Linear Bandits

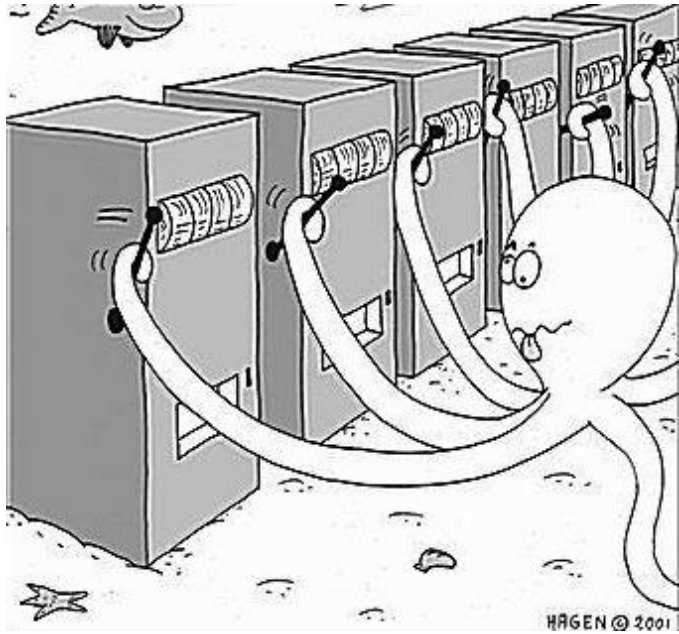
Principles of Online Decision-Making (CS-303)

Prof. Matthias Grossglauser

Information and Network Dynamics (INDY) lab
School of Computer and Communication Sciences (I&C)
EPFL

Recap: Multi-armed bandits

- A framework for online decision-making
- Given: k possible actions (arms)
- Play one arm at each round: $A_t \in [k]$, get reward X_t
- Goal: maximise cumulative reward: $\sum_{t=1}^{T} X_t$



- Each arm has its own reward distribution: $X_t \sim p_i$; $\mathbb{E}[X_t] = \mu_i$
- Challenge: initially, no information about rewards: μ_i unknown
- Must balance exploration (learning μ_i) and exploitation (maximising X_t)

Recap: the UCB algorithm

Notation

- Estimate of mean reward of arm i at time t: $\hat{\mu}_i(t)$
- Number of pulls of arm i at time t: $T_i(t)$
- Upper confidence bound of arm i at time t:

$$UCB_i(t,\delta) = \hat{\mu}_i(t) + \sqrt{\frac{2\log(1/\delta)}{T_i(t)}}$$

exploitation

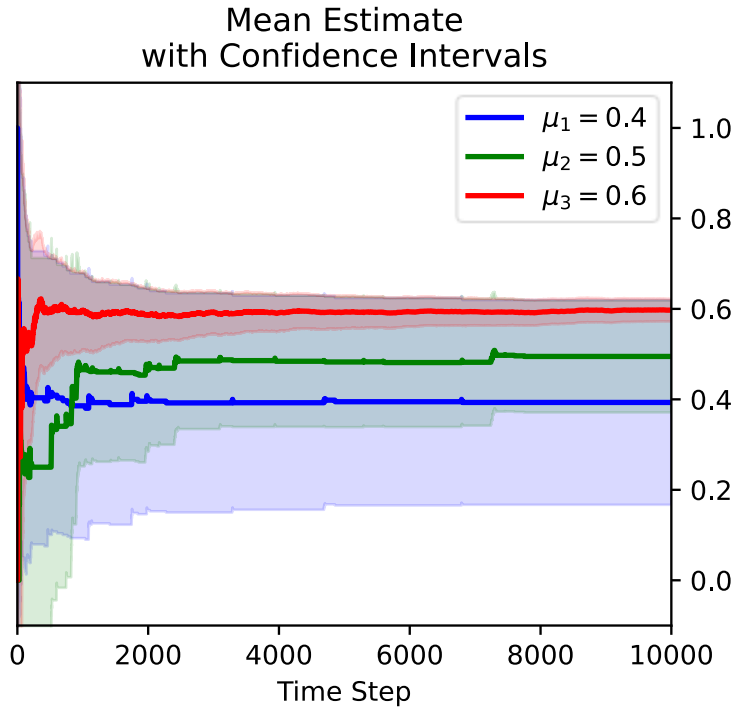
exploration

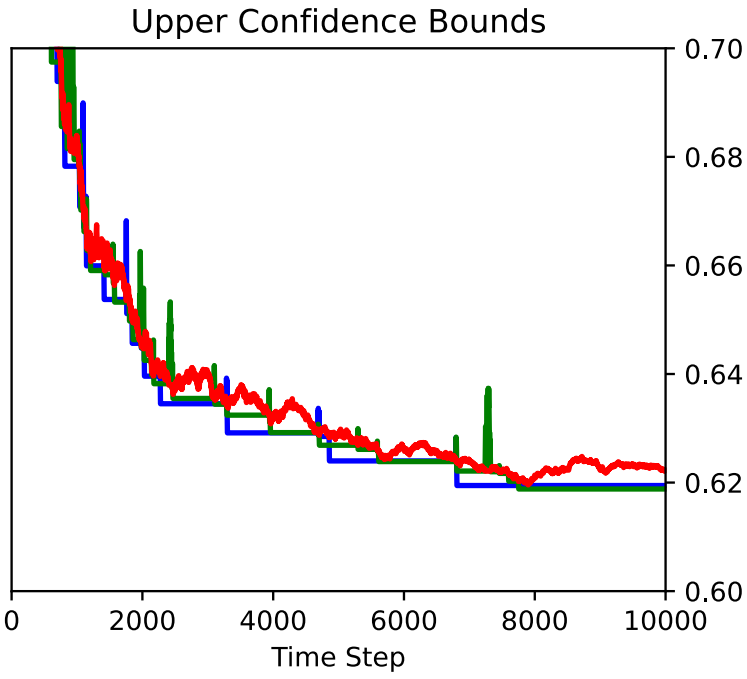
UCB Algorithm

For $t = 1, 2, \ldots, n$ do:

Choose $A_t = \arg\max_i UCB_i(t-1, \delta)$

Observe X_t and update $UCB_i(t,\delta)$

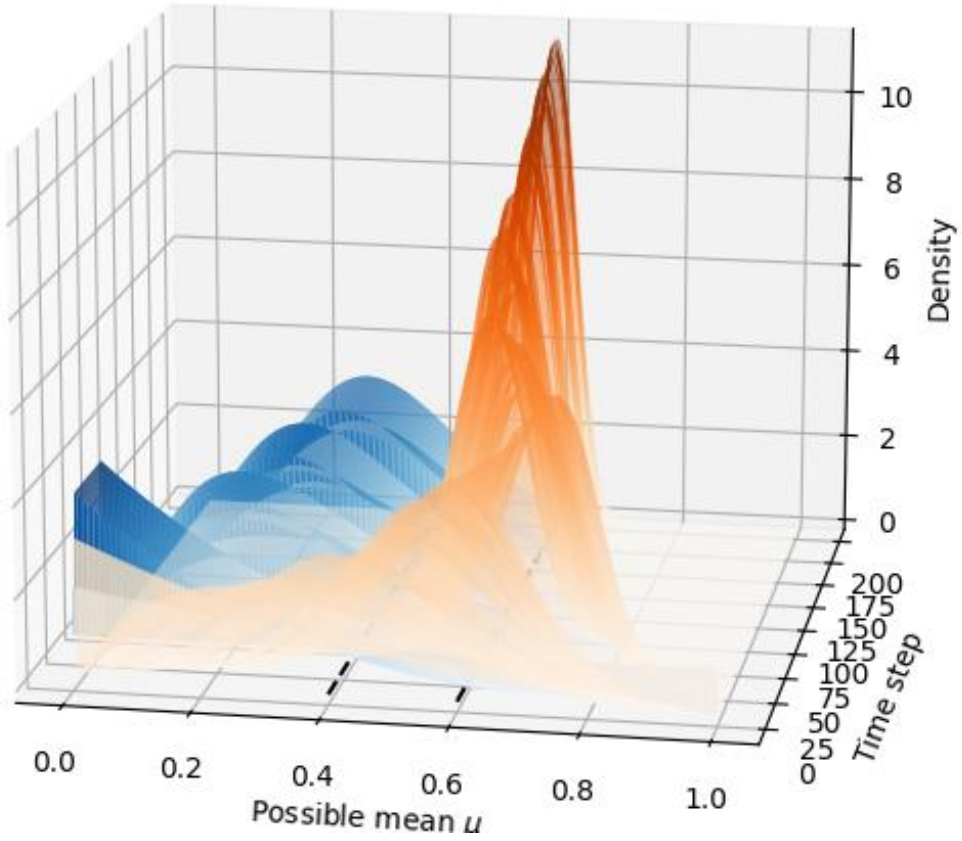




Recap: Thompson Sampling

Bayesian algorithm

- Maintain belief distribution $f_{i,t}(\cdot)$ for all arms
- Exploration happens through randomness in the sampling
- Over time, beliefs concentrate → exploitation



Thompson Sampling

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

  For $i = 1, 2, \ldots, k$ do:

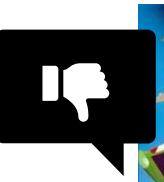
    sample $\tilde{\mu}_i(t) \sim f_{i,t}(\cdot)$

  Choose $A_t = \arg\max_i \left(\tilde{\mu}_i(t)\right)$

  Observe X_t, update $f_{A_t,t}(\cdot)$

Recap: regret bounds for multi-armed bandits

- Problem dependent bound: $R_n \lesssim \sum_{i:\Delta_i>0} \frac{16 \ln(n)}{\Delta_i}$
- Problem independent bound: $R_n \leq 8\sqrt{kn \ln n}$
- Regret grows with no. of arms: <u>cost of exploring each option</u>
- Information structure: Pulling one arm tells *nothing* about other arms
- Not always true in practice!



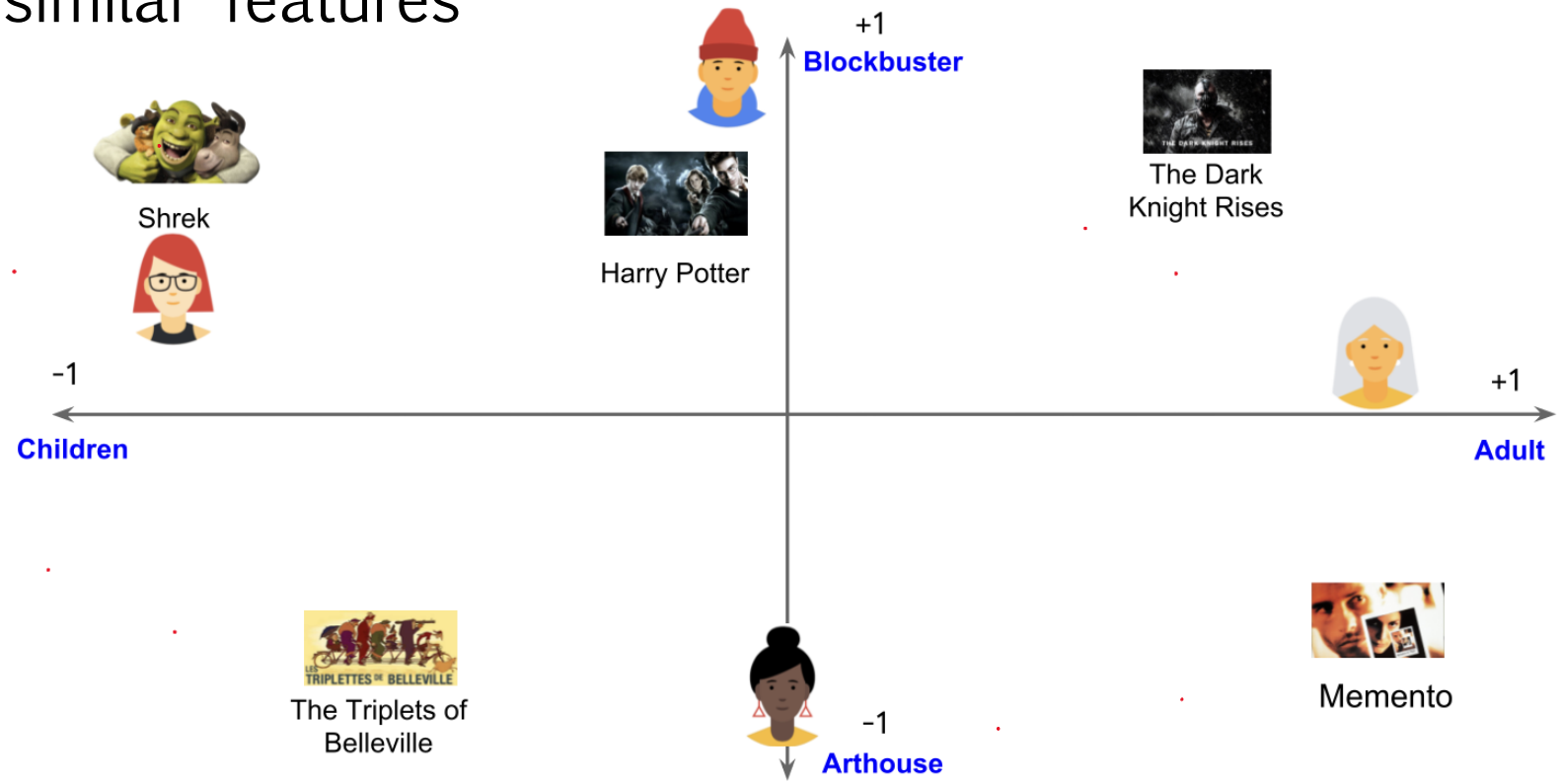
Netflix Scenario

a person does not like **Toy Story** and **Ratatouille**
➔ they probably don't like children movies
➔ they probably don't like **Finding Nemo**
➔ not worth recommending it (and regretting it!)

Modeling recommender systems

Consider a recommender system like Netflix

- items (movies) have features: genre, language, actor, director, ...
- similar movies ⇒ similar features



[Figure courtesy:
Google for Developers:
Recommendation Systems]

- Users have features too: preference over genres, languages, actors, etc
- Similar tastes ⇒ similar features

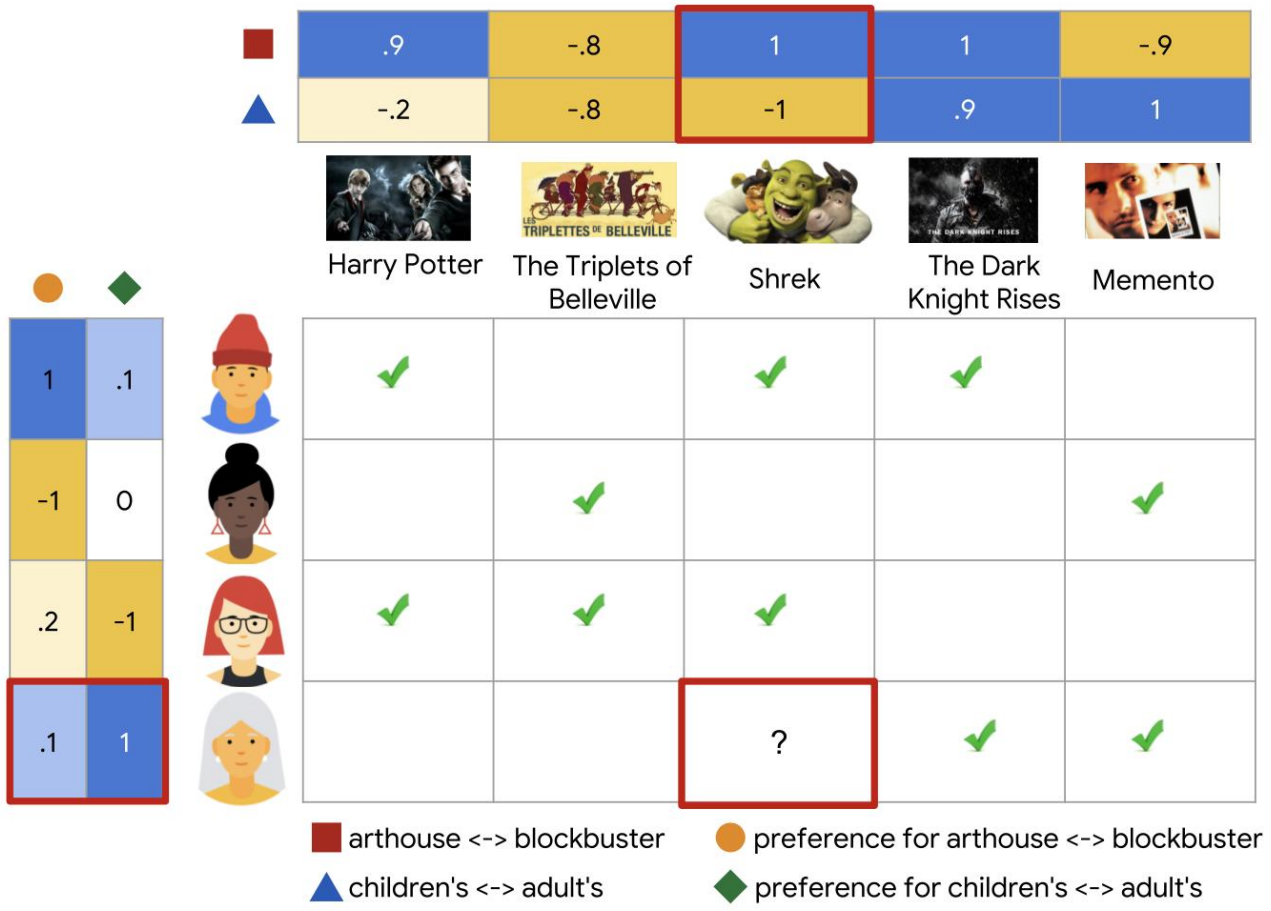
Modeling recommender systems

- User feature vector $\theta^* \in \mathbb{R}^d$
- Item feature vector $a \in \mathbb{R}^d$

- **Utility**: $\sum_i a_i \theta_i = a^\top \theta$
- Sum over features of a_i
- (users' preference for feature)

- **Reward**: $a^\top \theta + \varepsilon$ or $g(a^\top \theta) + \varepsilon$
  - Continuous: watch time
  - Discrete: rating given to movie
  - Binary: watched/not watched

- Reward is a noisy signal of utility

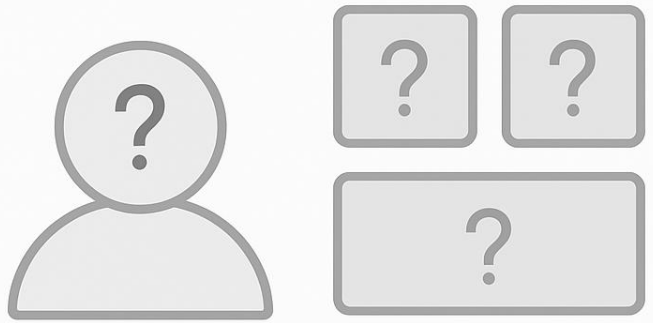


[Figure courtesy: Google for Developers: Recommendation Systems]

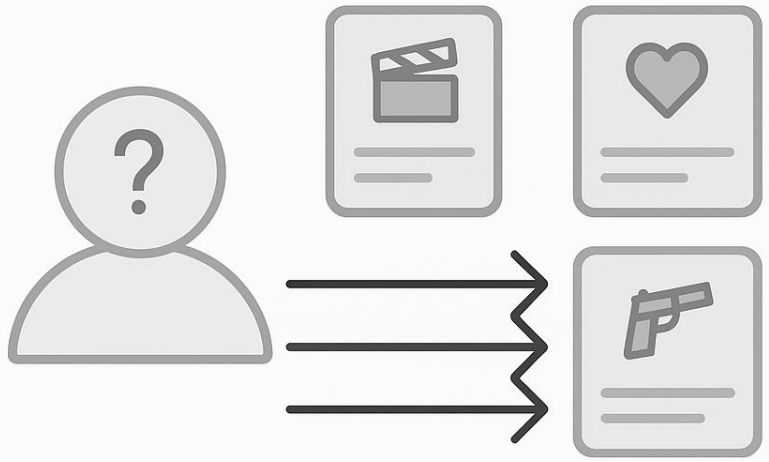
Modeling assumptions for recommender systems

1. Both features unknown:
- Collaborative filtering, matrix factorization
- Identify similar users, similar items
- Use preferences of similar users to recommend new items
- Advantage: no need to handcraft features
- Cold start problem: each new user needs to rate few items to get good recommendations; analogous for new items

2. Only user features unknown:
- Recommend diverse movies to start with
- With little data, can identify preferences: likes comedy, dislikes action, ...
- Use data to learn user features



Cold start problem



Use content features to bootstrap

Recommender systems as bandits

Suppose a new user signs up on Netflix

- Goal: recommend the best movies for this user over time
  - Maximise cumulative reward ↔ minimize cumulative regret
- Subgoal: learn the user preferences
  - Need to try out different movies
  - At the cost of showing a few bad movies
- Like multi-armed bandits:
  - Online learning problem, exploration-exploitation tradeoff
- Unlike multi-armed bandits:
  - Reward from one arm (recommending a movie) gives information about other arms!

The linear bandit model

- Arms represented as a set of vectors $\mathcal{A} = \{a_1, a_2, \dots, a_k\}$; $a_i \in \mathbb{R}^d$
- Arm played at time t: $A_t \in \mathcal{A}$
- Reward at time t: $X_t = A_t^\top \theta^* + \eta_t$
- Unknown parameter to be estimated: $\theta^* \in \mathbb{R}^d$
- Noise model η_t: independent for all t, σ-subgaussian
- Best arm: $a^* = \arg\max_{a \in \mathcal{A}} a^\top \theta^*$
- Instantaneous regret: $r_t = a^{*\top} \theta^* - A_t^\top \theta^* = (a^* - A_t)^\top \theta^*$

Reduction to MAB

$\{a_1, a_2, \dots, a_d\}$: orthonormal basis

$a_i = [0, \dots, 0, 1, 0 \dots, 0] \Rightarrow$

$X_t = a_i^\top \theta^* + \eta_t = \theta_i^* + \eta_t$

Flexibility of linear bandit model

- Arm features can include known aspects of user context
  - Age, gender, region
  - Explicitly mentioned interests
  - Behavioural stats (# of comedy movies watched)

- The action set $\mathcal{A}_t = \{a_1^t, a_2^t, \ldots, a_k^t\}$ can vary with time, e.g.
  - Same user's context can change (restaurants in a new city; recommending among some search results;...)
  - Different users at different times
  - New items appear, old ones disappear

- Interpreting θ^*: global weight vector over features
  - $r_t = A_t^\top \theta^*$
  - Positive θ_i^* for (young, animation) feature

  $\Rightarrow$ young people prefer animation movies

Linear bandit model extends easily to time-varying, arbitrary action set $\mathcal{A}_t$

Formulations of regret for linear bandits

- Instantaneous regret: $r_t = a^{*\top}\theta^* - A_t^\top\theta^* = (a^* - A_t)^\top\theta^*$
  - Note: does not depend on noise η_t
- Random cumulative regret: $\hat{R}_n(A_1, \ldots, A_t) = \sum_{t=1}^n \max_{a\in\mathcal{A}_t}\langle\theta_*, a - A_t\rangle = \sum_{t=1}^n r_t$
  - Depends on the choice of actions
- Regret: $R_n = \mathbb{E}[\hat{R}_n] = \mathbb{E}\left[\sum_{t=1}^n \max_{a\in\mathcal{A}_t}\langle\theta_*, a\rangle - \sum_{t=1}^n X_t\right]$
  - Taking the expectation → choice of actions marginalized out

Estimating θ^* via linear regression

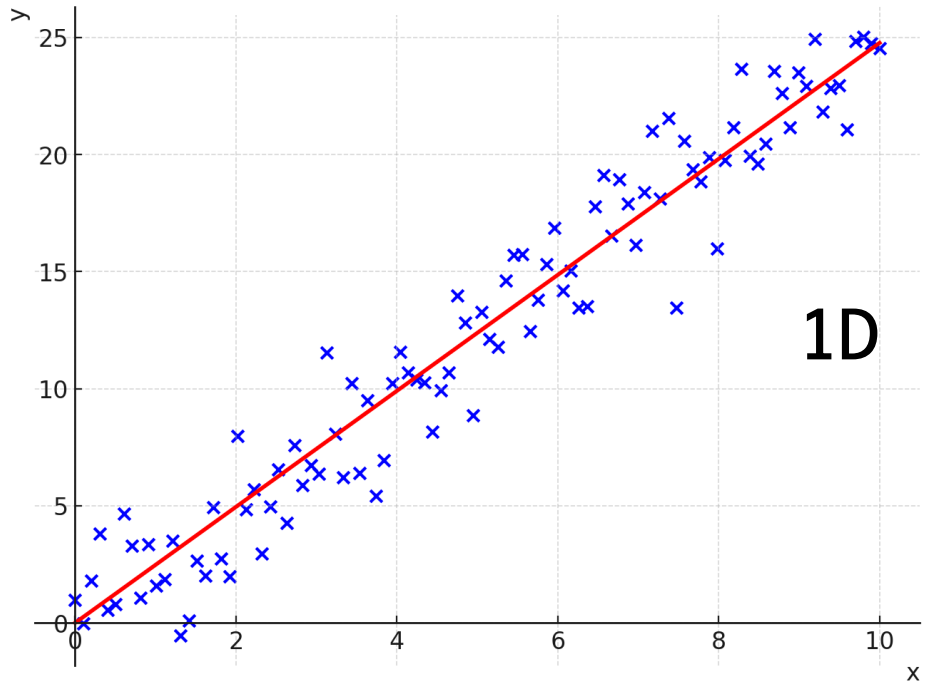
- Data: $(A_1, X_1, A_2, X_2, \ldots, A_t, X_t)$
- Model: $X_t = A_t^\top \theta^* + \eta_t$
- $\theta^* \in \mathbb{R}^d$ unknown, to be estimated
- $\eta_t \in \mathbb{R}$: 1-subgaussian random variable

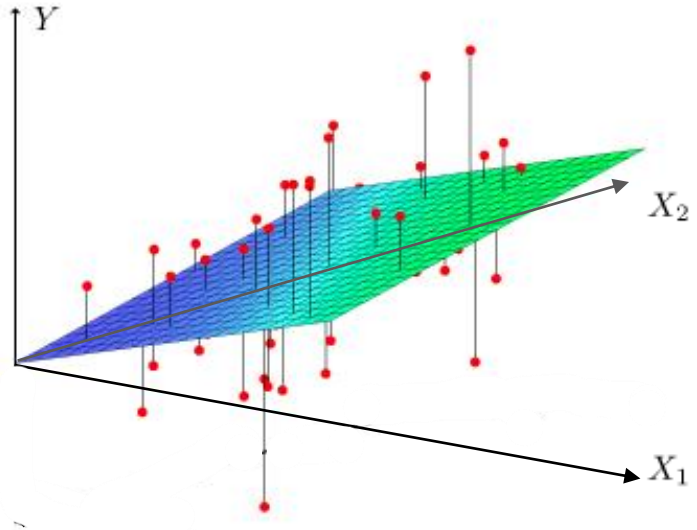
- Known as **linear regression**
- Least squares problem:

$$\min_\theta \sum_{i \in [t]} \left(X_i - A_i^\top \theta\right)^2$$

- Solution from calculus:

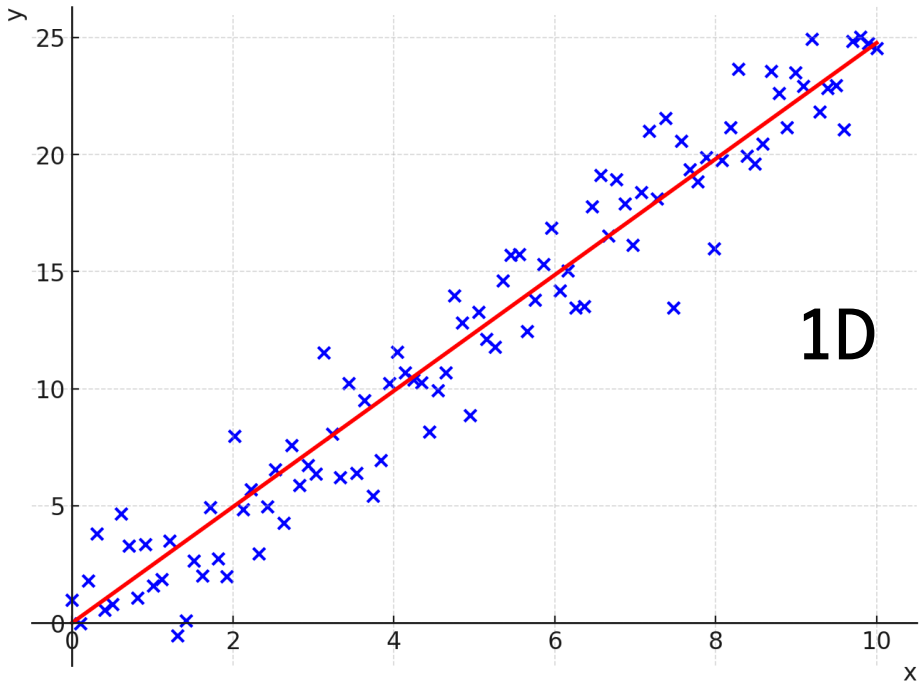
$$\hat{\theta}_t = V_t^{-1} \left(\sum_{i \in [t]} A_i X_i \right); \quad V_t = \sum_{i \in [t]} A_i A_i^\top$$

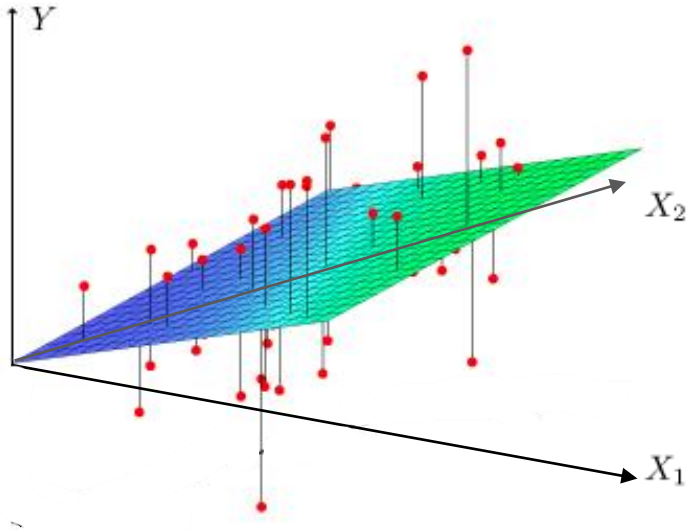




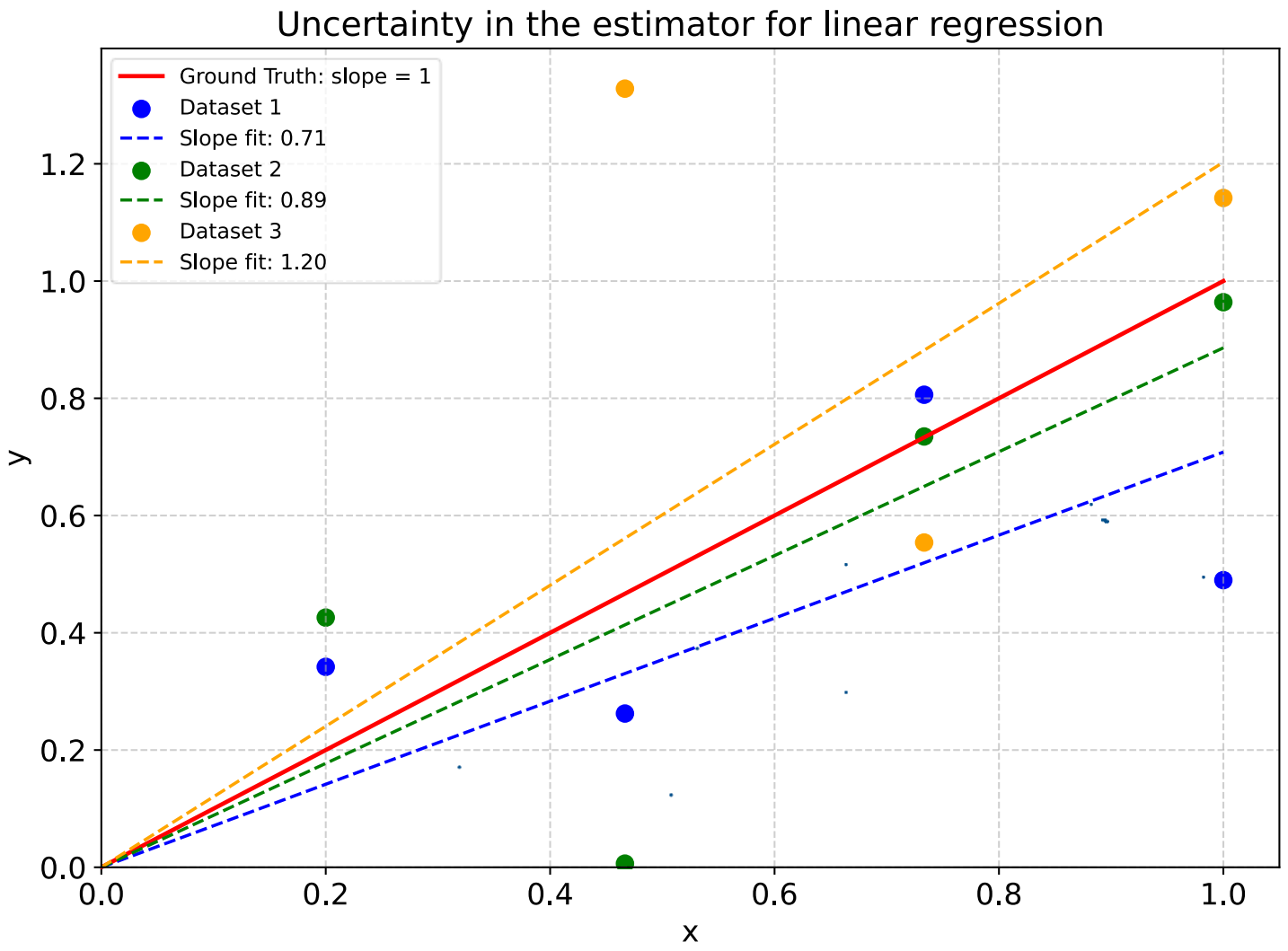
Estimating θ^* via linear regression

- $\theta^* \in \mathbb{R}^d$ unknown, to be estimated
- $\eta_t \in \mathbb{R}$: 1-subgaussian random variable
- $\hat{\theta}_t = V_t^{-1}\left(\sum_{i\in[t]} A_i X_i\right); \quad V_t = \sum_{i\in[t]} A_i A_i^{\top}$
- $\mathbb{E}[\hat{\theta}_t] = \mathbb{E}\left[V_t^{-1}\left[\underbrace{\sum_{i\in[t]} A_i A_i^{T}}_{V_t}\,\theta^* + \underbrace{A_i\eta_i}_{0}\right]\right] = \theta^*$





Uncertainty in linear regression: 1D case



Here, θ^* and $\hat{\theta}$ are scalars

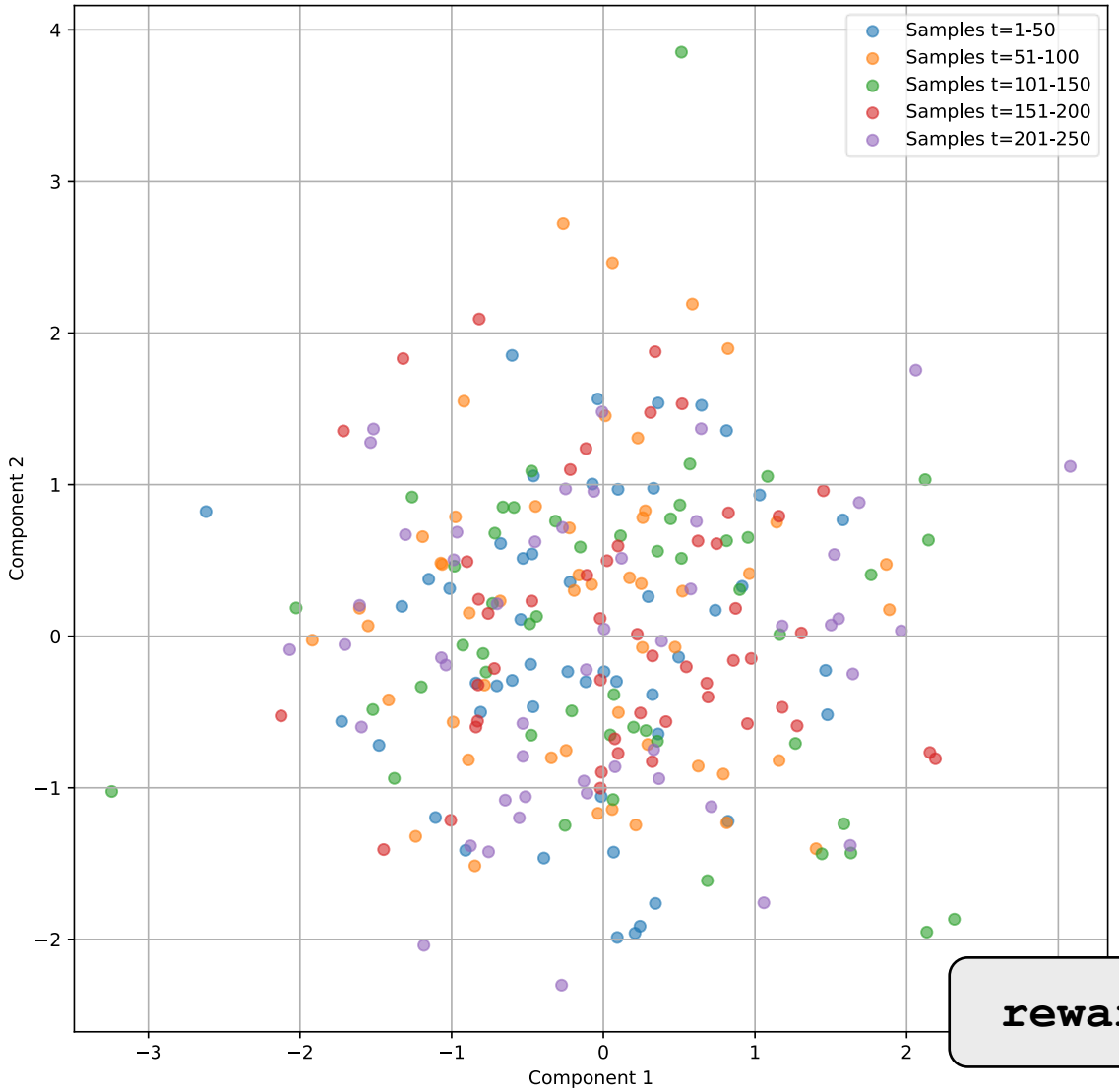
Uncertainty in $\hat{\theta}$ can be conveyed through confidence intervals

More noise in the data ⇒ larger confidence intervals

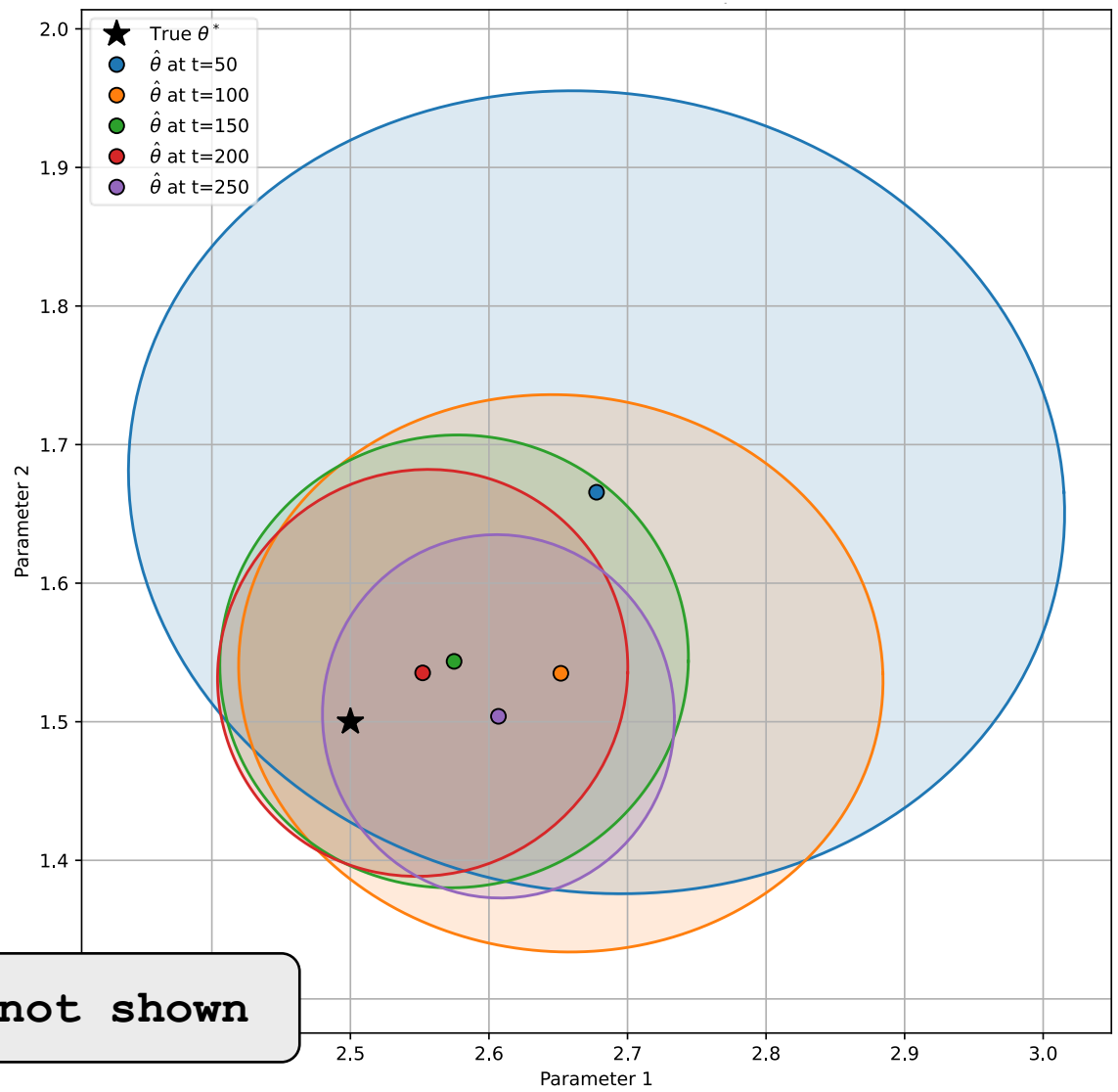
More data points ⇒ smaller confidence intervals

Uncertainty in linear regression: 2D case

Distribution of Points A_t



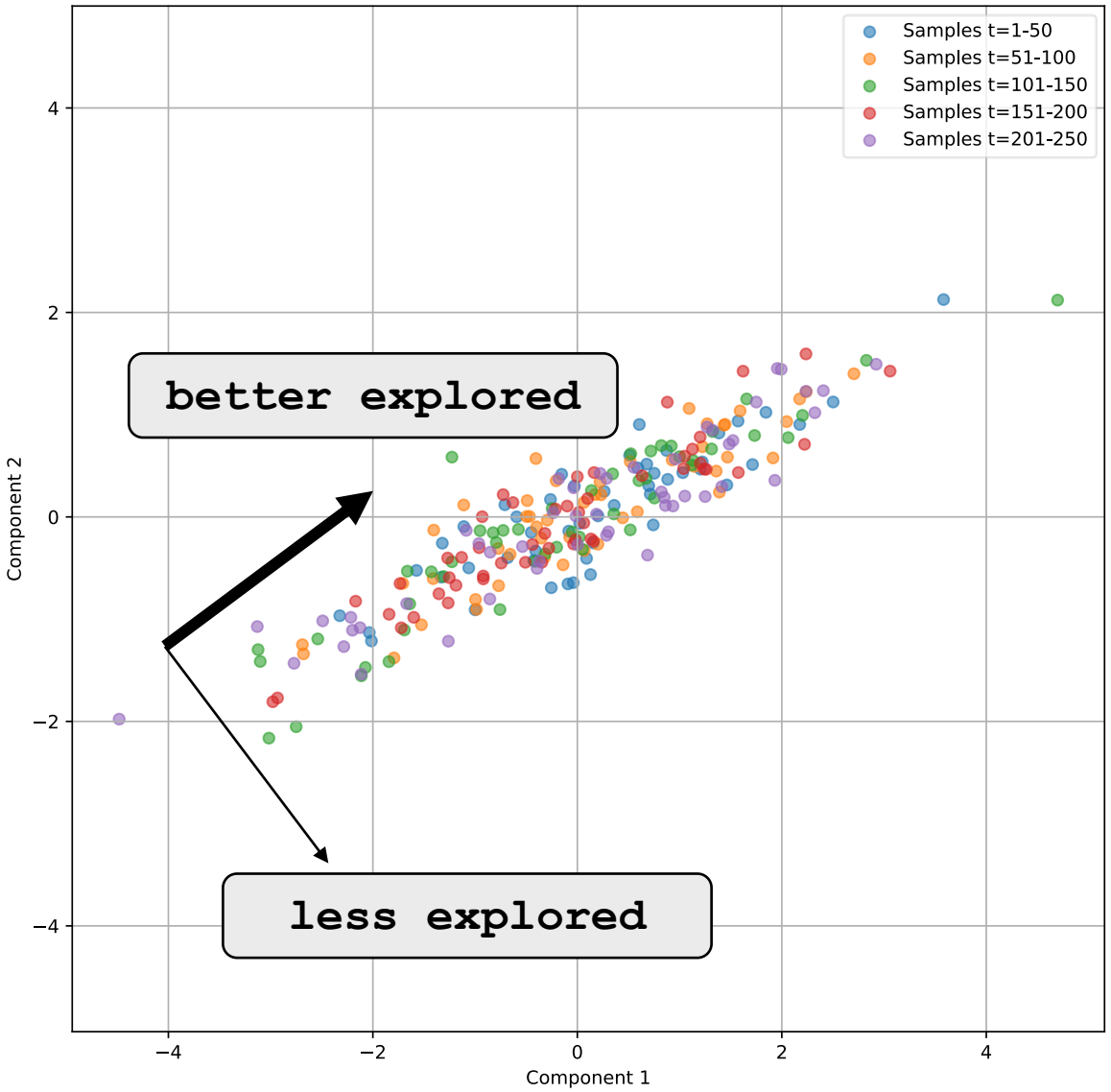
Confidence Ellipses $\mathcal{C}_t$ centred at $\hat{\theta}_t$



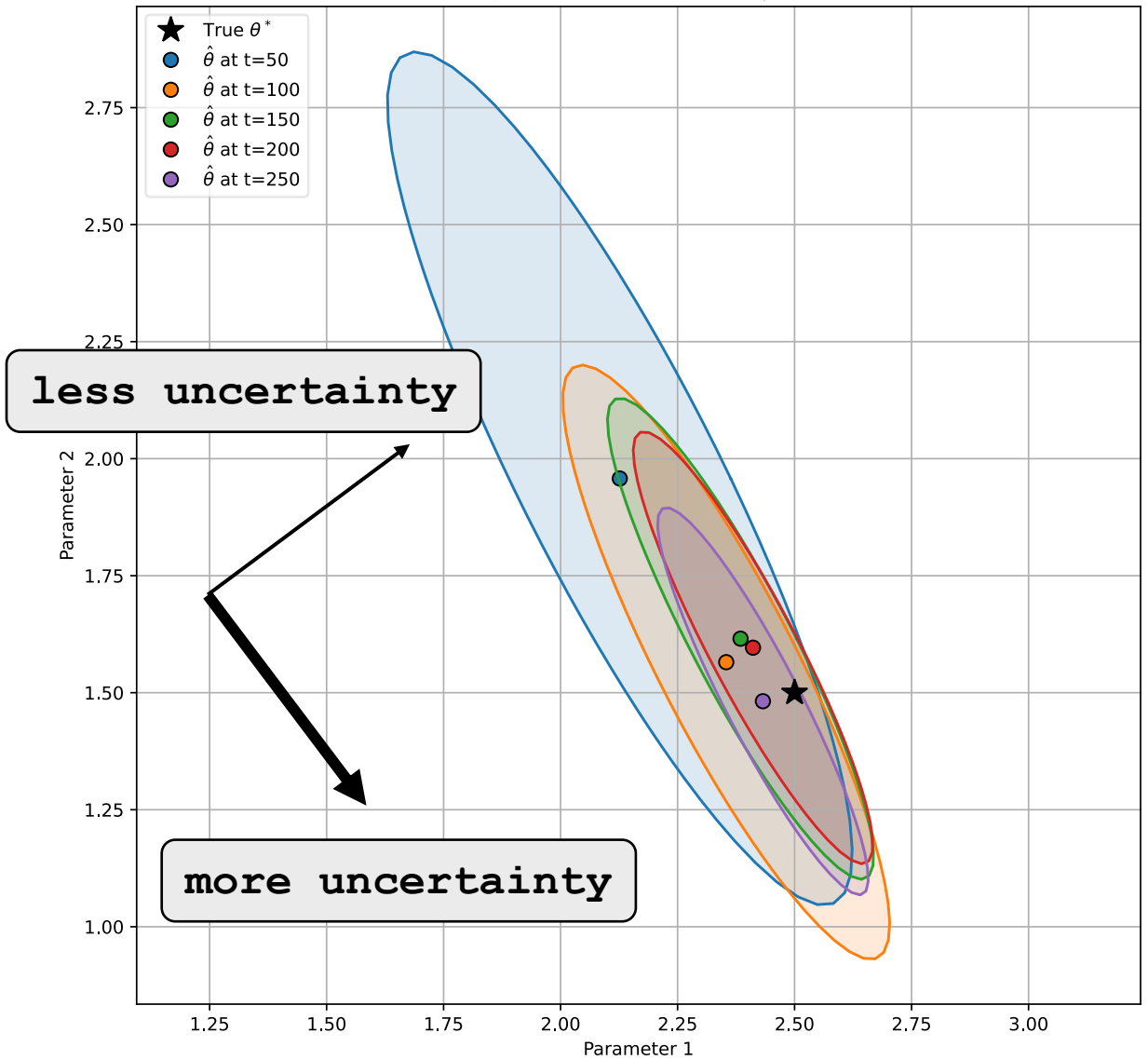
rewards not shown

Uncertainty in linear regression: 2D case

Distribution of Points A_t



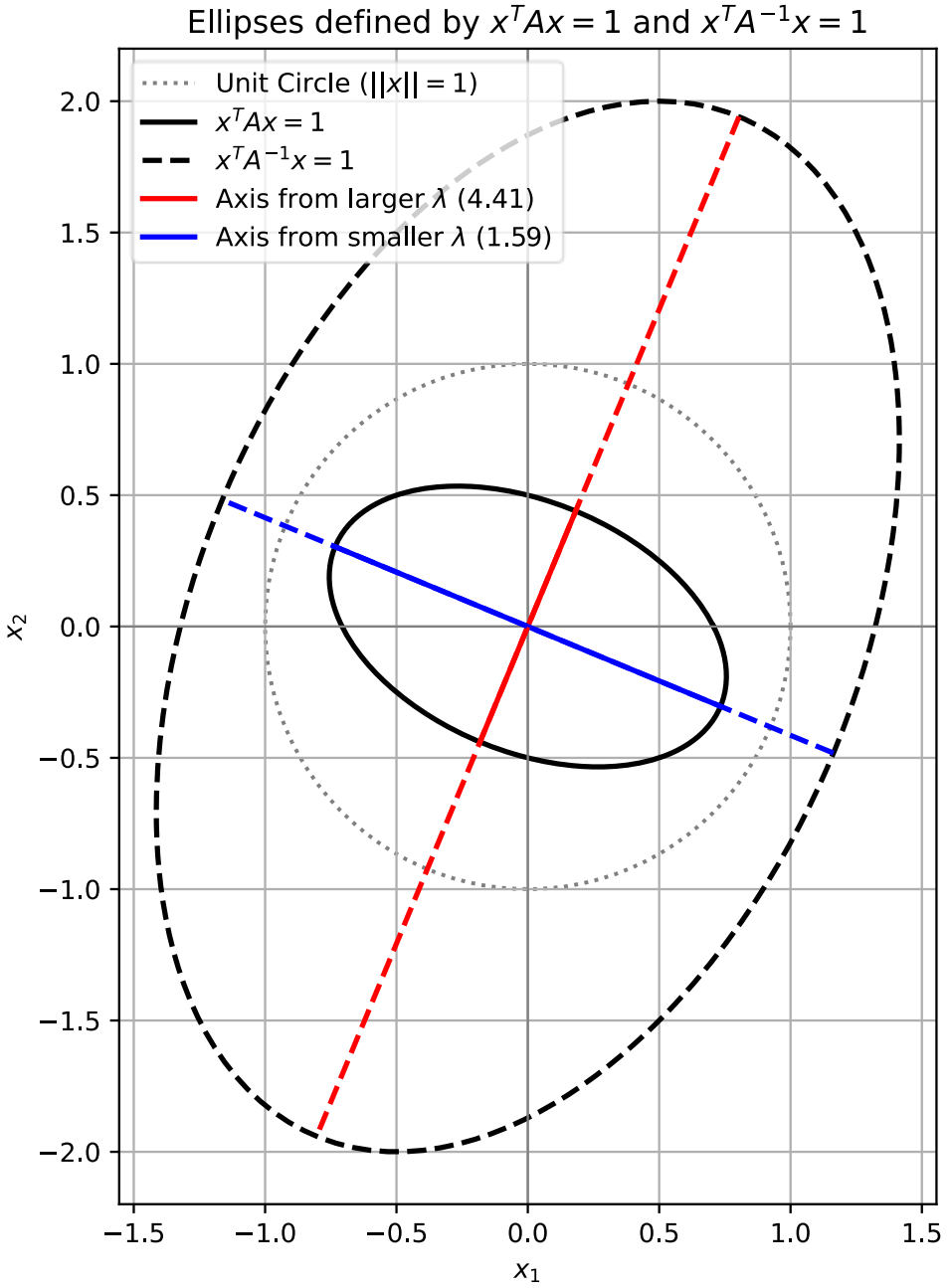
Confidence Ellipses $\mathcal{C}_t$ centred at $\hat{\theta}_t$



Why ellipses?

- The collection of data points represented by design matrix $V_t = \sum_{i \in [t]} A_i A_i^{\top}$
- More data points in a certain direction a $\Rightarrow$ eigenvalues of V_t larger in that direction
- More data points in a certain direction a $\Rightarrow$ better estimate of θ^* parallel to a
- Uncertainty of θ^* quantified by V_t^{-1} $\Rightarrow$ more uncertainty along larger eigenvalues
- Confidence sets:

$$\mathcal{C}_t = \left\{ \theta \in \mathbb{R}^d : \left\| \theta - \hat{\theta}_{t-1} \right\|_{V_{t-1}}^2 \leq \beta_t \right\}$$



Regularisation in linear regression

- Least squares problem:

$$\min_{\theta} \sum_{i \in [t]} \left(X_i - A_i^\top \theta\right)^2$$

- Solution from calculus:

$$\hat{\theta}_t = V_t^{-1} \left(\sum_{i \in [t]} A_i X_i \right); \quad V_t = \sum_{i \in [t]} A_i A_i^\top$$

- What if V_t is not invertible?
- Simple fix: use a regulariser. Solve:

$$\min_{\theta} \sum_{i \in [t]} \left(X_i - A_i^\top \theta\right)^2 + \lambda \|\theta\|^2$$

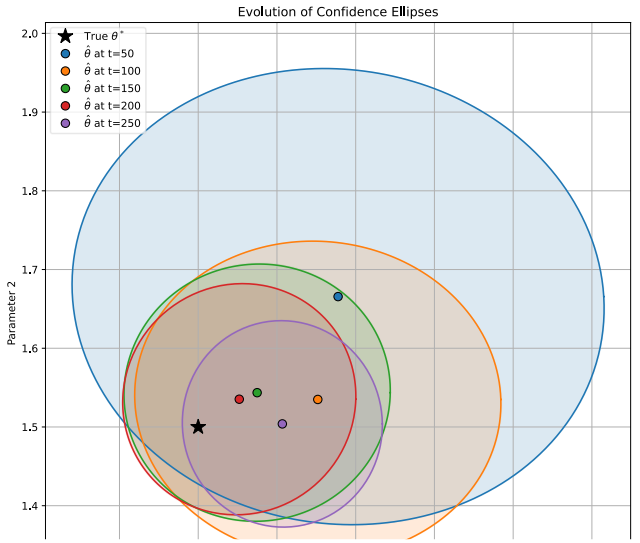
$$\hat{\theta}_t = V_t^{-1} \left(\sum_{i \in [t]} A_i X_i \right); \quad V_t = \lambda I + \sum_{i \in [t]} A_i A_i^\top$$

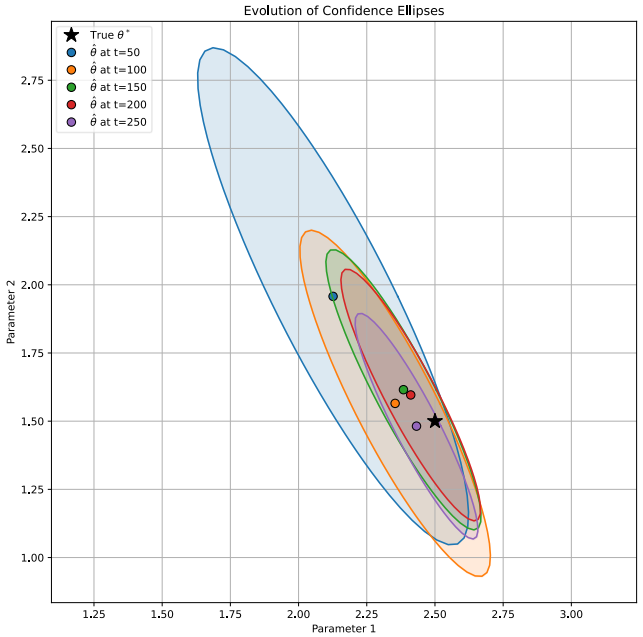
Impact of regulariser on confidence ellipses

isotropic

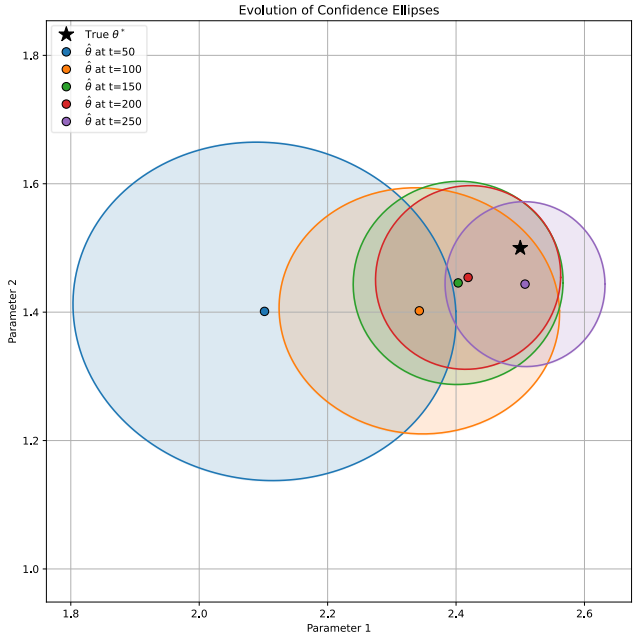
non-isotropic

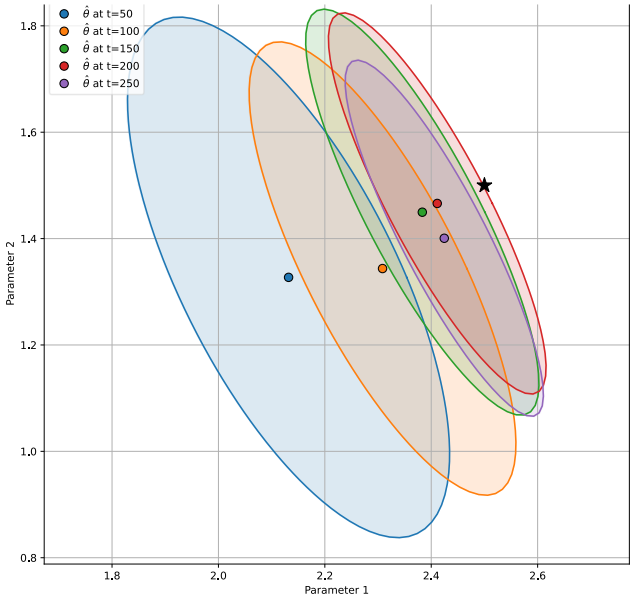
Without regulariser





With regulariser

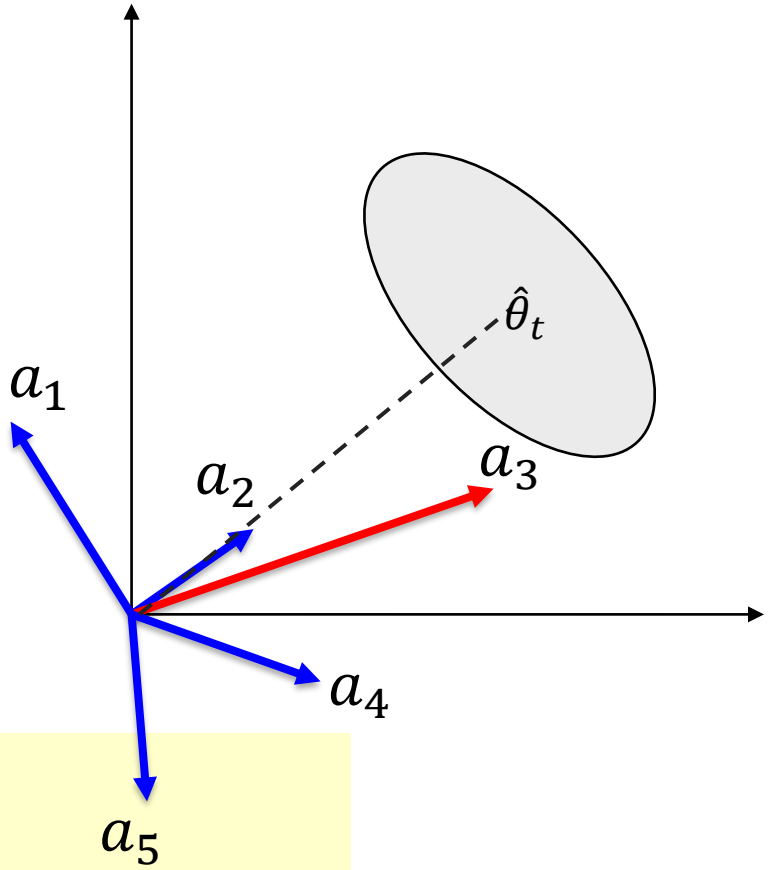




UCB for linear bandits

Notation:

- $\mathcal{C}_t = \left\{\theta \in \mathbb{R}^d : \left\|\theta - \hat{\theta}_{t-1}\right\|^2_{V_{t-1}} \leq \beta_{t-1}\right\}$
- $V_0 = \lambda I,\ V_t = \lambda I + \sum_{s\in[t]} A_s A_s^\top$
- $\hat{\theta}_t = V_t^{-1}\left(\sum_{s\in[t]} A_s X_s\right)$
- $1 \leq \beta_1 \leq \beta_2 \leq \cdots \leq \beta_n$



LinUCB Algorithm

For $t = 1, 2, \ldots, n$ do:

For all arms $a \in \mathcal{A}_t$

Compute $UCB_t(a) = \max_{\theta\in\mathcal{C}_t} a^\top\theta$

Play $A_t = \arg\max_{a\in\mathcal{A}_t} UCB_t(a)$

Observe X_t and update $V_t,\ \hat{\theta}_t,\ \mathcal{C}_t$

UCB computation for linear bandits

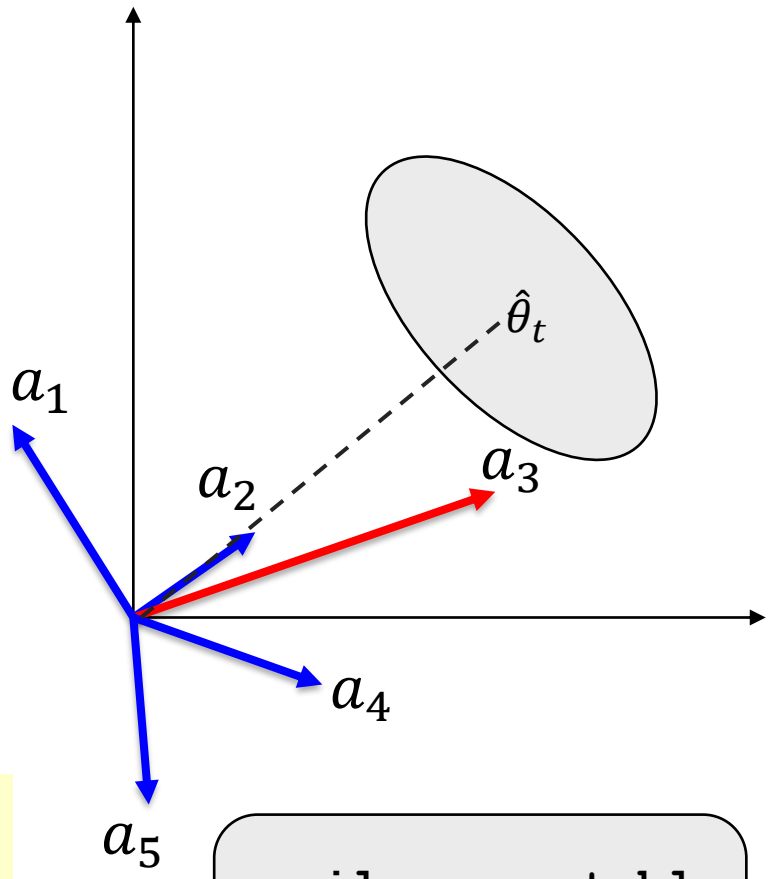
$$A_t = \arg\max_{a \in \mathcal{A}_t} UCB_t(a)\,;\; UCB_t(a) = \max_{\theta \in \mathcal{C}_t} a^\top \theta$$

$$\Rightarrow (A_t, \tilde{\theta}_t) = \arg\max_{a,\theta \in (\mathcal{A}_t, \mathcal{C}_t)} a^\top \theta$$

$$\Rightarrow A_t = \arg\max_{a \in \mathcal{A}_t} a^\top \hat{\theta}_{t-1} + \sqrt{\beta_t}\, \|a\|_{V_{t-1}^{-1}} \quad \text{IF}$$

ellipsoid →

$$\mathcal{C}_t = \left\{\theta \in \mathbb{R}^d : \left\|\theta - \hat{\theta}_{t-1}\right\|_{V_{t-1}}^2 \le \beta_{t-1}\right\}$$



easily computable individually for each arm $a \in \mathcal{A}_t$

LinUCB Algorithm

For $t = 1, 2, \ldots, n$ do:

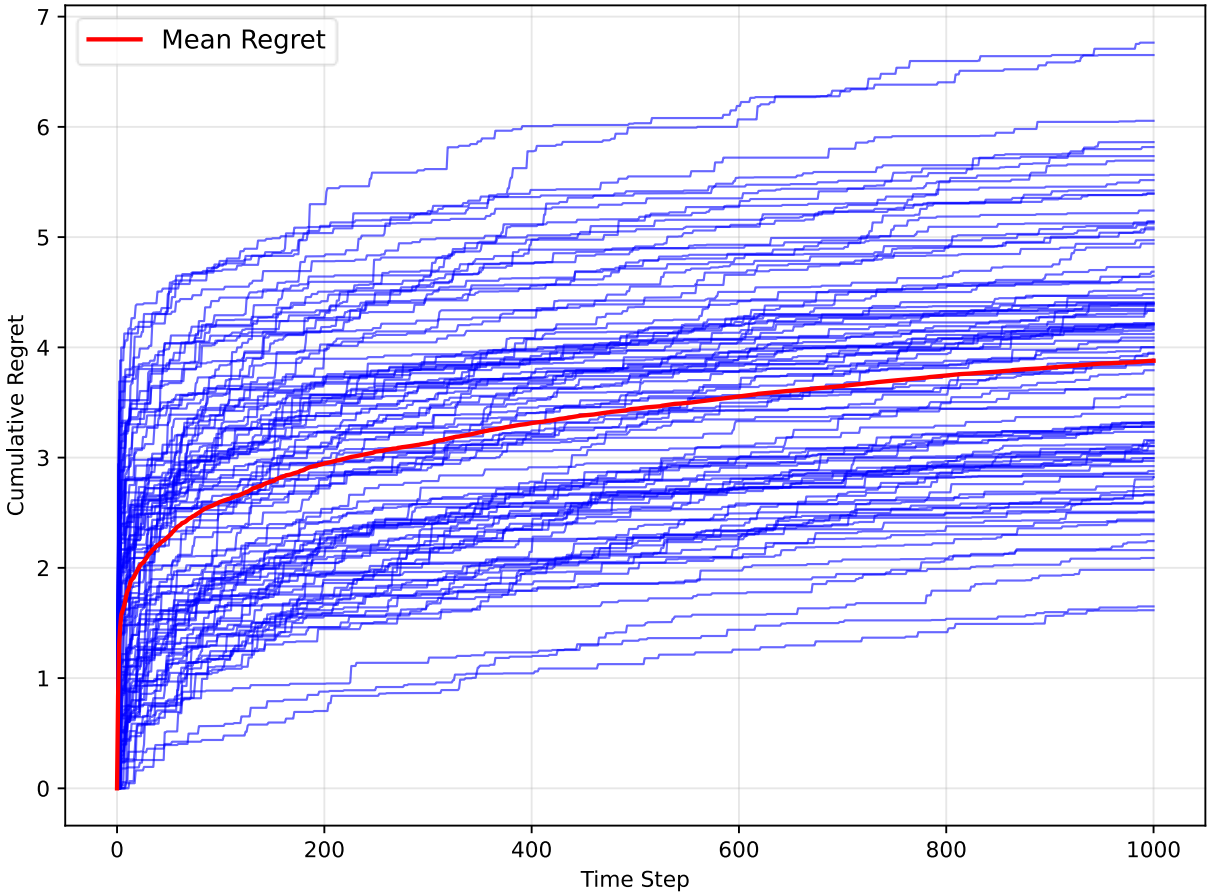
Play $A_t = \arg\max_{a \in \mathcal{A}_t} a^\top \hat{\theta}_{t-1} + \sqrt{\beta_t}\, \|a\|_{V_{t-1}^{-1}}$

Observe X_t and update V_t, $\hat{\theta}_t$, $\mathcal{C}_t$

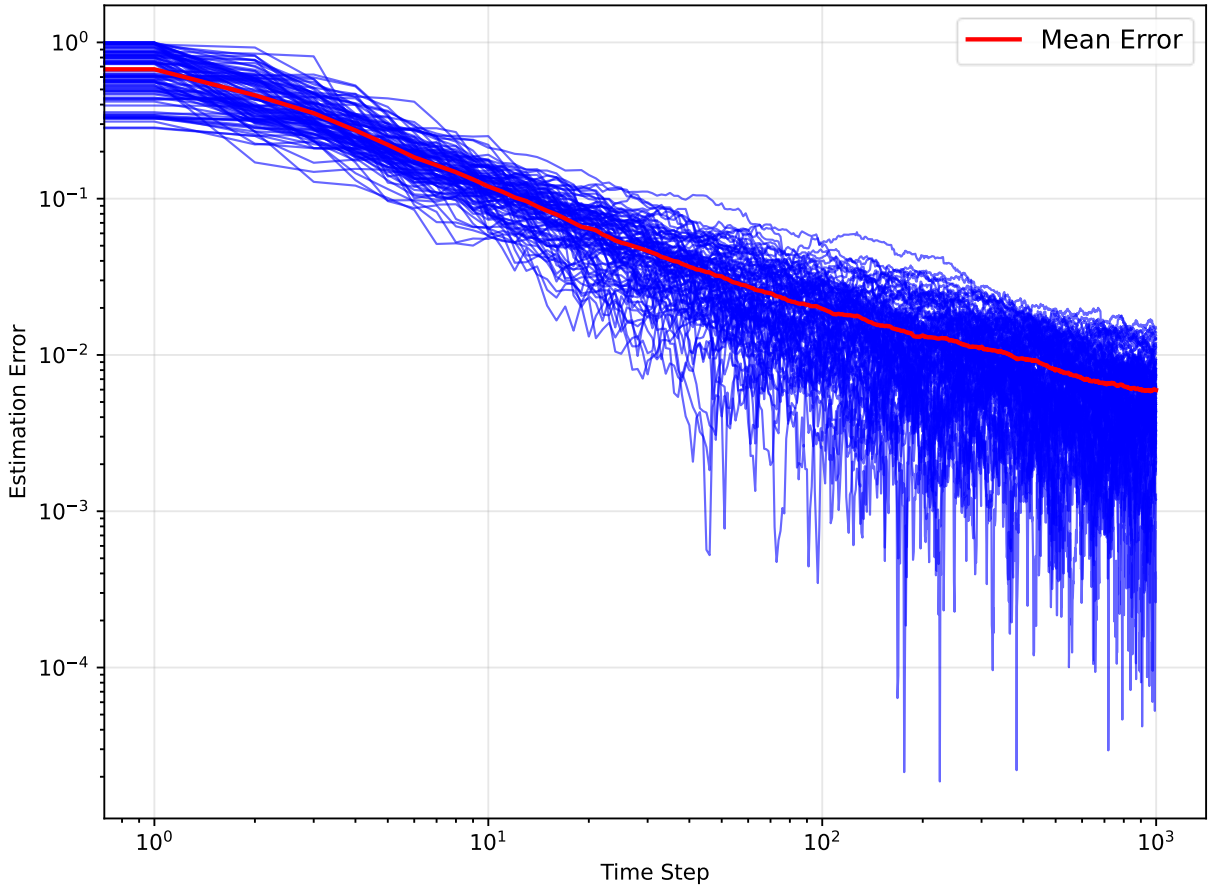
LinUCB: regret and error plots

LinUCB on a linear bandit with $k = 100$ arms and $d = 5$ dimensional features

Regret R_t



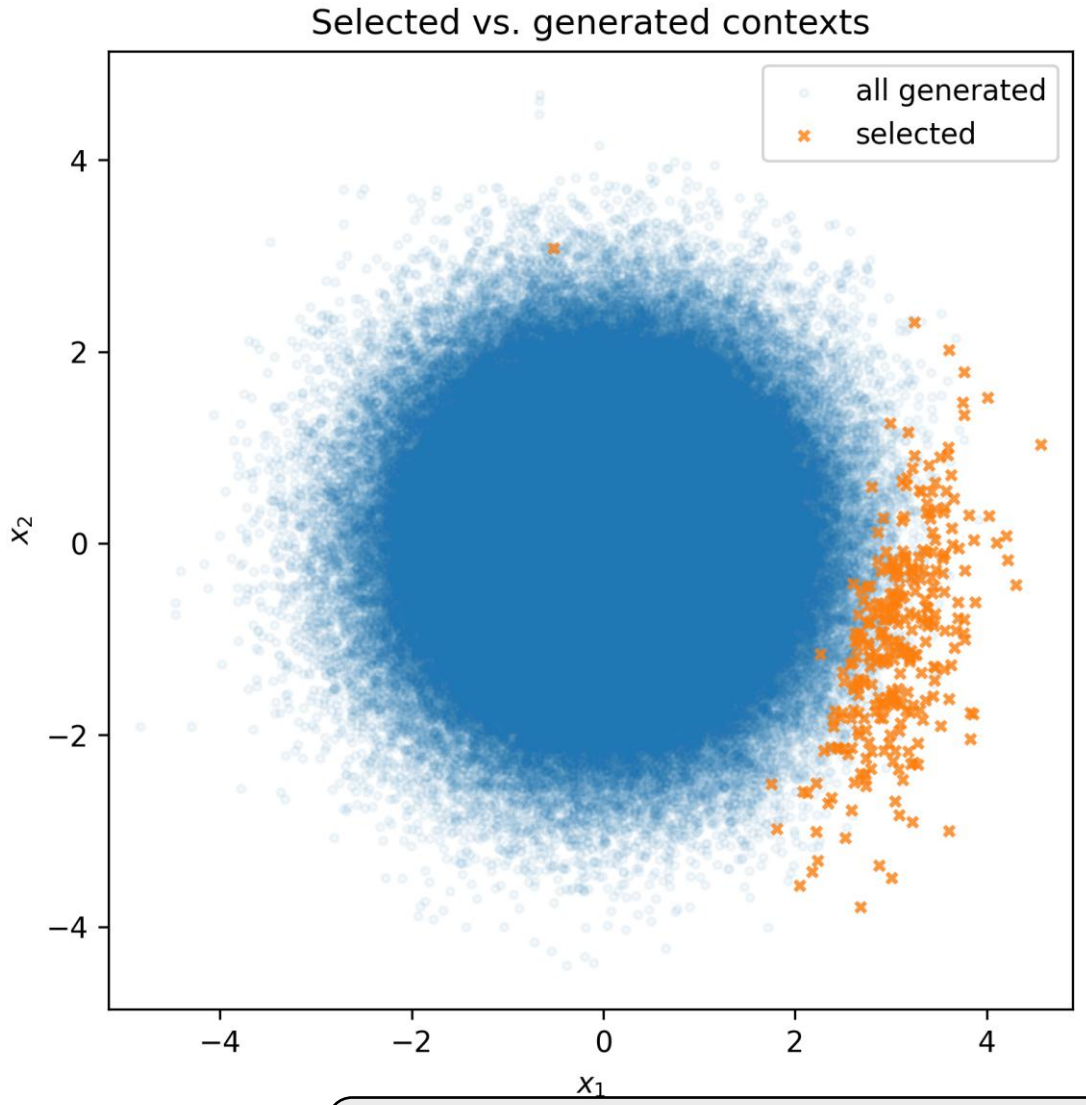
Error $\|\theta^* - \hat{\theta}_t\|$

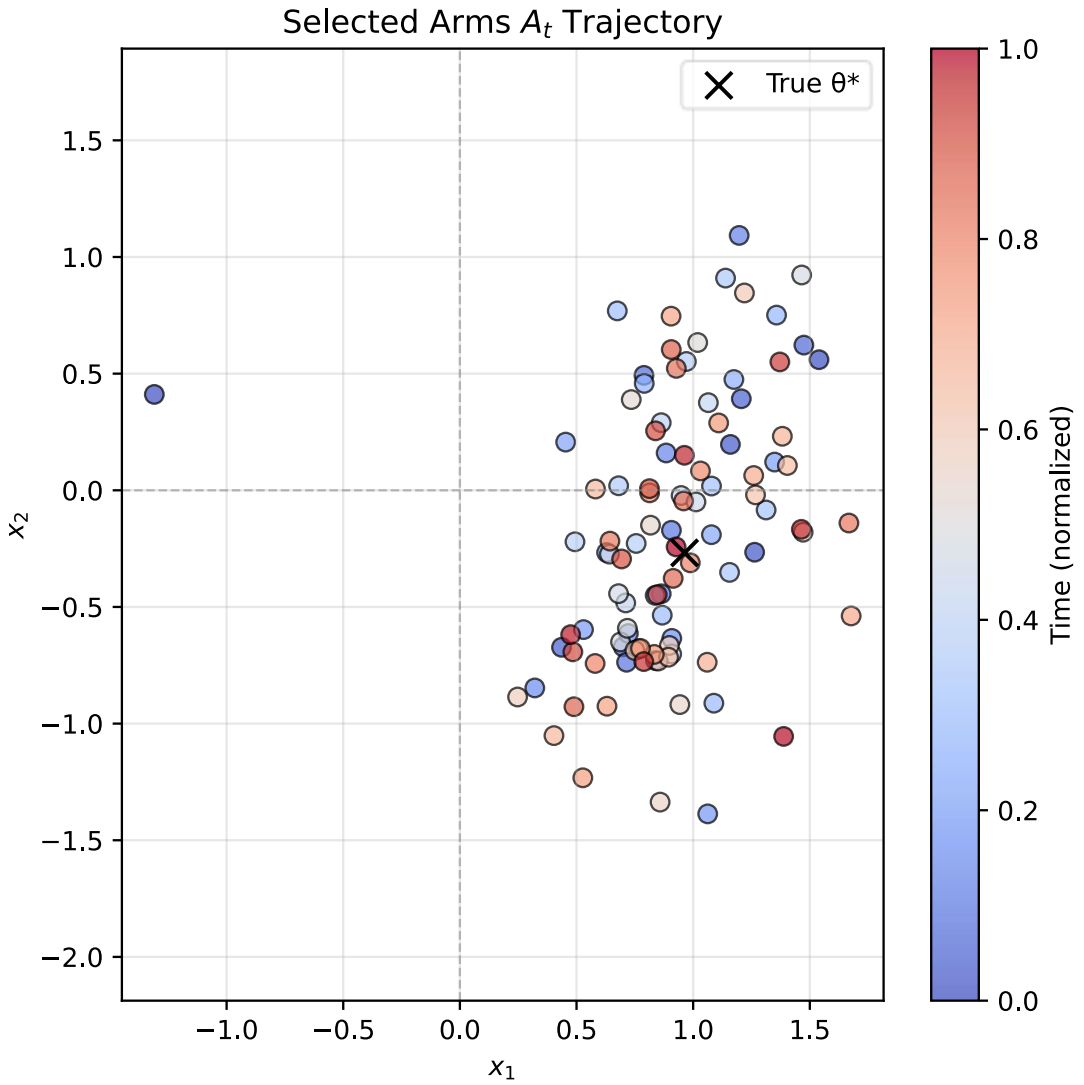


$R_t = O(\sqrt{t})$, similar to unstructured bandits

$\|\theta^* - \hat{\theta}_t\| = O(1/\sqrt{t})$, similar to linear regression

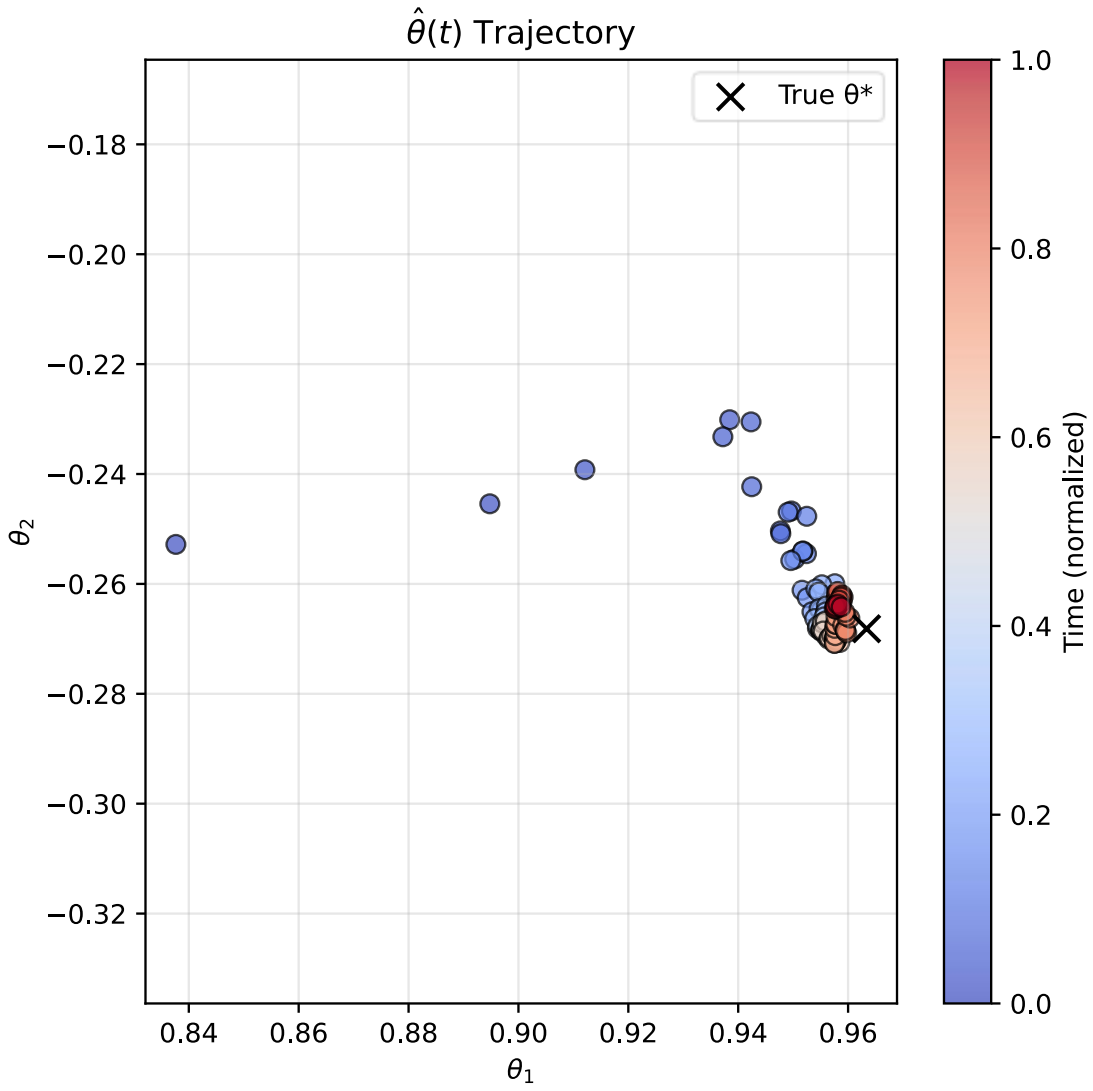
Actions over time

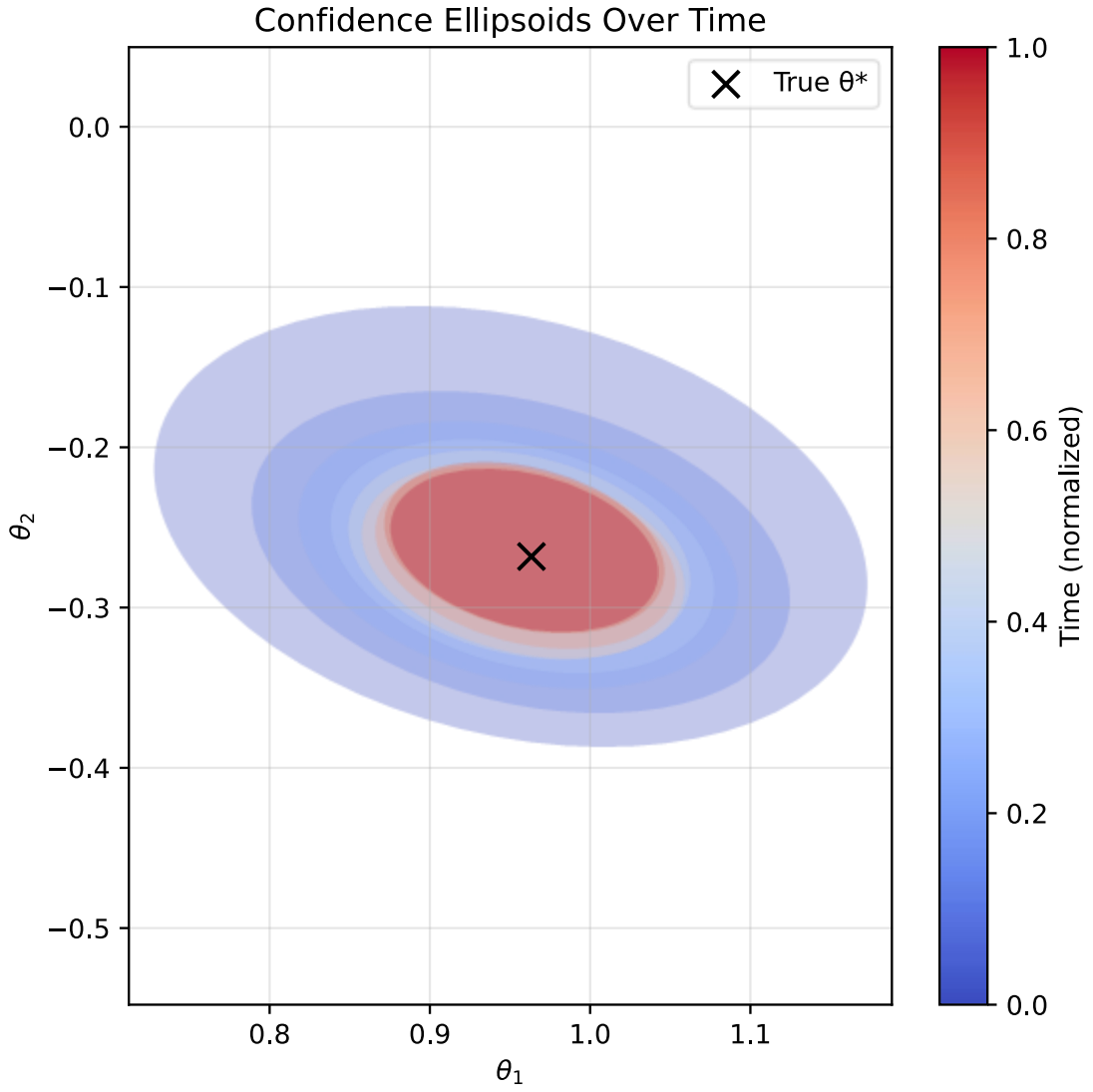




Selected arms align in the direction of θ^*

Confidence ellipsoids with time





Analysis: high-level view

Elliptic Potential Lemma:
Bound on $\sum_{t=1}^{n}\left(1 \wedge \|a_t\|^2_{V_t^{-1}}\right)$
as function of n, d, L, V_0

- Recall:
  - n: # epochs
  - d: dimension of a, θ
  - λV_0: regulariser
  - δ: bound on prob. of confidence ellipse violation

Define good event (prob. $1-\delta$)
parametrized by $\beta_1, \ldots, \beta_n$
and show that $\hat{R}_n \leq f(n, \beta_n, V_0, V_n)$

Define good event (prob. $1-\delta$)
parametrized by $\beta_1, \ldots, \beta_n$
and show that $\hat{R}_n \leq f(n, d, \beta_n, \lambda, L)$

Bound on $\left\|\hat{\theta}_t - \theta^*\right\|_{V_t}$
with probability at least $1-\delta$
as a function of $\delta, d, \lambda, \theta^*, V_t$

Setting $\delta = 1/n$ gives
bound on regret $R_n \leq Cd\sqrt{n}\log(nL)$

Analysis: high-level view

Elliptic Potential Lemma:
Bound on $\sum_{t=1}^{n}\left(1 \wedge \|a_t\|^2_{V_t^{-1}}\right)$
as function of n, d, L, V_0

- Recall:
  - n: # epochs
  - d: dimension of a, θ
  - λV_0: regulariser
  - δ: bound on prob. of confidence ellipse violation

Define good event (prob. $1-\delta$)
parametrized by $\beta_1, \ldots, \beta_n$
and show that $\hat{R}_n \leq f(n, \beta_n, V_0, V_n)$

Define good event (prob. $1-\delta$)
parametrized by $\beta_1, \ldots, \beta_n$
and show that $\hat{R}_n \leq f(n, d, \beta_n, \lambda, L)$

Bound on $\left\|\hat{\theta}_t - \theta^*\right\|_{V_t}$
with probability at least $1-\delta$
as a function of $\delta, d, \lambda, \theta^*, V_t$

Setting $\delta = 1/n$ gives
bound on regret $R_n \leq Cd\sqrt{n}\log(nL)$

Key concentration result

Theorem (20.5 in L&S):

- Suppose we have a sequence of datapoints $(A_1, X_1, A_2, X_2, \dots)$ satisfying
- $X_t = \langle \theta^*, A_t \rangle + \eta_t$, η_t independent and 1-subgaussian
- Each A_t may depend on past $(A_1, X_1, \dots, A_{t-1}, X_{t-1})$ (any bandit algorithm)
- Let $\hat{\theta}_t$ be least-squares estimator with regulariser weight λ
- Fix $\delta \in (0,1)$. Then, with probability at least $1 - \delta$, it **holds for all $t \in \mathbb{N}$**:

$$\left\| \hat{\theta}_t - \theta^* \right\|_{V_t} < \sqrt{\lambda} \|\theta^*\|_2 + \sqrt{2 \log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det V_t}{\lambda^d}\right)}$$

we shall not prove this here

Intuition for concentration bound

We will prove a simple concentration inequality for a single arm a

- Recall $\hat{\theta}_t = V_t^{-1}\left(\sum_{i\in[t]} A_i X_i\right)$ and $X_t = A_t^\top \theta^* + \eta_t$
- $\langle a, \hat{\theta}_t - \theta^* \rangle = \left\langle a, V_t^{-1}\left(\sum_{i\in[t]} A_i X_i\right) - \theta^* \right\rangle = \left\langle a, V_t^{-1}\left(\sum_{i\in[t]} A_i \left(A_i^\top \theta^* + \eta_i\right)\right) - \theta^* \right\rangle$
- Observe: $\sum_{i\in[t]} A_i \left(A_i^\top \theta^*\right) = V_t \theta^* \Rightarrow V_t^{-1}\left(\sum_{i\in[t]} A_i \left(A_i^\top \theta^*\right)\right) = \theta^*$
- $\langle a, \hat{\theta}_t - \theta^* \rangle = \left\langle a, V_t^{-1}\left(\sum_{i\in[t]} A_i \eta_i\right)\right\rangle = \left\langle a, V_t^{-1}\left(\sum_{i\in[t]} A_i\right)\right\rangle \eta_i = \sum_{i\in[t]} \langle a, V_t^{-1} A_i \rangle \eta_i$
- This is a linear combination of independent, 1-subgaussian r.v.s $\{\eta_i\}_{i\in[n]}$

Intuition for concentration bound

In the last slide, we showed:

- $\langle a, \hat{\theta}_t - \theta^* \rangle = \sum_{i \in [t]} \langle a, V_t^{-1} A_i \rangle \eta_i$
- a linear combination of independent, 1 sub-Gaussian r.v.s $\{\eta_i\}_{i \in [n]}$
  - $\eta \sim$ 1-subgaussian $\Rightarrow \alpha\eta \sim$ $|\alpha|$-subgaussian and
  - $\eta_1, \eta_2 \sim$ independent σ_1-(σ_2 -)subgaussian $\Rightarrow \eta_1 + \eta_2$ subgaussian with param $\sqrt{\sigma_1^2 + \sigma_2^2}$
  - $\sum_{i \in [t]} \langle a, V_t^{-1} A_i \rangle \eta_i$ is also a subgaussian random variable
- Applying concentration inequality for subgaussians, we get

$$\mathbb{P}\left(\langle a, \hat{\theta}_t - \theta^* \rangle \geq \sqrt{2 \|a\|^2_{V_t^{-1}} \log 1/\delta} \right) \leq \delta$$

$$\sum_{i \in [t]} \langle a, V_t^{-1} A_i \rangle^2 = \|a\|^2_{V_t^{-1}}$$

Simple bound is not sufficient

- The bound $\mathbb{P}\left(\langle a, \hat{\theta}_t - \theta^* \rangle \geq \sqrt{2 \log \frac{1}{\delta}\left(\sum_{i \in [t]} \langle a, V_t^{-1} A_i \rangle^2\right)}\right) \leq \delta$
- holds only for a single arm

- For regret guarantees, we need such a bound for all arms

- One option: union bound

- If the action set $\mathcal{A} = \{a_1, a_2, \ldots, a_k\}$ is small, this is a reasonable strategy
  - regret will scale with $\log k$
- For large action sets, we want a bound independent of k
- Need a bound on $\mathbb{P}\left(\left\|\hat{\theta}_t - \theta^*\right\|_{V_t} \geq \beta_t\right)$
- See L&S, Chapter 20

Analysis: high-level view

Elliptic Potential Lemma:
Bound on $\sum_{t=1}^{n}\left(1 \wedge \|a_t\|^2_{V_t^{-1}}\right)$
as function of n, d, L, V_0

- Recall:
  - n: # epochs
  - d: dimension of a, θ
  - λV_0: regulariser
  - δ: bound on prob. of confidence ellipse violation

Define good event (prob. $1-\delta$)
parametrized by $\beta_1, \ldots, \beta_n$
and show that $\hat{R}_n \leq f(n, \beta_n, V_0, V_n)$

Define good event (prob. $1-\delta$)
parametrized by $\beta_1, \ldots, \beta_n$
and show that $\hat{R}_n \leq f(n, d, \beta_n, \lambda, L)$

Bound on $\left\|\hat{\theta}_t - \theta^*\right\|_{V_t}$
with probability at least $1-\delta$
as a function of $\delta, d, \lambda, \theta^*, V_t$

Setting $\delta = 1/n$ gives bound on
regret $R_n \leq Cd\sqrt{n}\log(nL)$

Elliptic Potential Lemma

Lemma: Assume the following:

- $a_1, a_2, \ldots, a_n$ is a sequence in $\mathbb{R}^d$ such that $\|a_t\|_2 \leq L$ for all $t \in [n]$
- For all $t \in [n]$, $V_t = V_0 + \sum_{s \leq t} a_s a_s^\top$, and V_0 is positive definite

Then,

$$\sum_{t=1}^{n} \left(1 \wedge \|a_t\|^2_{V_{t-1}^{-1}}\right) \leq 2 \log\left(\frac{\det V_n}{\det V_0}\right) \leq 2d \log\left(\frac{\mathrm{Tr}\,(V_0) + nL^2}{d(\det V_0)^{1/d}}\right)$$

- $V_0 = \lambda I \Rightarrow \det V_0 = \lambda^d$, and $\mathrm{Tr}\,(V_0) = d\lambda$
- $\Rightarrow \log\left(\frac{\mathrm{Tr}\,(V_0)+nL^2}{d(\det V_0)^{1/d}}\right) = \log\left(\frac{d\lambda+nL^2}{d\lambda}\right)$

$$\Rightarrow \sum_{t=1}^{n} \left(1 \wedge \|a_t\|^2_{V_{t-1}^{-1}}\right) \leq 2d \log\left(\frac{d\lambda + nL^2}{d\lambda}\right)$$

Analysis: high-level view

Elliptic Potential Lemma:
Bound on $\sum_{t=1}^{n}\left(1 \wedge \|a_t\|^2_{V_t^{-1}}\right)$
as function of n, d, L, V_0

- Recall:
  - n: # epochs
  - d: dimension of a, θ
  - λV_0: regulariser
  - δ: bound on prob. of confidence ellipse violation

Define good event (prob. $1-\delta$)
parametrized by $\beta_1, \ldots, \beta_n$
and show that $\hat{R}_n \leq f(n, \beta_n, V_0, V_n)$

Define good event (prob. $1-\delta$)
parametrized by $\beta_1, \ldots, \beta_n$
and show that $\hat{R}_n \leq f(n, d, \beta_n, \lambda, L)$

Bound on $\left\|\hat{\theta}_t - \theta^*\right\|_{V_t}$
with probability at least $1-\delta$
as a function of $\delta, d, \lambda, \theta^*, V_t$

Setting $\delta = 1/n$ gives
bound on regret $R_n \leq Cd\sqrt{n}\log(nL)$

Regret guarantees for LinUCB

Theorem: Assume the following:

- $\max_{t\in[n]} \sup_{a,b\in\mathcal{A}_t} (a-b)^\top \theta^* \le 1$
- $\|a\|_2 \le L$ for all $a \in \cup_{t=.1}^{n} \mathcal{A}_t$
- $\beta_1, \beta_2, \ldots$ are chosen such that with probability $\ge 1-\delta$,

$$\forall\, t, \qquad \theta^* \in \mathcal{C}_t = \left\{\theta \in \mathbb{R}^d : \left\|\theta - \hat{\theta}_{t-1}\right\|_{V_{t-1}}^2 \le \beta_t\right\}$$

Then, with probability $\ge 1-\delta$,

$$\hat{R}_n \le \sqrt{8dn\,\beta_n \log\left(\frac{d\lambda + nL^2}{d\lambda}\right)}$$

we will prove this

Proof outline

- Proof can be divided into two steps:

1. Bounding instantaneous regret: $r_t \leq \sqrt{\beta_n}\|A_t\|_{V_{t-1}^{-1}}$, by using
   - Good event (concentration event),
   - Definition of UCB
   - Cauchy-Schwarz inequality

regret bounded by exploration bonus in UCB algorithm!

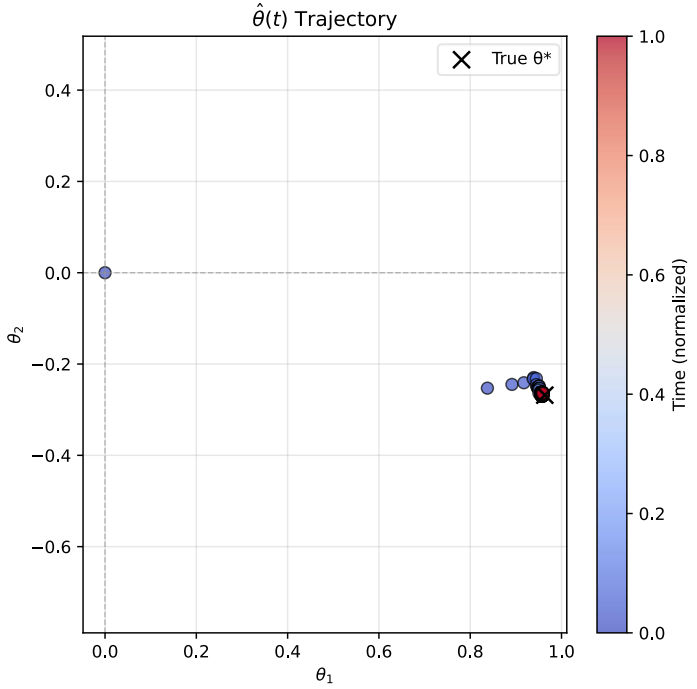
2. Bounding cumulative regret: $R_n \leq \sqrt{2dn\beta_n \log\left(\frac{d\lambda+nL^2}{d\lambda}\right)}$, by using
   - Elliptical potential lemma
   - Other algebraic inequalities

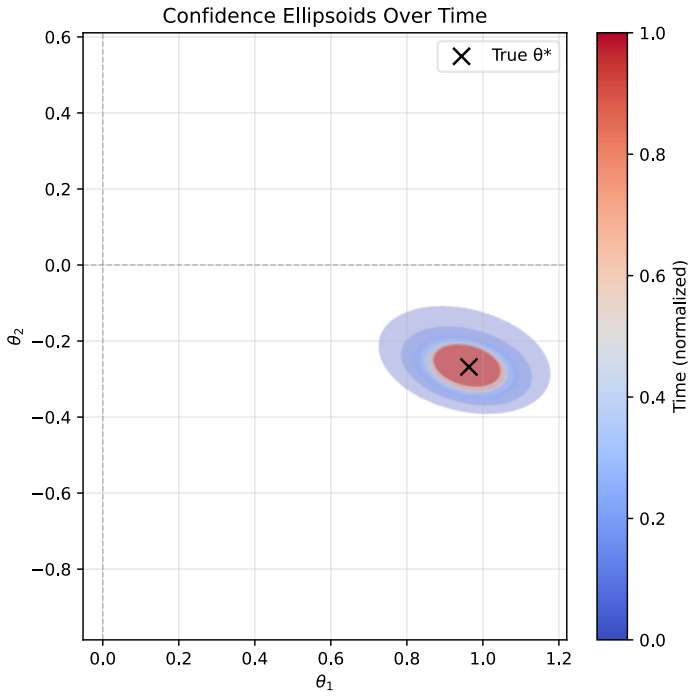
Understanding the good event

- Good event:

$$\forall\, t, \theta^* \in \mathcal{C}_t = \left\{\theta \in \mathbb{R}^d : \left\|\theta - \hat{\theta}_{t-1}\right\|^2_{V_{t-1}} \leq \beta_t\right\}$$

- Holds with probability $1 - \delta$
  - For now, we have assumed this
  - Intuitively, this can be reasoned as follows
- $\left\|\theta^* - \hat{\theta}_{t-1}\right\|^2_{V_{t-1}} \leq \beta_t$ is a concentration event
  - Recall $\beta_t = O(\log(t))$
  - We expect eigenvalues of $V_{t-1} \propto t - 1$
  - Put together, we get $\left\|\theta^* - \hat{\theta}_t\right\| = O\left(\sqrt{\frac{\log(t)}{t}}\right)$
  - With more data, the estimate $\hat{\theta}_t$ gets closer to θ^*
- We will assume throughout that good event holds





Using the UCB algorithm bound

- Define $A_t^* = \arg\max_{a \in \mathcal{A}_t} \langle \theta^*, a \rangle$ (best arm at time t)
- Instantaneous regret is (reward of best arm) – (reward of chosen arm)
- $\Rightarrow r_t = \langle \theta^*, A_t^* \rangle - \langle \theta^*, A_t \rangle$
- $\theta^* \in \mathcal{C}_t \Rightarrow \langle \theta^*, A_t^* \rangle \leq \max_{\theta \in \mathcal{C}_t} \langle \theta \ , A_t^* \rangle = UCB_t(A_t^*)$
- $\Rightarrow r_t = \langle \theta^*, A_t^* \rangle - \langle \theta^*, A_t \rangle \leq UCB_t(A_t^*) - \langle \theta^*, A_t \rangle$
- Arm A_t was played by UCB algorithm $\Rightarrow UCB_t(A_t^*) \leq UCB_t(A_t)$
- $\Rightarrow r_t = \langle \theta^*, A_t^* \rangle - \langle \theta^*, A_t \rangle \leq UCB_t(A_t^*) - \langle \theta^*, A_t \rangle \leq UCB_t(A_t) - \langle \theta^*, A_t \rangle$

$$UCB_t(a) = \max_{\theta \in \mathcal{C}_t} \langle \theta, a \rangle$$

Using Cauchy-Schwarz Inequality

- $r_t \leq UCB_t(A_t) - \langle \theta^*, A_t \rangle$
- **Define** $\tilde{\theta}_t = \arg\max_{\theta \in \mathcal{C}_t} \langle \theta, A_t \rangle$. This implies $UCB_t(A_t) = \langle \tilde{\theta}_t, A_t \rangle$
- $r_t \leq UCB_t(A_t) - \langle \theta^*, A_t \rangle = \langle \tilde{\theta}_t, A_t \rangle - \langle \theta^*, A_t \rangle = \langle \tilde{\theta}_t - \theta^*, A_t \rangle$
- Write: $\langle \tilde{\theta}_t - \theta^*, A_t \rangle = (\tilde{\theta}_t - \theta^*)^T A_t = (\tilde{\theta}_t - \theta^*)^T V_{t-1}^{1/2} V_{t-1}^{-1/2} A_t$
- By Cauchy-Schwarz inequality, $\langle \tilde{\theta}_t - \theta^*, A_t \rangle \leq \|\tilde{\theta}_t - \theta^*\|_{V_{t-1}} \|A_t\|_{V_{t-1}^{-1}}$
  - $\langle \tilde{\theta}_t - \theta^*, A_t \rangle \leq \sqrt{(\tilde{\theta}_t - \theta^*)^T V_{t-1}^{\frac{1}{2}} V_{t-1}^{\frac{1}{2}} (\tilde{\theta}_t - \theta^*)} \sqrt{A_t^T V_{t-1}^{-\frac{1}{2}} V_{t-1}^{-\frac{1}{2}} A_t}$
- $r_t \leq \langle \tilde{\theta}_t, A_t \rangle - \langle \theta^*, A_t \rangle = \langle \tilde{\theta}_t - \theta^*, A_t \rangle \leq \|\tilde{\theta}_t - \theta^*\|_{V_{t-1}} \|A_t\|_{V_{t-1}^{-1}}$
- Why did we do this step?
  - Because we have a simple bound on $\|\tilde{\theta}_t - \theta^*\|_{V_{t-1}}$!

Using the good event

- $r_t \le \left\|\tilde{\theta}_t - \theta^*\right\|_{V_{t-1}} \|A_t\|_{V_{t-1}^{-1}}$
- Recall $\tilde{\theta}_t = \arg\max_{\theta \in \mathcal{C}_t} \langle \theta, A_t \rangle$. This implies $\tilde{\theta}_t \in \mathcal{C}_t$
- We know $\mathcal{C}_t = \left\{\theta \in \mathbb{R}^d : \left\|\theta - \hat{\theta}_{t-1}\right\|_{V_{t-1}}^2 \le \beta_t\right\}$
- Together, we get $\left\|\tilde{\theta}_t - \theta^*\right\|_{V_{t-1}} \le \sqrt{\beta_t}$
- $\Rightarrow r_t \le \left\|\tilde{\theta}_t - \theta^*\right\|_{V_{t-1}} \|A_t\|_{V_{t-1}^{-1}} \le \sqrt{\beta_t} \|A_t\|_{V_{t-1}^{-1}}$
- Finally, we chose $\beta_t \le \beta_n$ for all $t \le n$
- $\Rightarrow r_t \le \left\|\tilde{\theta}_t - \theta^*\right\|_{V_{t-1}} \|A_t\|_{V_{t-1}^{-1}} \le \sqrt{\beta_t} \|A_t\|_{V_{t-1}^{-1}} \le \sqrt{\beta_n} \|A_t\|_{V_{t-1}^{-1}}$

From instantaneous regret to cumulative regret

- $r_t \leq \sqrt{\beta_n}\|A_t\|_{V_{t-1}^{-1}}$
- Summing up instantaneous regret to get cumulative regret:
- $R_n = \sum_{t=1}^{n} r_t \leq \sqrt{\beta_n}\left(\sum_{t=1}^{n}\|A_t\|_{V_{t-1}^{-1}}\right)$
- Challenge: how to bound $\sum_{t=1}^{n}\|A_t\|_{V_{t-1}^{-1}}$
- We want to prove: $R_n \leq \sqrt{8dn\,\beta_n \log\left(\frac{d\lambda+nL^2}{d\lambda}\right)}$
- We need to prove: $\sum_{t=1}^{n}\|A_t\|_{V_{t-1}^{-1}} \leq \sqrt{8dn\,\log\left(\frac{d\lambda+nL^2}{d\lambda}\right)}$

Intuition behind series summation

- Why can we expect $\sum_{t=1}^{n} \|A_t\|_{V_{t-1}^{-1}} \approx \sqrt{nd}$?
- $V_t = \sum_{s=1}^{t} A_s A_s^\top$
- Suppose: all A_s are of norm L,
- And A_s chosen uniformly among unit vectors in d dimensions
- Eigenvalues of $V_t \sim \frac{tL^2}{d}$
- Eigenvalues of $V_t^{-1} \sim \frac{d}{tL^2}$
- $\|A_t\|^2_{V_{t-1}^{-1}} \sim \frac{d}{tL^2} \times L^2 = \frac{d}{t}$
- $\sum_{t=1}^{n} \|A_t\|_{V_{t-1}^{-1}} \sim \sum_{t=1}^{n} \sqrt{\frac{d}{t}} \approx \sqrt{nd}$

- This is made rigorous by Elliptical Potential Lemma
- Omitting the rest of the proof

Analysis: high-level view

Elliptic Potential Lemma:
Bound on $\sum_{t=1}^{n}\left(1 \wedge \|a_t\|^2_{V_t^{-1}}\right)$
as function of n, d, L, V_0

- Recall:
  - n: # epochs
  - d: dimension of a, θ
  - λV_0: regulariser
  - δ: bound on prob. of confidence ellipse violation

Define good event (prob. $1-\delta$)
parametrized by $\beta_1, \ldots, \beta_n$
and show that $\hat{R}_n \leq f(n, \beta_n, V_0, V_n)$

Define good event (prob. $1-\delta$)
parametrized by $\beta_1, \ldots, \beta_n$
and show that $\hat{R}_n \leq f(n, d, \beta_n, \lambda, L)$

Bound on $\left\|\hat{\theta}_t - \theta^*\right\|_{V_t}$
with probability at least $1-\delta$
as a function of $\delta, d, \lambda, \theta^*, V_t$

Setting $\delta = 1/n$ gives
bound on regret $R_n \leq Cd\sqrt{n}\log(nL)$

Regret guarantees for LinUCB

Theorem: Assume the following:

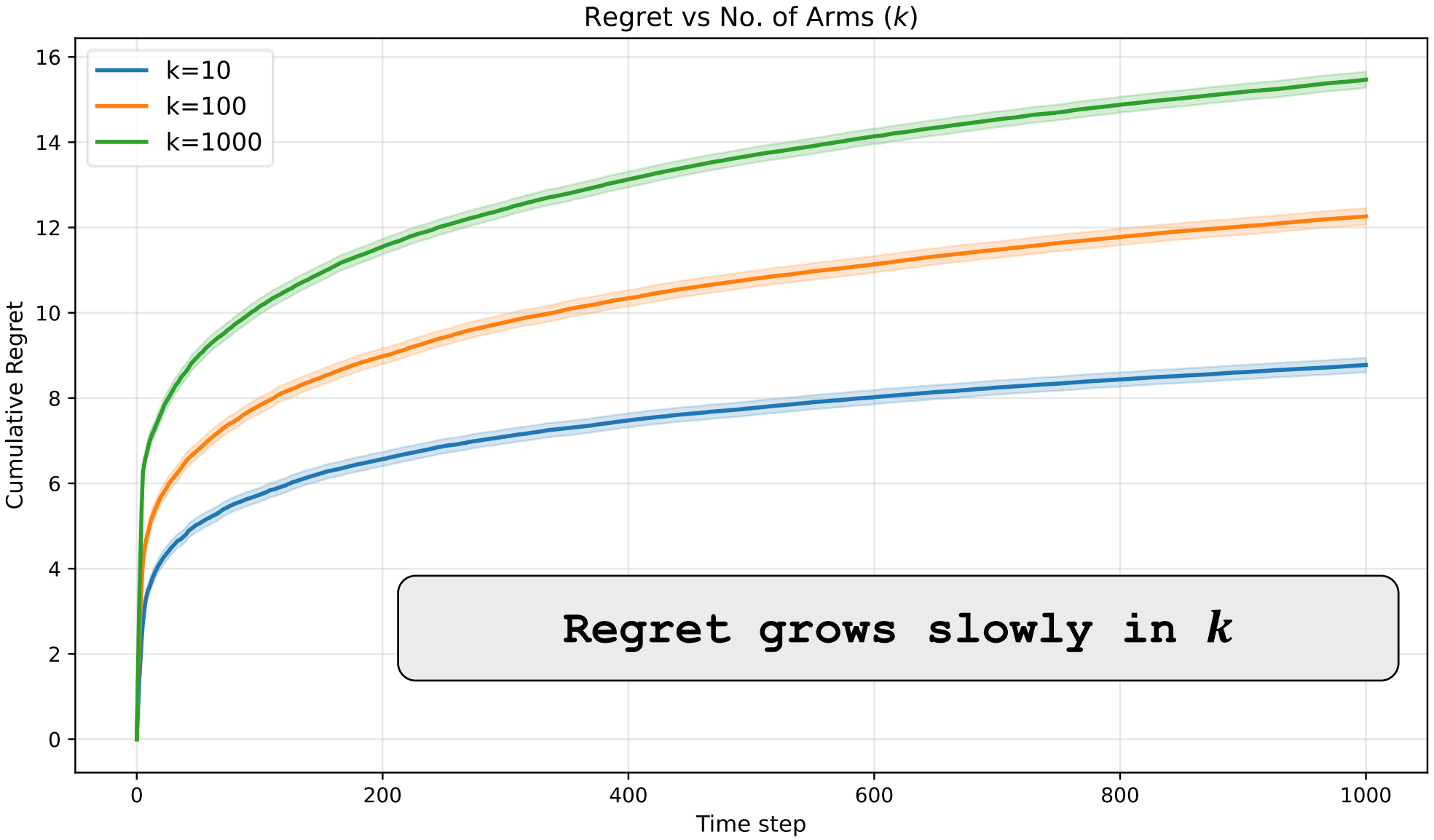
- $\max_{t\in[n]} \sup_{a,b\in\mathcal{A}_t} (a-b)^\top \theta^* \le 1$
- $\|a\|_2 \le L$ for all $a \in \bigcup_{t=.1}^{n} \mathcal{A}_t$
- $\beta_1, \beta_2, \ldots$ are chosen such that with probability $\ge 1-\delta$, $\delta = 1/n$

$$\forall\, t, \qquad \theta^* \in \mathcal{C}_t = \left\{\theta \in \mathbb{R}^d : \left\|\theta - \hat{\theta}_{t-1}\right\|_{V_{t-1}}^2 \le \beta_t\right\}$$

  - $\beta_t = O\left(\log\left(\frac{1}{\delta}\right) + d\,\log\left(\frac{d\lambda+nL^2}{d\lambda}\right)\right)$ suffices

$$R_n = O\left(d\sqrt{n}\,\log(nL)\right)$$

Regret: variation with k



Linear Bandit

$d = 5$

$a_i^t \sim \mathcal{N}(0, I_d)$

Regret: variation with d

Linear Bandit

$k = 100$

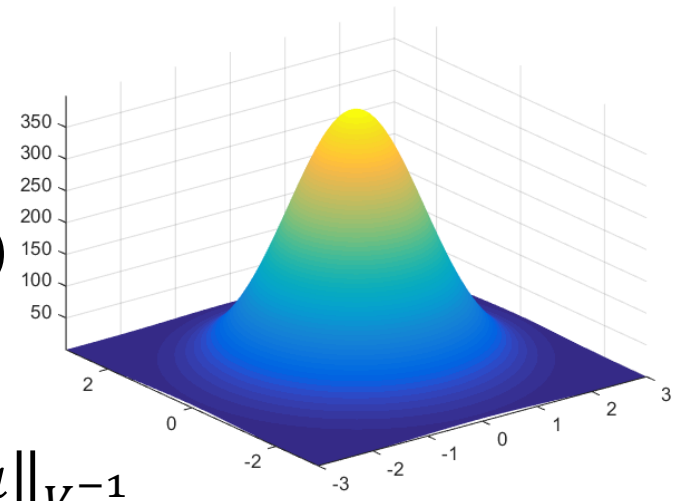
$a_i^t \sim \mathcal{N}(0, I_d)$

Thompson Sampling for linear bandits

Key ideas in Thompson sampling:

- Maintain belief distribution over unknown parameters
- At each time t,
  - Sample parameters from belief distribution (exploration)
  - Play assuming sample is true parameter (exploitation)

- In linear bandits, unknown parameter is θ^*
- Maintain a multivariate Gaussian belief $f_t(\theta^*) \sim \mathcal{N}(\mu_t, \Sigma_t)$
- After sampling $\tilde{\theta}_t \sim f_t(\theta^*)$, play $A_t = \arg\max_{a \in \mathcal{A}_t} a^\top \tilde{\theta}_t$
- <u>Computationally faster</u> than $A_t = \arg\max_{a \in \mathcal{A}_t} a^\top \hat{\theta}_{t-1} + \sqrt{\beta_t}\, \|a\|_{V_{t-1}^{-1}}$



- Heuristic: $\mu_t = \hat{\theta}_t = V_t^{-1}\left(\sum_{i \in [t]} A_i X_i\right)$; $\Sigma_t^{-1} = V_t = \sum_{i \in [t]} A_i A_i^\top$

Generalised linear bandits

- Linear bandit reward model: $X_t = A_t^\top \theta^* + \eta_t$
- Reward can be unbounded
- Unrealistic for scenarios where reward is binary: click or no click
- Better model: $X_t = g(A_t^\top \theta^*) + \eta_t$
- $g(\cdot)$: $\mathbb{R} \to [0,1]$: nonlinear link function. E.g., sigmoid: $g(t) = 1/(1+e^{-t})$

- Captures: $X_t = 1$ w.p. $g(A_t^\top \theta^*)$, $X_t = 0$ otherwise
- Often used in practice, e.g., online advertisement placement

- All principles of LinUCB and Thompson Sampling carry over!
- Only change: $\hat{\theta}_t$ estimated by logistic regression
- Tight regret bounds being recently discovered (2020 onwards)

Conclusion

- Introduced the linear bandit model
  - Arms have known feature vectors $a \in \mathbb{R}^d$
  - Reward is a noisy linear function of features $X_t = A_t^\top \theta^* + \eta_t$
  - Model for recommender systems
- Different information structure
  - Unknown parameter is $\theta^* \in \mathbb{R}^d$
  - Pulling one arm conveys information about other arm rewards
  - Regret scales as $O(d\sqrt{n})$ rather than $O(\sqrt{kn})$
- Parameter estimation by linear regression
  - Importance of regulariser
  - Uncertainty in estimate quantified by $V_t = \sum_{i \in [t]} A_i A_i^\top$
  - Uncertainty intervals → uncertainty ellipsoids
- LinUCB algorithm
  - Extends optimism in the face of uncertainty principle
  - Provable regret guarantees – key point: independent of k!

Reading:

Chapters 19 & 20 of L&S (some parts very technical)