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.

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 $\psi(c, a)$.
- 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:

$$\Omega(nMk)$$

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

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

- **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:

$$C_t = \left\{ \theta \in \mathbb{R}^d : \left\| \theta - \hat{\theta}_{t-1} \right\|_{V_{t-1}} \leq \beta_t \right\}$$

- This is an ellipsoid:
  - Centered at the current estimate $\hat{\theta}_{t-1}$.
  - Shaped by the matrix V_{t-1}.
  - Radius β_t is chosen based on subgaussian concentration bounds.
- We will derive and prove this construction rigorously in the next slides.

Stochastic linear bandit

- **Proof/Derivations (Preview)**:
- To build the confidence set C_t, we proceed as follows:
- Use **regularized least squares** to estimate θ^*:

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

- Define the **design matrix** (regularized Gram matrix):

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

- Then, using **self-normalized martingale inequalities**, we can show that with high probability:

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

- This directly yields the **confidence ellipsoid**:

$$C_t = \left\{\theta \in \mathbb{R}^d : \left\|\theta - \hat{\theta}_{t-1}\right\|_{V_{t-1}} \leq \beta_t\right\}$$

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.

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 $q_1, \ldots, q_d$
  - $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_d)$ 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

Let $\lambda > 0$ be a regularization parameter, and let β_t be a confidence radius computed via concentration inequalities. Let d denote the dimension of the feature space.

$$V_0 \leftarrow \lambda I_d$$
$$b_0 \leftarrow 0 \in \mathbb{R}^d$$

1. Observe context $C_t \in \mathcal{C}$ and corresponding action set:
$$\mathcal{A}_t = \{\psi(C_t, 1), \dots, \psi(C_t, k)\} \subset \mathbb{R}^d$$
2. Compute the regularized least squares estimate:
$$\hat{\theta}_{t-1} = V_{t-1}^{-1} b_{t-1}$$
3. For each $a \in \mathcal{A}_t$, compute:
$$\mathrm{UCB}_t(a) = \langle \hat{\theta}_{t-1}, a \rangle + \beta_t \cdot \|a\|_{V_{t-1}^{-1}}$$

1. Select action:
$$A_t \in \arg\max_{a \in \mathcal{A}_t} \mathrm{UCB}_t(a)$$
2. Observe reward:
$$X_t = \langle \theta^*, A_t \rangle + \eta_t$$
3. Update:
$$V_t \leftarrow V_{t-1} + A_t A_t^\top$$
$$b_t \leftarrow b_{t-1} + A_t X_t$$

With probability at least $1 - \delta$, the following inequality holds for all t:
$$\|\hat{\theta}_{t-1} - \theta^*\|_{V_{t-1}} \leq \beta_t$$
where:
$$\beta_t = \sqrt{\lambda} S + \sqrt{2 \log \left(\frac{\det(V_t)^{1/2}}{\delta \cdot \lambda^{d/2}} \right)}$$
assuming $\|\theta^*\|_2 \leq S$.

Frame Title

Symbol	Meaning
$\psi(C_t, a)$	Feature vector for context-action pair (C_t, a)
V_t	Regularized design matrix
b_t	Accumulated response-weighted features
$\hat{\theta}_t$	Ridge regression estimate of θ^*
β_t	Confidence radius
$\|x\|_V^2$	Mahalanobis norm: $x^\top V x$
$\mathrm{UCB}_t(a)$	Optimistic reward estimate
A_t	Selected action at round t