$\|\mathbf{x}\|_M = \sqrt{\mathbf{x}^T M \mathbf{x}}$$

Notably, when $M = I$ (the identity matrix), the Mahalanobis norm reduces to the standard Euclidean norm:

$$\|\mathbf{x}\|_M = \|\mathbf{x}\|_2$$

Properties of the Mahalanobis Norm:

- $\|\mathbf{x}\|_M \geq 0$
- $\|a\mathbf{x}\|_M = |a| \cdot \|\mathbf{x}\|_M$ (where a is a scalar)
- $\|\mathbf{x} + \mathbf{y}\|_M \leq \|\mathbf{x}\|_M + \|\mathbf{y}\|_M$ (generalized triangle inequality)
- $\langle \mathbf{x}, \mathbf{y} \rangle \leq \|\mathbf{x}\|_M \cdot \|\mathbf{y}\|_{M^{-1}}$

  (generalized Cauchy-Schwarz)

  Cauchy-Schwarz The original Cauchy-Schwarz inequality states that for any two vectors $\mathbf{u}$ and $\mathbf{v}$:

  $$|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\| \cdot \|\mathbf{v}\|$$

  This means the absolute value of the inner product of two vectors is less than or equal to the product of their magnitudes (norms). In the context of the Mahalanobis norm, we use the generalized version, which allows us to work with the Mahalanobis norm defined by a positive definite matrix M. Here, we compare the inner product $\langle \mathbf{x}, \mathbf{y} \rangle$ with $\|\mathbf{x}\|_M \cdot \|\mathbf{y}\|_{M^{-1}}$, which incorporates the properties of M and its inverse.

  Use of M^{-1} You may wonder why we use $\|\mathbf{y}\|_{M^{-1}}$ instead of $\|\mathbf{y}\|_M$ on the right-hand side. If we were to use $\|\mathbf{y}\|_M$, the right side would yield $\sqrt{M\mathbf{x}^2} \cdot \sqrt{M\mathbf{y}^2}$, resulting in an expression that contains two instances of the matrix M. We ensure the dimensions balance correctly by using M^{-1} for the second term.

- $\max_{\|\mathbf{x}\|_M < 1} \langle \mathbf{x}, \mathbf{y} \rangle = \|\mathbf{y}\|_{M^{-1}}$

In our analysis, we will utilize the Mahalanobis norm to define the covariance matrix for our data. Specifically, we express the update for the covariance matrix $V_{t+1}(\lambda)$ as follows:

$$V_{t+1}(\lambda) = \sum_{s=1}^{t-1} \Phi_s \Phi_s^T + \lambda I$$

Here, $V_{t-1}(\lambda)$ appears in the lemma, where the summation represents the data covariance matrix accumulated up to step $t-1$. The addition of λI serves as a regularization term, ensuring the matrix remains positive definite. ‘

Lemma Example

Consider the following example where $\hat{\theta}_t \in \mathbb{R}^2$, and the covariance matrix $V_{t-1}(0)$ with $\lambda = 0$ is computed as:

$$V_{t-1}(0) = \sum_{s=1}^{t-1} \Phi_s \Phi_s^T = (t-2)\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + 1\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} t-2 & 0 \\ 0 & 1 \end{pmatrix}$$

Side Note: The individual matrices were derived from:

$$\Phi_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \Phi_1 \Phi_1^T = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

$$\Phi_{t-1} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \Phi_{t-1} \Phi_{t-1}^T = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

Now, the confidence set $\Theta_t(0)$ for θ can be written as:

$$\Theta_t(0) = \left\{ \theta : \sqrt{\begin{pmatrix} \theta_1 - \hat{\theta}_1^t & \theta_2 - \hat{\theta}_2^t \end{pmatrix} \begin{pmatrix} t-2 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \theta_1 - \hat{\theta}_1^t \\ \theta_2 - \hat{\theta}_2^t \end{pmatrix}} \leq \sqrt{2} \right\}$$

Expanding the quadratic form:

$$\Theta_t(0) = \left\{ \theta : \frac{(t-2)}{2}(\theta_1 - \hat{\theta}_1^t)^2 + \frac{1}{2}(\theta_2 - \hat{\theta}_2^t)^2 \leq 1 \right\}$$

This equation represents an ellipse in the (θ_1, θ_2)-plane, with the center at $(\hat{\theta}_1^t, \hat{\theta}_2^t)$. The lengths of the semi-axes are determined by the coefficients of θ_1 and θ_2, which reflect the scaling due to the covariance matrix.
To explicitly show this is an ellipse, we can rewrite it in standard ellipsoid form:

$$\Theta_t(0) = \left\{ \theta : \frac{(\theta_1 - \hat{\theta}_1^t)^2}{\sqrt{\frac{2}{(t-2)^2}}} + \frac{(\theta_2 - \hat{\theta}_2^t)^2}{\sqrt{2}^2} \leq 1 \right\}$$

Explanation: - The center of the ellipse is $(\hat{\theta}_1^t, \hat{\theta}_2^t)$, meaning the estimate $\hat{\theta}^t$ is at the center. - The size of the ellipse is determined by the scaling terms, $t-2$ and 1, which influence the axes' lengths. Specifically, the axis along θ_1 is scaled by $t-2$, while the axis along θ_2 is scaled by 1.
This ellipse represents the confidence region for the parameter θ^*, with different scaling along the two axes, reflecting the uncertainty in each dimension.

2.3 Implications of Confidence Bounds

For directions in which we have more data available, the uncertainty in θ^* along those directions is reduced. This is an important observation that arises from the structure of the covariance matrix and how it affects our confidence intervals.

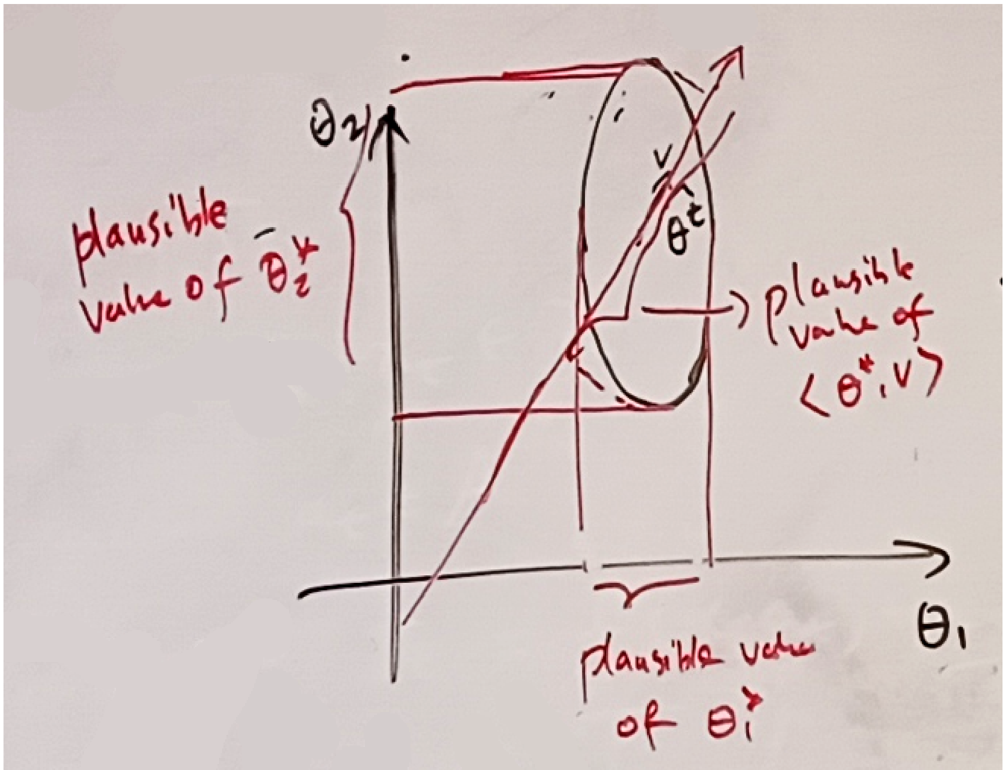

Figure 2: Graph of the confidence ellipse derived from the lemma example. Along the θ_1 (x-axis), the width is $\sqrt{\frac{2}{(t-2)^2}}$, while along the θ_2 (y-axis), the width is 2. This shows that with fewer data points in the direction of θ_2, the plausible range of θ_2 values is larger compared to θ_1, where we have more data.

In Figure 2, the confidence ellipse shows the uncertainty regions for θ_1 and θ_2. On the θ_1-axis, the smaller width represents higher certainty in this direction due to having more data points. In contrast, the θ_2-axis has a larger width, reflecting higher uncertainty since only one data point is available in that direction.

2.4 Proof Sketch: Bound Derivation

To formalize this, consider the following proof sketch based on the data observed up to time $t-1$. We will assume that λ is very small, so that $V_{t-1}(\lambda) \approx V_{t-1}(0)$. Below, we use $\hat{\theta}_t$ and V_{t-1} as shorthands of $\hat{\theta}_t(\lambda)$ and $V_{t-1}(\lambda)$, respectively.

The estimate $\hat{\theta}_t$ can be expressed as:

$$\hat{\theta}_t = V_{t-1}^{-1}\left(\sum_{s=1}^{t-1} \phi_s r_s\right)$$

Breaking this down into two components:

$$= V_{t-1}^{-1}\left(\sum_{s=1}^{t-1} \phi_s \phi_s^T \theta^* + \sum_{s=1}^{t-1} \phi_s \epsilon_s\right)$$

(See the handwritten notes for a full derivation.)

Which gives us:

$$\hat{\theta}_t \approx \theta^* + V_{t-1}^{-1} \left(\sum_{s=1}^{t-1} \phi_s \epsilon_s \right)$$

The error term, $V_{t-1}^{-1} \sum_{s=1}^{t-1} \phi_s \epsilon_s$, accounts for the residuals from noisy data.

2.4.1 Measuring Distance Using Mahalanobis Norm

To measure the distance between $\hat{\theta}_t$ and θ^*, we use the Mahalanobis norm, as introduced earlier:

$$\|\hat{\theta}_t - \theta^*\|_{V_{t-1}} = \left\| V_{t-1} \sum_{s=1}^{t-1} \phi_s \epsilon_s \right\|_{V_{t-1}}$$

This simplifies to:

$$= \left\| \sum_{s=1}^{t-1} \phi_s \epsilon_s \right\|_{V_{t-1}}$$

With high probability (w.h.p), this bound is:

$$\leq \sqrt{2 \ln(T) + \alpha \ln(1 + t/d)}$$

$$= B_t(\lambda) = \mathcal{O}(\sqrt{d})$$

Additional Notes: This bound was obtained using the Self-Normalized Tail Inequality. For further reading, refer to Abbasi-Yadkori, Pal, and Szepesvári's work on the topic.

Aside: Mahalanobis Norm Simplification

To simplify expressions like $\|M^{-1}x\|_M$, we use the following logic:

$$\|M^{-1}x\|_M = \sqrt{(M^{-1}x)^T M (M^{-1}x)} = \sqrt{x^T M^{-1} x}$$

Here, the M^{-1} and M matrices cancel out, leaving $\sqrt{x^T M^{-1} x}$, which is the form we end up with for the Mahalanobis norm.

2.5 Novelty and Elliptical Potential

The term $||\Phi(x_t, a)||_{V_{t-1}^{-1}}$ quantifies how "novel" the feature vector $\Phi(x_t, a)$ is concerning the observations $\Phi_1, \ldots, \Phi_{t-1}$ that have been previously encountered. This relationship is visualized through the concept of "Elliptical Potential," as depicted in Figure 3.

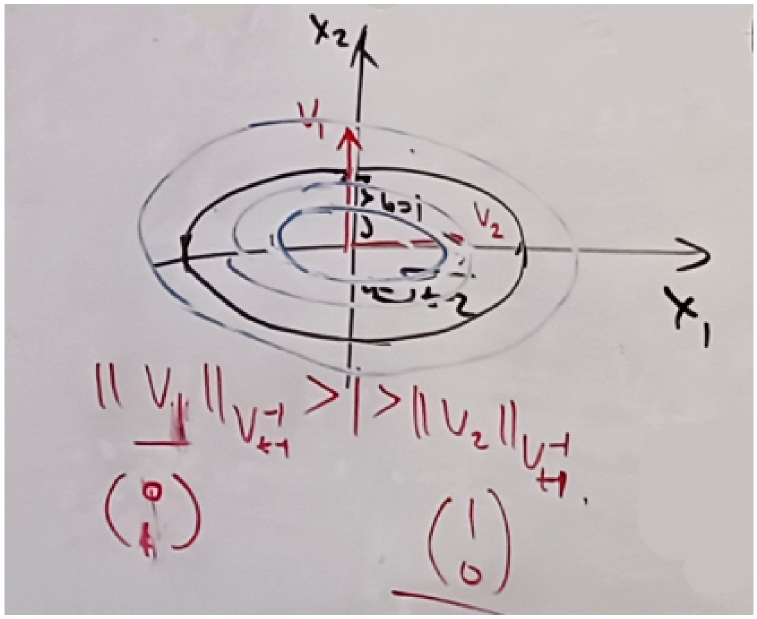

Figure 3: Visualization of the elliptical potential defined by the condition $\{x : ||x||_{V_{t-1}^{-1}} \leq 1\}$. Points inside the ellipse indicate lower novelty, while points outside suggest higher novelty, illustrating the confidence interval for the feature vectors.

The contours of points of equal elliptical potential can be expressed mathematically as:

$$\{x : ||x||_{V_{t-1}^{-1}} = c\}$$

for different values of c; a larger c corresponds to a larger ellipse.

This representation implies that any point x satisfying this condition lies within the ellipse, while those that do not satisfy it lie outside.

To illustrate this concept, we reference a previous example where the inverse covariance matrix is defined as:

$$V_{t-1}^{-1} = \begin{pmatrix} \frac{1}{t-2} & 0 \\ 0 & 1 \end{pmatrix}$$

From this, we can calculate the corresponding elliptical potential as follows:

$$(x_1, x_2) \begin{pmatrix} \frac{1}{t-2} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \leq 1 \implies \frac{x_1^2}{t-2} + x_2^2 \leq 1$$

Here, $\frac{1}{t-2}$ represents a^2 and 1 represents b^2 in the ellipse equation, which gives us $a = \sqrt{t-2}$ and $b = 1$. This is plotted in Figure 3 as the black line.

This example demonstrates that as we accumulate more data on x_1, the ellipse becomes thinner along that axis, indicating a smaller confidence interval and therefore reduced elliptical potential. Inversely, features that have not been observed as frequently will yield a wider ellipse, reflecting greater uncertainty in their estimates.

Additionally, the contours represented in Figure 3 provide insights into the novelty of the feature vectors. The contours inside the black line correspond to points where the V_{t-1}-norm is less than 1, indicating regions of lower novelty. In contrast, points lying outside the black line have an V_{t-1}-norm greater than 1, suggesting areas of higher novelty. These contours serve as a visual representation of the varying levels of confidence we hold in our parameter estimates based on the observed data.

In summary, the elliptical potential and its corresponding contours provide a framework for understanding how novelty is assessed in the context of linear contextual bandits, ultimately guiding decision-making based on the confidence intervals established through past observations.

Final Remark: this ellipse has a shape exactly "opposite" to the confidence set. Do you think it makes sense?

2.6 The UCB Algorithm for Linear Contextual Bandits (again)

This is a repeat from before with some added information

For $t = 1, 2, \ldots, T$:

- Construct the confidence set $\mathbf{\Theta}_t$ for θ^*.
  - **Hope:** $\theta^* \in \mathbf{\Theta}_t$ - This implies that we believe the true parameter θ^* is contained within our constructed set, allowing us to make reliable inferences about the reward function based on this assumption.
  - For this rest of this iteration, assume that $\theta^* \in \Theta_t$ is the only information we know about the ground truth reward predictor θ^*. In other words, Θ_t is the set of all "plausible" values of θ^*.
- Observe the context x_t.
- For every action a, find the highest plausible reward. This is determined by calculating the upper confidence bound $\mathrm{UCB}_t(x_t, a)$. The process is as follows:
  1. First, compute the maximum inner product between θ and the context-action pair $\phi(x_t, a)$, constrained to the confidence set $\mathbf{\Theta}_t$:
  $$\max_{\theta \in \mathbf{\Theta}_t} \langle \theta, \phi(x_t, a) \rangle = \mathrm{UCB}_t(x_t, a)$$
  2. Then, approximate the reward using the current estimate $\hat{\theta}_t$ to obtain a more explicit expression of $\mathrm{UCB}_t(x_t, a)$ by solving the maximization problem.
  $$\begin{aligned}\mathrm{UCB}_t(x_t, a) &= \max_{\theta : \|\theta - \hat{\theta}_t\|_{V_{t-1}} \le \beta_t(1)} \langle \theta, \phi(x_t, a) \rangle \\ &= \max_{z : \|z\|_{V_{t-1}} \le \beta_t(1)} \langle \hat{\theta}_t, \phi(x_t, a) \rangle + \langle z, \phi(x_t, a) \rangle \qquad \text{(Change of variable } \theta = \hat{\theta}_t + z\text{)} \\ &= \langle \hat{\theta}_t, \phi(x_t, a) \rangle + \beta_t(1) \|\phi(x_t, a)\|_{V_{t-1}^{-1}}\end{aligned}$$
  3. Finally, introduce the Exploration Bonus, which accounts for the uncertainty in the estimate:
  $$\mathrm{UCB}_t(x_t, a) = \langle \hat{\theta}_t, \phi(x_t, a) \rangle + \beta_t(1) \|\phi(x_t, a)\|_{V_{t-1}^{-1}}$$
  This Exploration Bonus term at the end encourages exploring actions where we have less confidence, as it grows with the uncertainty $\|\phi(x_t, a)\|_{V_{t-1}^{-1}}$.
- Take the action:
  $$a_t = \arg\max_{a \in \mathcal{A}} \mathrm{UCB}_t(x_t, a)$$
  This step selects the action a_t that maximizes the upper confidence bound, balancing between exploiting known information and exploring uncertain options.

3 Regret Analysis

In this section, we analyze the regret associated with the Linear Upper Confidence Bound (LinUCB) algorithm. This section was discussed quickly as time was limited.

3.1 Theorem

The following theorem encapsulates the regret bounds for LinUCB:

Theorem 2. *For a given event E such that $P(E) \geq 1 - \frac{1}{T}$, we have the following bound on the cumulative regret:*

$$Reg(T) \leq \tilde{O}(d\sqrt{T}),$$

where d represents the dimensionality of the feature space and T is the number of time steps. This result is generally unimprovable.

Additionally, when applying LinUCB to a multi-armed bandit (MAB) problem, we observe that:

$$\text{Reg}(T) \leq A\sqrt{T},$$

where A is a constant that reflects the problem's specific characteristics. This is notable because the standard Native UCB approach yields a regret of $\sqrt{AT}$, which is better than the regret incurred by LinUCB.

3.2 Proof

The cumulative regret over T time steps can be expressed as:

$$\text{PReg}(T) = \sum_{t=1}^{T} \text{reg}_t,$$

where reg_t denotes the regret at time t.

3.2.1 Claim 1

For each time step t, we assert the following:

$$\text{reg}_t \leq 2b_t(a_t),$$

where $b_t(a_t)$ represents the bound on the confidence interval for the action a_t. This claim is similar to the analysis of upper confidence bounds in the multi-armed bandit (MAB) context. For a detailed understanding, refer to the handwritten notes accompanying this section.

3.2.2 Claim 2

$$\sum_{t=1}^{T} b_t(a_t) \leq \tilde{O}(d\sqrt{T}).$$

This observation can be rewritten as:

$$\beta_t(1) \sum_{t=1}^{T} ||\Phi(x_t, a_t)||_{V_{t-1}^{-1}},$$

which provides insight into how novel the feature vector $\Phi(x_t, a_t)$ is in relation to the previously observed feature vectors $\Phi_1, \ldots, \Phi_{t-1}$.

This concept ties directly into the earlier discussed Elliptical Potential, where the novelty of the observed feature vector influences the bounds on the regret. In our next lecture we will introduce the Elliptical Potential Lemma that provide a bound on this.