Recap : Stochastic Contextual Bandit setup

- Stochastic contextual bandits generalize adversarial contextual bandits by adding a stochastic reward model.
- At each round t:
  - The learner observes a context C_t.
  - Then chooses an action $A_t \in [k]$.
- The reward is given by:

$$X_t = r(C_t, A_t) + \eta_t$$

  where:
  - r is the expected reward function.
  - η_t is a noise term.
- The noise η_t is conditionally 1-subgaussian given past observations.
- This implies:

$$\mathbb{E}[X_t \mid \mathcal{F}_t] = r(C_t, A_t)$$

  and the noise has zero mean.

Terminology

- Let:
  - $C_t \in \mathcal{C}$: context at round t
  - $A_t \in [k]$: action chosen at round t
  - $X_t \in \mathbb{R}$: reward received
  - $r : \mathcal{C} \times [k] \to \mathbb{R}$: unknown expected reward function
  - η_t: noise
- Then the stochastic contextual bandit model assumes:

$$X_t = r(C_t, A_t) + \eta_t$$

- With the assumptions:
  - $\mathbb{E}[\eta_t \mid \mathcal{F}_t] = 0$
  - η_t is conditionally 1-subgaussian, i.e.,

$$\mathbb{E}\left[\exp(\lambda\eta_t) \mid \mathcal{F}_t\right] \leq \exp\left(\frac{\lambda^2}{2}\right) \quad \forall\lambda \in \mathbb{R}$$

- Here, the filtration $\mathcal{F}_t$ is defined as:

$$\mathcal{F}_t = \sigma(C_1, A_1, X_1, \ldots, C_{t-1}, A_{t-1}, X_{t-1}, C_t, A_t)$$

- Because this filtration captures all the past data and the current action, but not the current reward.

A motivating example

An online ad platform must select an advertisement to display to a user each time they visit a webpage. The objective is to maximize the cumulative **click-through rate (CTR)** over time.

At each round $t = 1, 2, \ldots, n$:

- The user is described by a **context vector** $c_t \in C$, encoding features such as:
  - Demographics (e.g., age, gender, location)
  - Device type (mobile, desktop)
  - Temporal features (hour, weekday)
  - Behavioral data (past clicks, interests)
- The learner must select one advertisement from a finite action set $\mathcal{A} = \{1, 2, \ldots, k\}$. Each ad $a \in \mathcal{A}$ has its own features, such as:
  - Product category
  - Target audience
  - Visual design style

Feature map ψ , and Regret

Assume a known joint feature map:

$$\psi : \mathcal{C} \times \mathcal{A} \to \mathbb{R}^d$$

that encodes the interaction between user and ad.
We assume the reward model is linear:

$$r(c_t, a) = \langle \theta^*, \psi(c_t, a) \rangle$$

where $\theta^* \in \mathbb{R}^d$ is an unknown parameter vector capturing latent preferences.

Feature map ψ , and Regret

After choosing action $A_t \in \mathcal{A}$, the learner observes:

$$X_t = r(c_t, A_t) + \eta_t \in \{0, 1\}$$

where:

- $X_t = 1$ if the user clicks the ad,
- $X_t = 0$ otherwise,
- η_t is 1-subgaussian noise (zero-mean, bounded variance).

Feature map ψ , and Regret

- **Generalization across ads:** The feature map allows learning across similar user-ad pairs, reducing sample complexity.
- **Sample-efficient exploration:** LinUCB selects actions optimistically using confidence ellipsoids around θ^*.
- **Scalability:** Only a single $\theta^* \in \mathbb{R}^d$ needs to be learned, rather than a separate reward for every (c, a) pair.

Linearity assumption

- If the true reward function $r(c, a)$ were known, the learner could act optimally at each round.
- **Regret** measures the performance gap due to not knowing r.
- In the worst case, estimating $r(c, a)$ for every pair (c, a) is infeasible — especially when the context space is large.
- A powerful workaround: assume rewards are linear in a feature map $\psi(c, a)$ - concat,NN,etc..
- This yields the **stochastic linear contextual bandit** model.
- Smoothness of r can be controlled via bounds on $\|\theta^*\|$.

Feature map ψ , and Regret

- Let:
  - $C_t \in \mathcal{C}$: context at round t
  - $A_t \in [k]$: chosen action at round t
  - $\psi : \mathcal{C} \times [k] \to \mathbb{R}^d$: feature map
  - $\theta^* \in \mathbb{R}^d$: unknown parameter vector
  - $X_t = r(C_t, A_t) + \eta_t$: reward with 1-subgaussian noise η_t
- Assume the **linear reward model**:

$$r(c,a) = \langle \theta^*, \psi(c,a) \rangle \quad \text{for all } (c,a) \in \mathcal{C} \times [k]$$

- Define the optimal action at round t as:

$$A_t^* = \arg\max_{a \in [k]} r(C_t, a)$$

- Then the cumulative **regret** over n rounds is defined as:

$$R_n = \mathbb{E}\left[\sum_{t=1}^{n} \left(r(C_t, A_t^*) - r(C_t, A_t)\right)\right] = \mathbb{E}\left[\sum_{t=1}^{n} \left(\max_{a \in [k]} r(C_t, a) - X_t\right.\right.$$

- This measures the cumulative performance gap caused by not knowing the reward function r

A lower bound on regret

- **Lower Bound (Tabular Case)**:
- If you must learn $r(c, a)$ for all M contexts and k actions, the worst-case cumulative regret is:

$$\Omega(\sqrt{nMk})$$

- This becomes infeasible when M is large (e.g., $M = 2^{100}$).

Stochastic linear bandit

Stochastic linear bandit

Stochastic linear bandit

- Linear contextual bandits simplify into the **stochastic linear bandit** setting.
- All that matters is the **feature vector** — not the specific identity of the action.
- At round t, the learner selects an action:

$$A_t \in \mathcal{A}_t \subset \mathbb{R}^d$$

- The reward is linear in the chosen action:

$$X_t = \langle \theta^*, A_t \rangle + \eta_t$$

  where:
  - $\theta^* \in \mathbb{R}^d$ is an unknown parameter vector,
  - η_t is 1-subgaussian noise.
- Pseudo-regret and expected regret are both defined over these chosen actions A_t.
- **Special cases include:**
  - Finite-armed bandits
  - Contextual bandits
  - Combinatorial linear bandits

Stochastic linear bandit

- We simplify Eq. (19.1):

$$r(c,a) = \langle \theta^*, \psi(c,a) \rangle \quad \Rightarrow \quad X_t = \langle \theta^*, A_t \rangle + \eta_t$$

- Where:
  - $A_t \in \mathcal{A}_t \subset \mathbb{R}^d$: decision/action in round t
  - $\theta^* \in \mathbb{R}^d$: unknown parameter vector
  - η_t: 1-subgaussian noise, i.e.,

$$\mathbb{E}\left[e^{\lambda \eta_t} \mid \mathcal{F}_t\right] \leq \exp\left(\frac{\lambda^2}{2}\right) \quad \text{for all } \lambda \in \mathbb{R}$$

- **Pseudo-Regret:**

$$\widehat{R}_n = \sum_{t=1}^{n} \left(\max_{a \in \mathcal{A}_t} \langle \theta^*, a \rangle - \langle \theta^*, A_t \rangle \right)$$

- **Expected Regret:**

$$R_n = \mathbb{E}[\widehat{R}_n] = \mathbb{E}\left[\sum_{t=1}^{n} \left(\max_{a \in \mathcal{A}_t} \langle \theta^*, a \rangle - X_t \right) \right]$$

Stochastic linear bandit

- **UCB (Upper Confidence Bound)** is a powerful method in stochastic bandits.
- It can be generalized to linear bandits using the **optimism in the face of uncertainty (OFU)** principle.
- The generalization involves:
  - Estimating θ^* with a confidence set $C_t \subset \mathbb{R}^d$
  - Selecting actions by solving:

$$A_t = \arg\max_{a \in \mathcal{A}_t} \max_{\theta \in C_t} \langle \theta, a \rangle$$

- The resulting algorithm is known as **LinUCB** or **OFUL (Optimism in the Face of Uncertainty for Linear bandits)**.
- **Key challenge:** Constructing the confidence set C_t such that:
  - It contains θ^* with high probability,
  - While remaining as small as possible to ensure good exploration-exploitation balance.

Stochastic linear bandit

Stochastic linear bandit

Stochastic linear bandit

- **Define Confidence Set:**
  Let $C_t \subset \mathbb{R}^d$ be a confidence set such that:
  $$\mathbb{P}\left[\theta^* \in C_t\right] \geq 1 - \delta$$
- **Define UCB Estimate:**
  For any action $a \in \mathbb{R}^d$, define:
  $$\mathrm{UCB}_t(a) = \max_{\theta \in C_t} \langle \theta, a \rangle \tag{19.2}$$
  This gives an upper bound on the expected reward of a, under uncertainty about θ^*.
- **LinUCB Selection Rule:**
  $$A_t = \arg\max_{a \in \mathcal{A}_t} \mathrm{UCB}_t(a) \tag{19.3}$$
  This rule selects the action that has the highest optimistic reward estimate based on the current confidence set C_t.

Stochastic linear bandit

- **Where is the challenge?**
- Choosing the confidence set C_t is non-trivial:
  - It is no longer a simple scalar interval (as in basic bandits).
  - It must contain the true parameter θ^* with high probability.
  - Yet, it must also be as small (tight) as possible to avoid unnecessary exploration.
- We will later construct C_t as an ellipsoid:
- **Proof/Derivations (Preview)**:
- To build the confidence set C_t, we proceed as follows:
- Use **regularized least squares** to estimate θ^*:

$$\hat{\theta}_t = \arg\min_{\theta \in \mathbb{R}^d} \sum_{s=1}^{t} (X_s - \langle \theta, A_s \rangle)^2 + \lambda \|\theta\|^2$$

Stochastic linear bandit

why λ?

Stochastic linear bandit

contour plots.

We aim to minimize the empirical squared loss:

$$L(\theta) = \sum_{s=1}^{t} (X_s - A_s^\top \theta)^2$$

$$\nabla_\theta L(\theta) = -2A^\top (X - A\theta) = 0$$

$$A^\top A\theta = A^\top X$$

Stochastic linear bandit

Stochastic linear bandit

- Let's compute the gradient of the regularized least squares loss function $L(\theta)$, and set it to zero to solve for $\hat{\theta}_t$.
- Expand the loss function:

$$L(\theta) = \sum_{s=1}^{t} \left(X_s - A_s^\top \theta \right)^2 + \lambda \|\theta\|^2$$

- Take the gradient with respect to θ:

$$\nabla_\theta L(\theta) = -2 \sum_{s=1}^{t} A_s \left(X_s - A_s^\top \theta \right) + 2\lambda\theta$$

- Set the gradient to zero:

$$\sum_{s=1}^{t} A_s A_s^\top \theta + \lambda\theta = \sum_{s=1}^{t} A_s X_s$$

- Group terms:

$$\left(\lambda I + \sum_{s=1}^{t} A_s A_s^\top \right) \theta = \sum_{s=1}^{t} A_s X_s$$

Stochastic linear bandit

- Define the design matrix:

$$V_t := \lambda I + \sum_{s=1}^{t} A_s A_s^\top \tag{19.6}$$

- Then the solution is:

$$\hat{\theta}_t = V_t^{-1} \sum_{s=1}^{t} A_s X_s \tag{19.5}$$

- **Note:** Some sources define the estimate as $\hat{\theta}_{t-1}$ to emphasize that data only up to time $t-1$ is used to choose action A_t.

$$\hat{\theta}_t - \theta_* = V_t^{-1} \sum_{s=1}^{t} A_s \eta_s - \lambda V_t^{-1} \theta_*$$

this is the error , we need to bound it.

Stochastic linear bandit

Stochastic linear bandit

$$\leq: \beta_{1,t} := \sqrt{2\log\left(\frac{\det(V_t)^{1/2}}{\delta \cdot \lambda^{d/2}}\right)}$$

Abbasi-Yadkouri(2011)

Stochastic linear bandit

$$\leq \lambda \, {\theta^*}^\top \theta^* = \lambda \left\| \theta^* \right\|_2^2$$

Stochastic linear bandit

$$\left\|\hat{\theta}_t - \theta^*\right\|_{V_t}^2 \leq \beta_t^2 := 2\log\left(\frac{\det(V_t)^{1/2}}{\delta \cdot \lambda^{d/2}}\right) + 2\lambda \left\|\theta^*\right\|_2^2$$

with prob. atleast 1-δ

$$\mathsf{V}_{t-1} = Q\Lambda Q^\top$$

$$\sum_{i=1}^{d} \left(\frac{\beta_t}{\lambda_i} z_i\right)^2 \leq 1$$

$$\mathsf{Vol}(E) = \frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)} \cdot \frac{\beta^{d/2}}{\sqrt{\det(M)}}$$

UCB using optimization over an ellipsoid

We are interested in computing:

$$\mathsf{UCB}_t(a) = \max_{\theta \in C_t} \langle \theta, a \rangle$$

where the confidence set is defined as:

$$C_t = \left\{ \theta \in \mathbb{R}^d : \left\| \theta - \hat{\theta}_t \right\|_{V_t} \leq \beta_t \right\}$$

This is equivalent to solving:

$$\max_{\theta \in \mathbb{R}^d} \langle \theta, a \rangle \quad \text{subject to} \quad (\theta - \hat{\theta}_t)^\top V_t (\theta - \hat{\theta}_t) \leq \beta_t^2$$

Stochastic linear bandit

Let $z := \theta - \hat{\theta}_t$, so that $\theta = \hat{\theta}_t + z$. Then:

$$\langle \theta, a \rangle = \langle \hat{\theta}_t + z, a \rangle = \langle \hat{\theta}_t, a \rangle + \langle z, a \rangle$$

The constraint becomes:

$$z^\top V_t z \leq \beta_t^2$$

Thus, the problem reduces to:

$$\max_{z \in \mathbb{R}^d} \langle z, a \rangle \quad \text{subject to} \quad z^\top V_t z \leq \beta_t^2$$

Stochastic linear bandit

This is a linearly constrained quadratic program. By the Cauchy-Schwarz inequality in the Mahalanobis norm:

$$\langle z, a \rangle \leq \|z\|_{V_t} \cdot \|a\|_{V_t^{-1}}$$

with equality when:

$$z = \alpha V_t^{-1} a \quad \text{for some scalar } \alpha > 0$$

To satisfy the constraint $\|z\|_{V_t} = \beta_t$, compute:

$$\|z\|_{V_t}^2 = z^\top V_t z = \alpha^2 a^\top V_t^{-1} a = \beta_t^2 \Rightarrow \alpha = \frac{\beta_t}{\|a\|_{V_t^{-1}}}$$

Thus, the optimal z is:

$$z^\star = \frac{\beta_t}{\|a\|_{V_t^{-1}}} V_t^{-1} a$$

Stochastic linear bandit

Substituting back:

$$\max_{\theta \in C_t} \langle \theta, a \rangle = \langle \hat{\theta}_t + z^\star, a \rangle = \langle \hat{\theta}_t, a \rangle + \langle z^\star, a \rangle$$

Compute:

$$\langle z^\star, a \rangle = \frac{\beta_t}{\|a\|_{V_t^{-1}}} a^\top V_t^{-1} a = \beta_t \cdot \|a\|_{V_t^{-1}}$$

Therefore:

$$\boxed{\mathsf{UCB}_t(a) = \langle \hat{\theta}_t, a \rangle + \beta_t \cdot \|a\|_{V_t^{-1}}}$$

Stochastic linear bandit

Stochastic linear bandit

Stochastic linear bandit

- Given a positive definite matrix V, the set:

$$\left\{x \in \mathbb{R}^d : (x-\mu)^\top V(x-\mu) \leq \beta\right\}$$

is an **ellipsoid** centered at μ.

- Let $V = Q\Lambda Q^\top$ be the eigendecomposition of V, where:
  - Q is an orthonormal matrix of eigenvectors $q_1, \dots, q_d$
  - $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_d)$ contains the eigenvalues
- Then the Mahalanobis norm becomes:

$$\|x-\mu\|_V^2 = (x-\mu)^\top Q\Lambda Q^\top (x-\mu) = \sum_{i=1}^{d} \lambda_i \langle q_i, x-\mu\rangle^2$$

- This shows:
  - The ellipsoid is stretched/scaled along directions q_i
  - Axis length in direction q_i is proportional to $\frac{1}{\sqrt{\lambda_i}}$
- Therefore:
  - As V_t increases in all directions (i.e., more data accumulated),
  - The ellipsoid shrinks — indicating improved certainty and better learning
  - Provided β_t doesn't grow too fast, the confidence set contracts

Continued

- **Why center at $\hat{\theta}_{t-1}$?**
  - $\hat{\theta}_{t-1}$ is the best guess of θ^* using data from rounds 1 to $t-1$
  - The confidence set C_t is built using this historical data
  - It allows us to apply the UCB principle:

    "With high probability, $\theta^* \in C_t$"

  - So we choose the action a that maximizes the most optimistic reward:

$$\mathrm{UCB}_t(a) = \max_{\theta \in C_t} \langle \theta, a \rangle$$

LinUCB

We assume the following:

- Feature map: $\psi : \mathcal{C} \times [k] \to \mathbb{R}^d$
- Regularization parameter: $\lambda > 0$
- Confidence radius β_t computed via concentration bounds
- Design matrix: $V_t \in \mathbb{R}^{d \times d}$
- Weighted feature-reward vector: $b_t \in \mathbb{R}^d$

Frame Title

$$V_0 = \lambda I_d, \quad b_0 = 0 \in \mathbb{R}^d$$

For each round $t = 1, 2, \ldots, n$:

1. **Observe context:** $C_t \in \mathcal{C}$
   Construct the set of feature vectors:
   $$\mathcal{A}_t = \{x_{t,1} = \psi(C_t, 1), \ldots, x_{t,k} = \psi(C_t, k)\} \subset \mathbb{R}^d$$
2. **Estimate the parameter:**
   $$\hat{\theta}_{t-1} = V_{t-1}^{-1} b_{t-1}$$
3. **Compute the UCB score for each arm** $a \in [k]$**:**
   $$\mathrm{UCB}_t(a) = \langle \hat{\theta}_{t-1}, x_{t,a} \rangle + \beta_t \cdot \|x_{t,a}\|_{V_{t-1}^{-1}}$$
4. **Select the action with highest UCB:**
   $$A_t = \arg\max_{a \in [k]} \mathrm{UCB}_t(a)$$

1. **Observe stochastic reward:**

$$X_t = \langle \theta^*, x_{t,A_t} \rangle + \eta_t, \quad \text{where } \eta_t \text{ is 1-subgaussian}$$

2. **Update:**

$$V_t = V_{t-1} + x_{t,A_t} x_{t,A_t}^\top$$

$$b_t = b_{t-1} + x_{t,A_t} X_t$$

With probability at least $1 - \delta$, the true parameter θ^* lies in the ellipsoid:

$$\|\hat{\theta}_{t-1} - \theta^*\|_{V_{t-1}} \leq \beta_t$$

A typical choice for β_t (assuming $\|\theta^*\|_2 \leq S$) is:

$$\beta_t = \sqrt{\lambda} S + \sqrt{2 \log \left(\frac{1}{\delta} \cdot \frac{\det(V_t)^{1/2}}{\lambda^{d/2}} \right)}$$

Frame Title

Frame Title

Frame Title

Frame Title

Frame Title

Frame Title