High-Dimensional Linear (Semi-)Bandits, Efficient Algorithms and Optimality

Raymond Zhang

Supervised by: **Richard Combes** and **Sheng Yang**

Friday 12 December 2025

Université Paris Saclay, CentraleSupélec / L2S, France

Outline

Linear Bandits

Linear Bandit Model

- At each time $t \in \{1, 2, \cdots, T\}$:
- The learner chooses an action $A_t \in \mathcal{A} \subset \mathbb{R}^d$
- Receives reward:

$$Y_t = A_t^\top \mu + Z_t,$$

  where $\mu \in \mathbb{R}^d$ is an unknown parameter, and Z_t is σ^2-sub-Gaussian noise.
- Goal: maximize cumulative reward (or minimize regret):

$$\mathcal{R}_T^\pi(\mu) = \sum_{t=1}^{T} a^{\star\top} \mu - \mathbb{E}[\sum_{t=1}^{T} A_t^\top \mu]$$

  where $a^\star = \arg\max_{a \in \mathcal{A}} a^\top \mu$.

Action Set and Regret

Linear Bandits

Lower bounds

Locally Asymptotic Minimax Lower Bound

Consider, $\mathcal{A} = \{a \in \mathbb{R}^{d+1} : \|a\|_M \leqslant 1\}$ and $B > 0$.

Theorem (5.1, page 152)

For any algorithm π, any $T \geqslant 1$ and any $\mu \in \mathbb{R}^{d+1}, \|\mu\|_{M^{-1}} = B$, there exists $\mu' \in \mathbb{R}^{d+1}, \|\mu'\|_{M^{-1}} = B$ such that :

$$\|\mu - \mu'\|_{M^{-1}}^2 \leq \min(\sigma d B/\sqrt{T}\,, 4B^2),$$

and :

$$\mathcal{R}_T^\pi(\mu') \geq \min\left(\frac{\sigma d\sqrt{T}}{16}, \frac{BT}{4}\right).$$

Previous lower bounds only showed :

$$\forall \pi, \exists \mu \in \mathcal{B}(0, \frac{d}{4\sqrt{3T}}), \mathcal{R}_T^\pi(\mu) \geqslant c\sigma d\sqrt{T}\,.$$

Lower Bound

Linear Bandits

Algorithms

Optimistic Linear Bandit Algorithms

- At each time $t \in \{1, 2, \cdots, T\}$:
- The learner chooses an action $A_t \in \mathcal{A}$

$$A_t = \arg \max_{\substack{a \in \mathcal{A} \\ \theta \in \mathcal{C}_{t-1}}} a^\top \theta$$

- Where $\mathcal{C}_{t-1}$ is a confidence set for μ at time $t-1$.

$$\mathbb{P}(\forall t, \mu \in \mathcal{C}_{t-1}) \geqslant 1 - \delta$$

- $\mathcal{C}_{t-1}$ is an ellipsoid[1] centered at the OLS or Ridge estimator $\widehat{\mu}_{t-1}$.

$$\mathcal{C}_t := \{\theta \in \mathbb{R}^d, \|\widehat{\mu}_t - \theta\|_{V_t} \leqslant \beta_t(\delta)\}$$

[1] Abbasi-yadkori, Yasin, Dávid Pál, and Csaba Szepesvári. Improved Algorithms for Linear Stochastic Bandits

Concentration and LinUCB

$\mathcal{NP}$-Hardness

In general, this bilinear optimization problem is $\mathcal{NP}$-hard to approximate.

Proposition (6.5, page 179)

For $p > 2$, when $\mathcal{A} = \{a \in \mathbb{R}^d : \|a\|_p \leqslant 1\}$, and W is a positive definite matrix, the problem :

$$\max_{\substack{a \in \mathcal{A} \\ \|\theta\|_W \leqslant 1}} a^\top \theta \,,$$

is $\mathcal{NP}$-hard to approximate below an approximation ratio $\varepsilon_p > 0$.

By a reduction from an operator norm computation problem[2].

[2] Bhattiprolu, Vijay, Mrinal Kanti Ghosh, Venkatesan Guruswami, Euiwoong Lee, and Madhur Tulsiani. Inapproximability of Matrix $p \to q$ Norms

$\mathcal{A}$ finite

We have :

$$\max_{\substack{\|a\|_M \leqslant 1 \\ \|\theta - c\|_W \leqslant 1}} a^\top \theta \qquad (P_B)$$

By maximizing over θ first, we have :

$$\max_{a \in \mathcal{A}} a^\top c + \|a\|_{W^{-1}} \, . \qquad (P_A)$$

This is a convex function !...**Maximizing** over a convex set.

If $\mathcal{A}$ is finite and not too big one can do an exhaustive search.

When $\mathcal{A}$ is an ellipsoid

We have :

$$\max_{\substack{\|a\|_M \leqslant 1 \\ \|\theta - c\|_W \leqslant 1}} a^\top \theta \qquad (P_B)$$

By a change of variable and maximizing over a first, we have :

$$\max_{\|\phi - b\|_\Lambda \leqslant 1} \|\phi\|_2 \qquad (P''_B)$$

Where $b = UM^{-1/2}c$ and $U\Lambda U^T = M^{\frac{1}{2}} W M^{\frac{1}{2}}$. Λ is diagonal.

This is a convex function !...**Maximizing** over a convex set.
There can be 2^d local maxima.

Commercial Solver

We recall our 3 equivalent problems :

$$\max_{\substack{\|a\|_M \leqslant 1 \\ \|\theta - c\|_W \leqslant 1}} a^\top \theta \qquad (P_B)$$

$$\max_{\|a\|_M \leqslant 1} a^\top c + \|a\|_{W^{-1}} \, . \qquad (P_A)$$

$$\max_{\|\phi - b\|_\Lambda \leqslant 1} \|\phi\|_2 \qquad (P''_B)$$

Giving those problems to commercial solver like Gurobi does not work in practise. (see fig 5.4.4 page 162). There may be exponentially many local maxima.

How to solve (P''_B)

- Write the KKT conditions of (P''_B), and examine all possible cases.
- Discard some pathological cases.
- Use the symmetry of the ellipsoid parametrized by Λ and b. To find the best solution among the 2^d candidates.

Theorem (6.6, page 180)

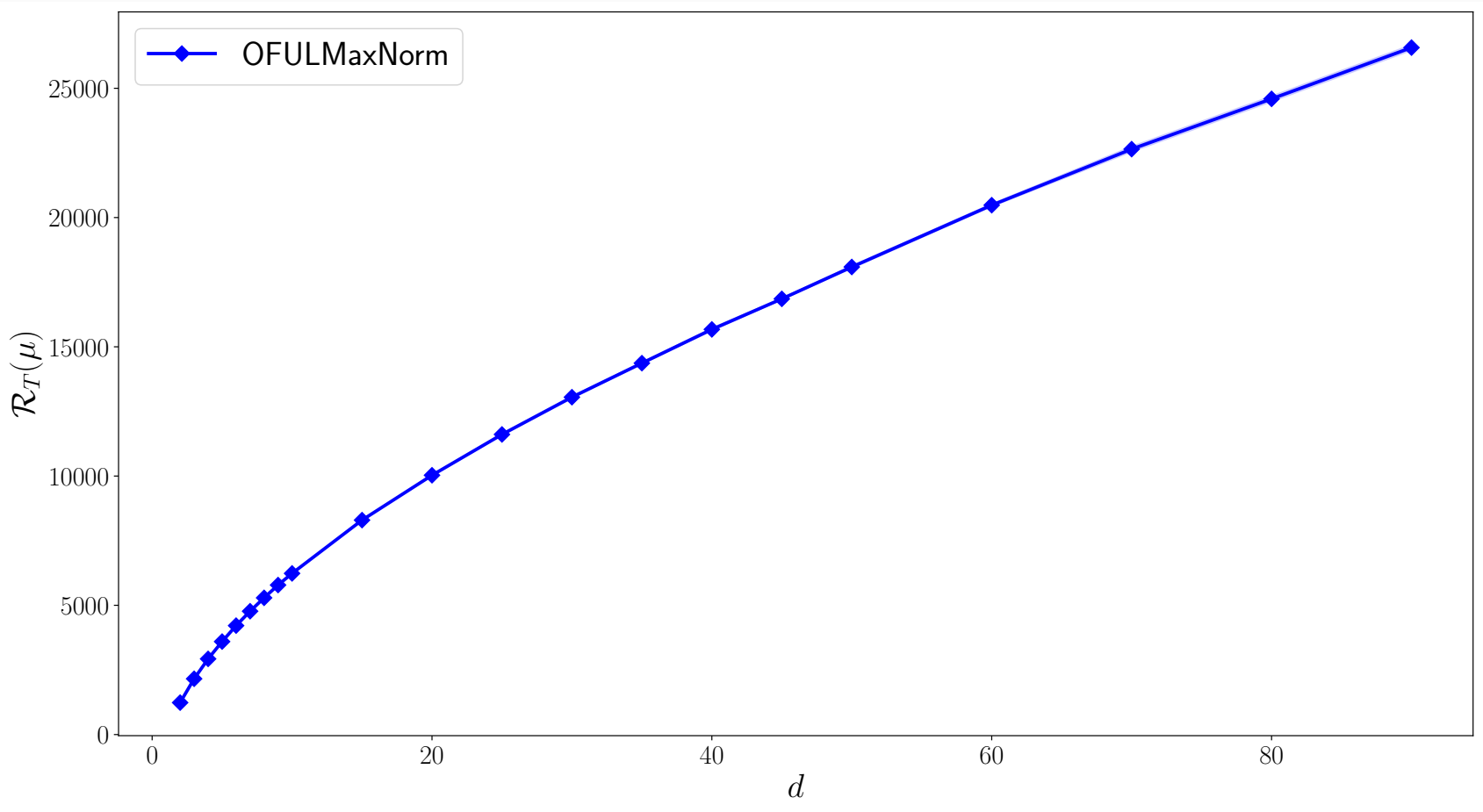
`MaxNorm` outputs $(\widehat{a}, \widehat{\theta})$ an ε-optimal solution to P_B, in the sense that $\widehat{a}^\top \widehat{\theta} \geqslant a_{\text{opt}}^\top \theta_{\text{opt}} - \varepsilon$ in time :

$$O\left(d^3 + d \log_2\left(\frac{f(\Lambda, b)}{\varepsilon}\right)\right).$$

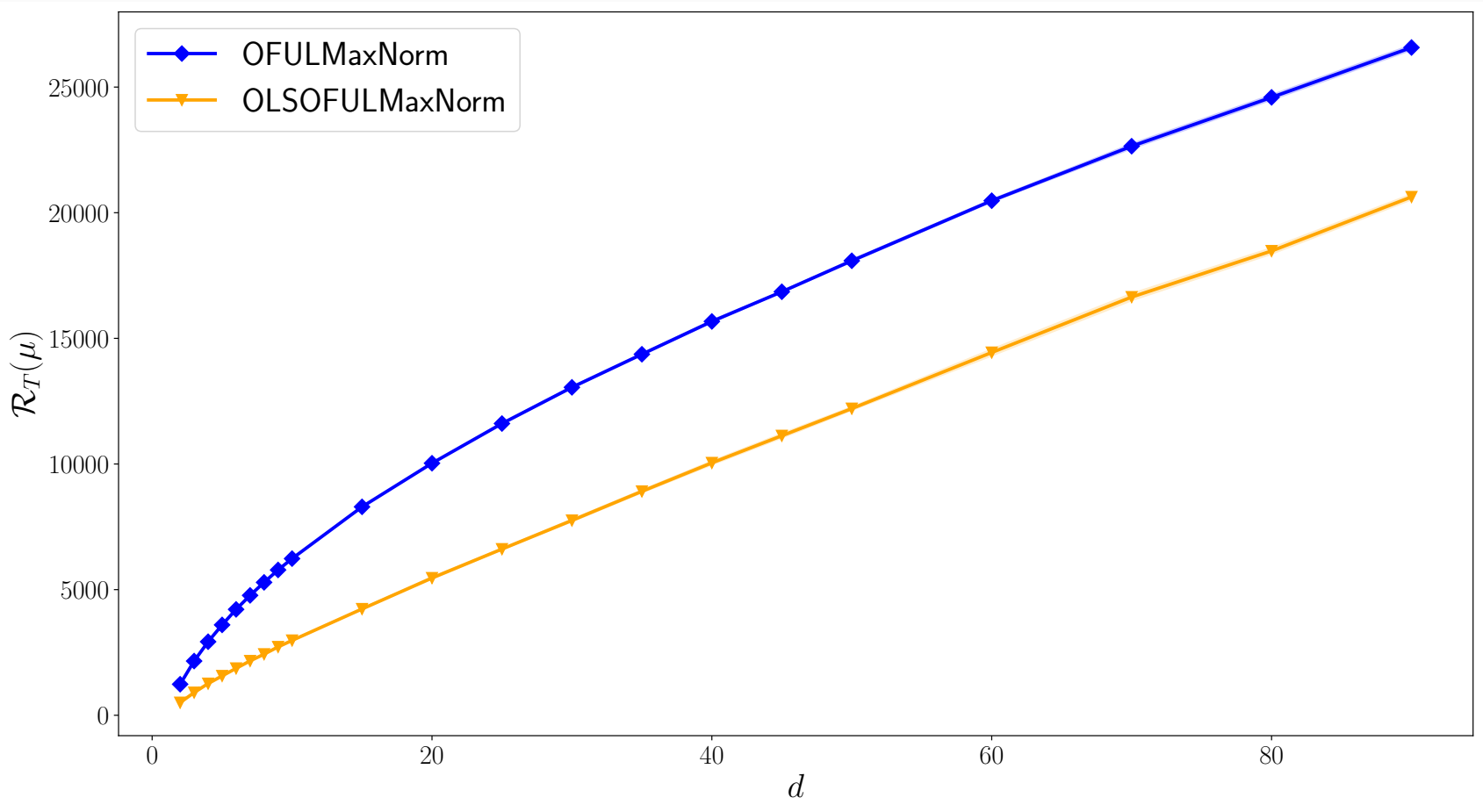
Where $f(\Lambda, b)$ is a nice function of the eigenvalues of Λ and b.

We can also efficiently solve P_A by a case disjunction and a change of variable to make it convex. Then use an interior point method.

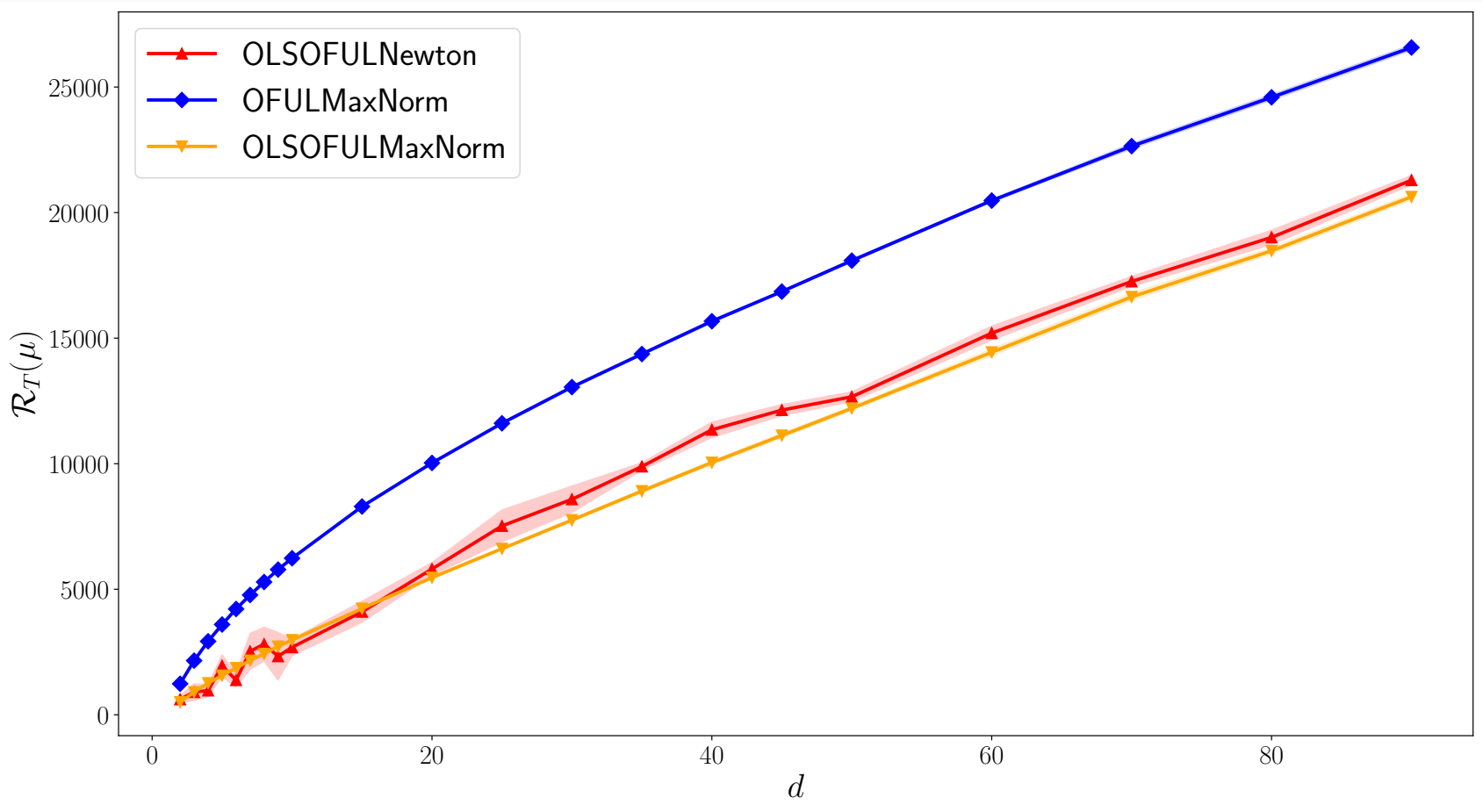
Regret and Dimension



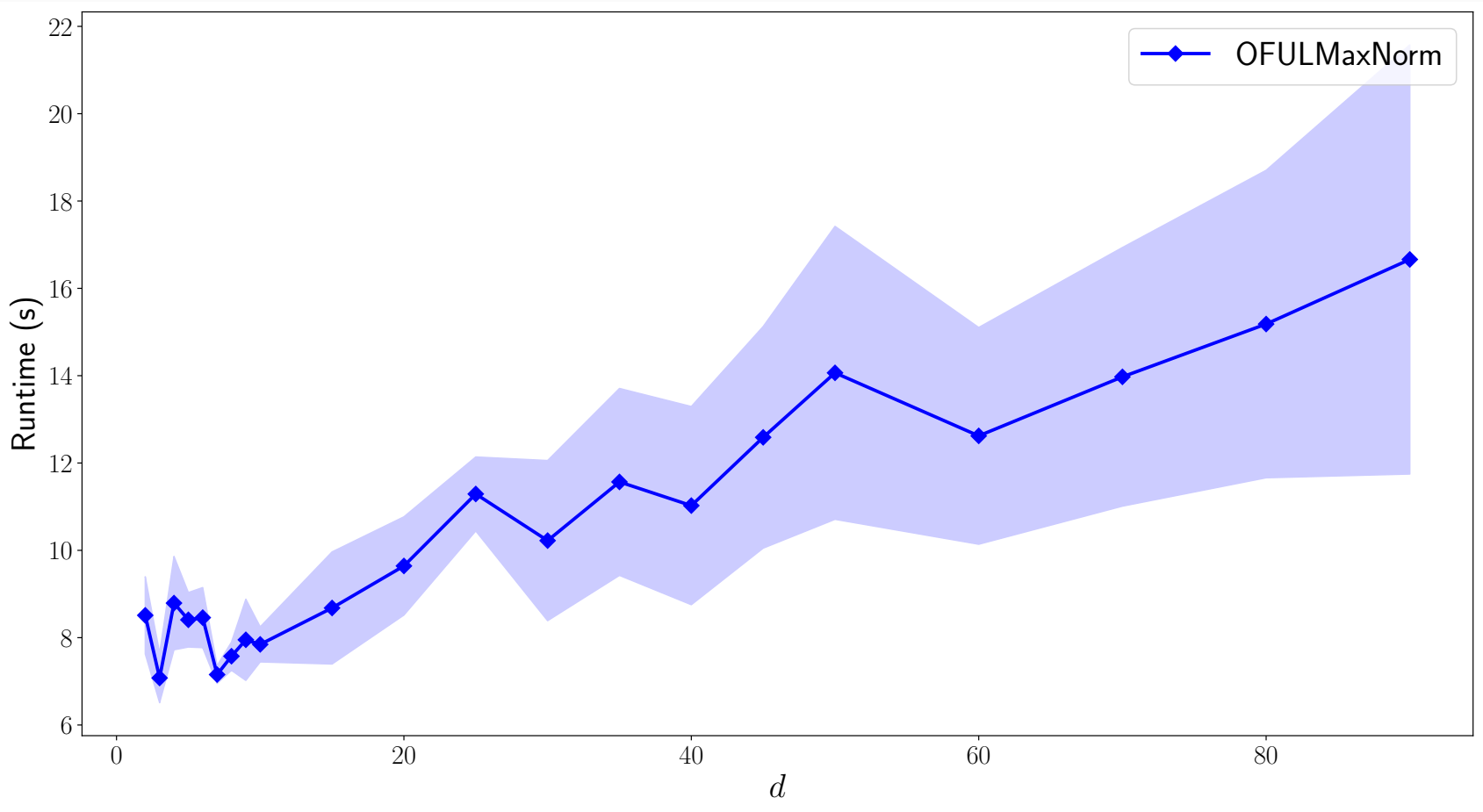
Regret and Dimension



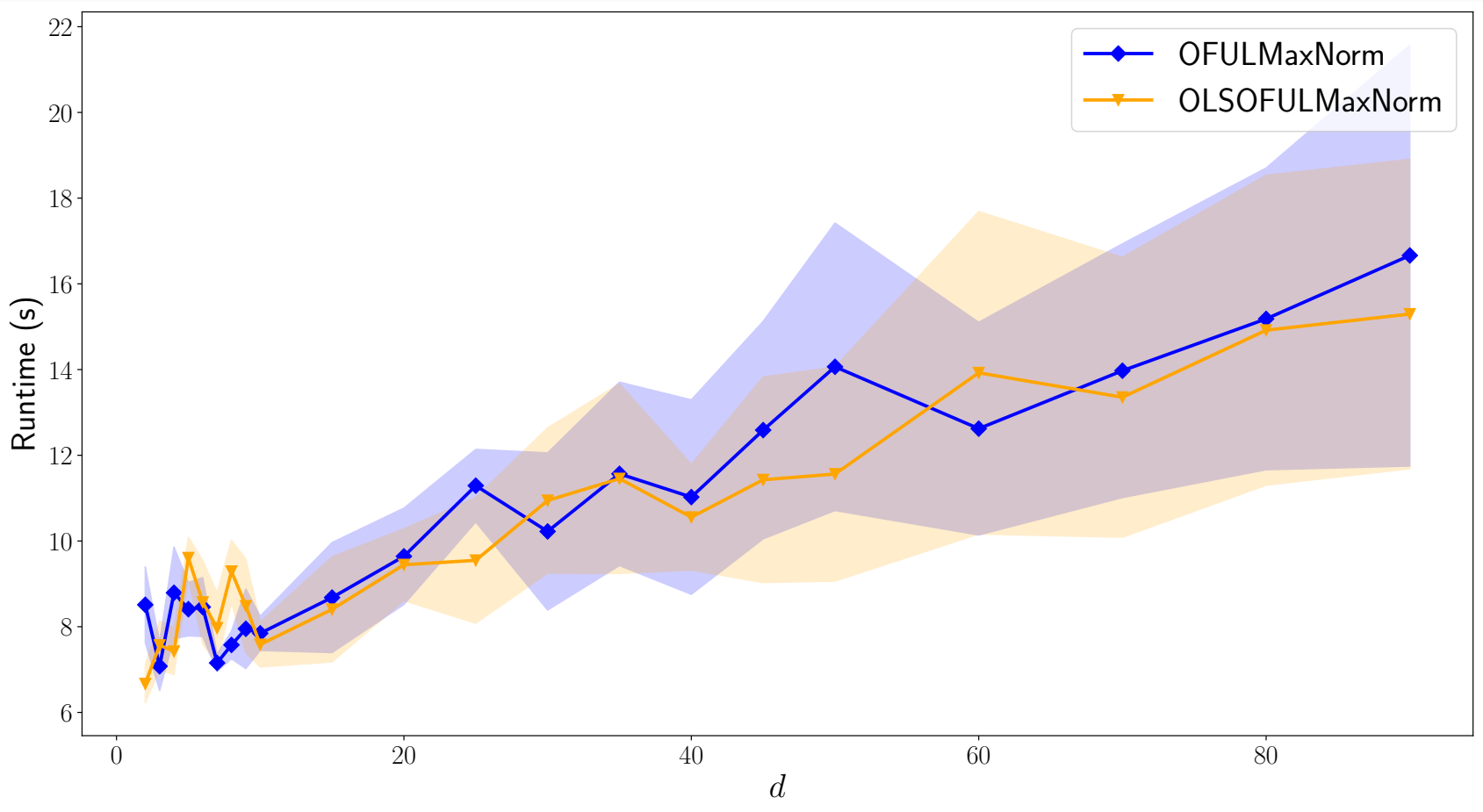
Regret and Dimension



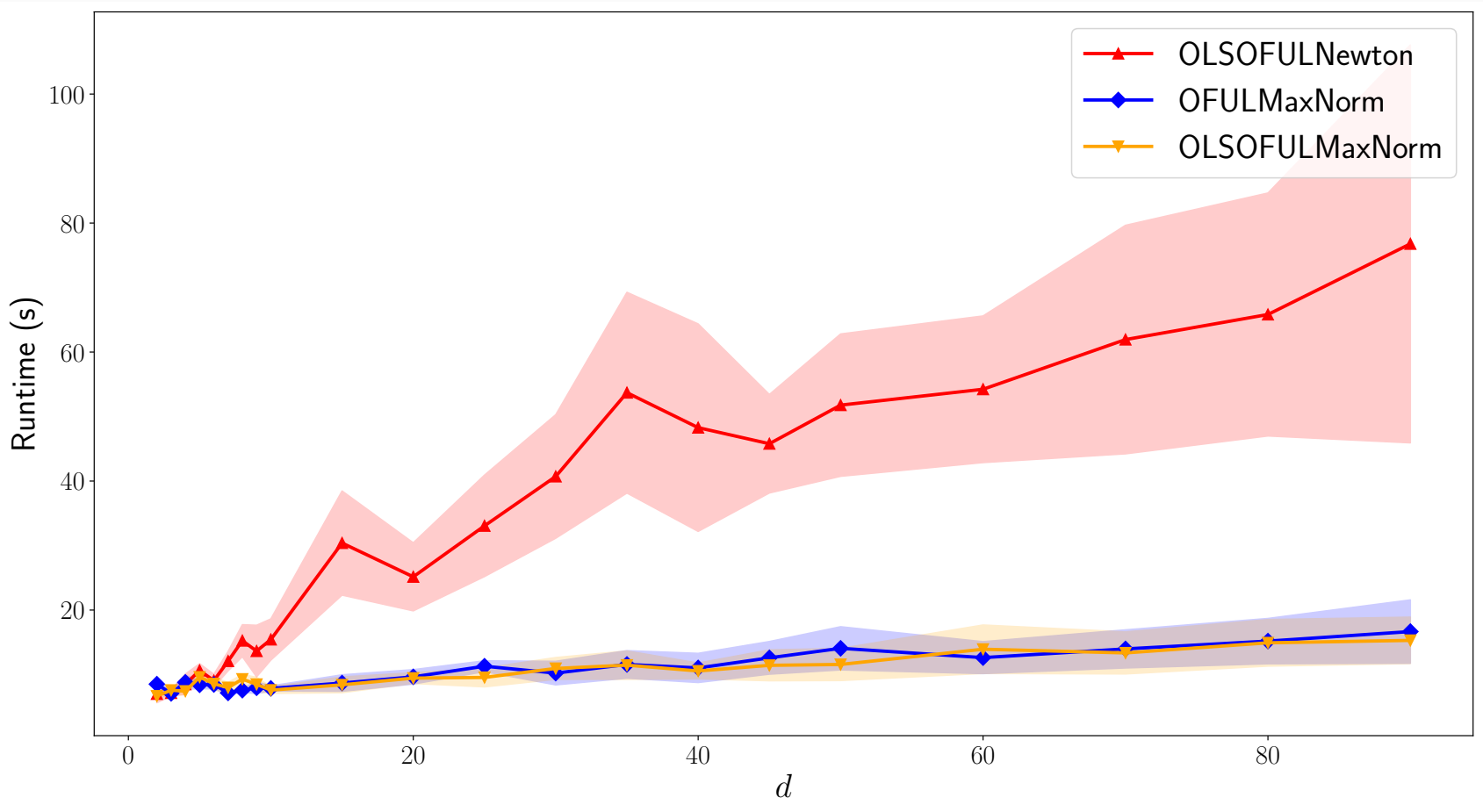
Running Time and Dimension



Running Time and Dimension



Running Time and Dimension



Optimistic Algorithms limitations

On Ellipsoids, existing optimistic algorithms have limitations :

- The confidence function β_t depends on $\|\mu\|_2$ or an upper bound on it.
- Need the knowledge of $\|\mu\|_2$.
- Regret upper bounds depend on $\|\mu\|_2$.
- Still a computation cost of d^3 at each round on Ellipsoids
- Not provably minimax optimal

Their regret is upper bounded by :

$$\lim_{T,d\to+\infty} \frac{\mathcal{R}_T}{\sigma d\sqrt{T}\ln(T)} \leqslant C\,.$$

The minimax lower bound is :

$$\mathcal{R}_T = \Omega\left(\sigma d\sqrt{T}\right)$$

EETC Algorithm

Three phases Algorithm :

- (E) Find an estimator $\widehat{B}$ of the norm $\|\mu\|_{M^{-1}}$ such that with high probability. :

$$c_1\|\mu\|_{M^{-1}} < \widehat{B} < c_2\|\mu\|_{M^{-1}}.$$

- (E) Explore on the canonical basis for $N_e = \frac{d\sigma\sqrt{T}}{\widehat{B}}$ rounds. Compute the OLS estimator $\widehat{\mu}$.
- (C) Commit to the estimated best action $a^\star(\widehat{\mu}) = \frac{M^{-1}\widehat{\mu}}{\|\widehat{\mu}\|_{M^{-1}}}$.

Phase 1 : Norm Adaptation

Time doubling strategy to estimate $\|\mu\|_{M^{-1}}$:

- Define time period of length

$$n_i = 2^i d$$

- Define confidence required after period i :

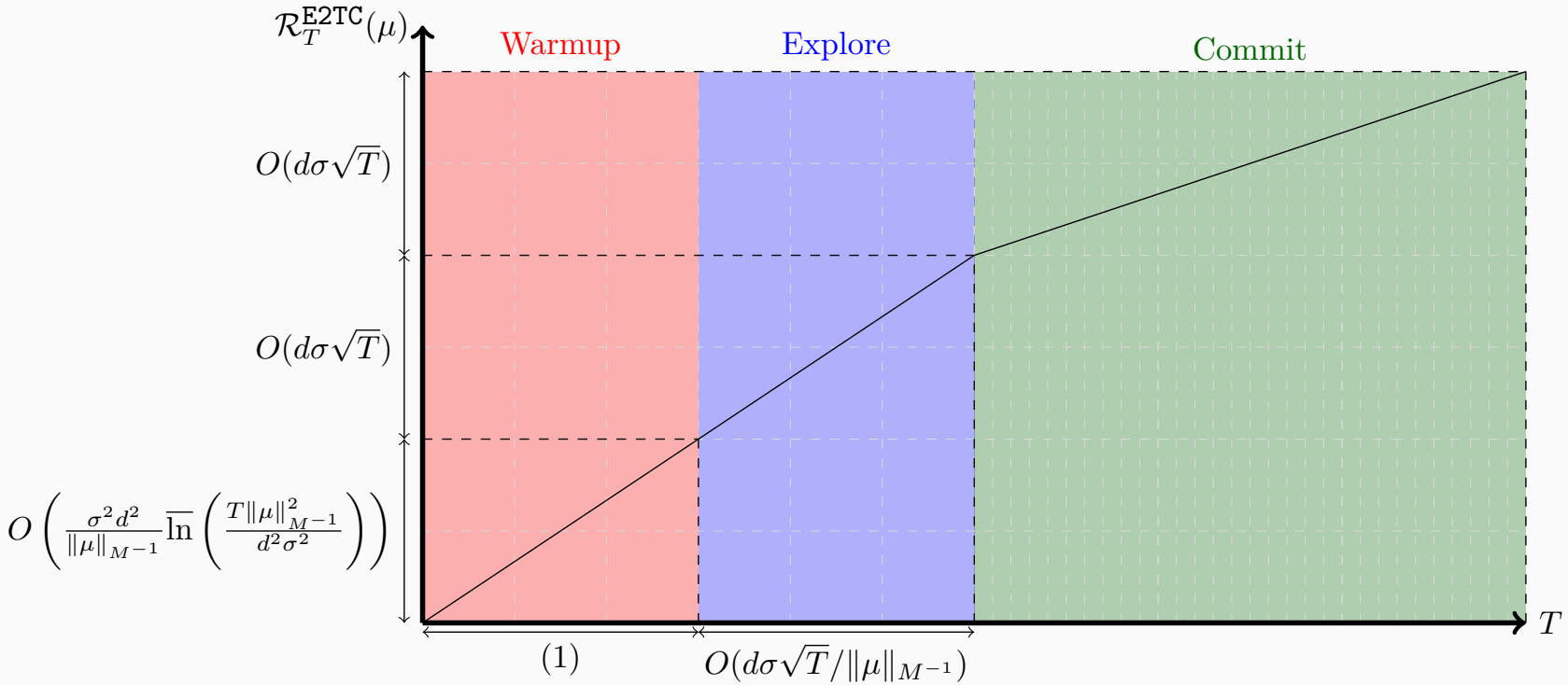
$$\delta_i = \min(dn_i/T, 1)$$

- During period i, play each $(M^{-1/2}e_j)_{j\in[d]}$ n_i/d times.
- At the end of period i, compute the OLS estimator $\widehat{\mu}_i$ and check if :

$$\|\widehat{\mu}_i\|_{M^{-1}} \geqslant 3U(\delta_i, n_i)$$

- Stop if true, and output $\widehat{B} = \|\widehat{\mu}_i\|_{M^{-1}}$.

Typical Run of EETC



E2TC Minimax Optimality

Corollary (5.7, page 158)

E2TC is locally asymptotically minimax optimal: for any $\mu \in \mathbb{R}^d$,

$$\mathcal{R}_T(\mu) = \mathcal{O}\Big(\min\left(\sigma d\sqrt{T} + d\|\mu\|_{M^{-1}}, T\|\mu\|_{M^{-1}}\right)\Big) .$$

For minimax optimality, it was really important to set increasing confidence δ_i during phase 1.

$$\delta_i = \min(dn_i/T, 1) \text{ and } n_i = 2^i d .$$

Regret and Big Norm

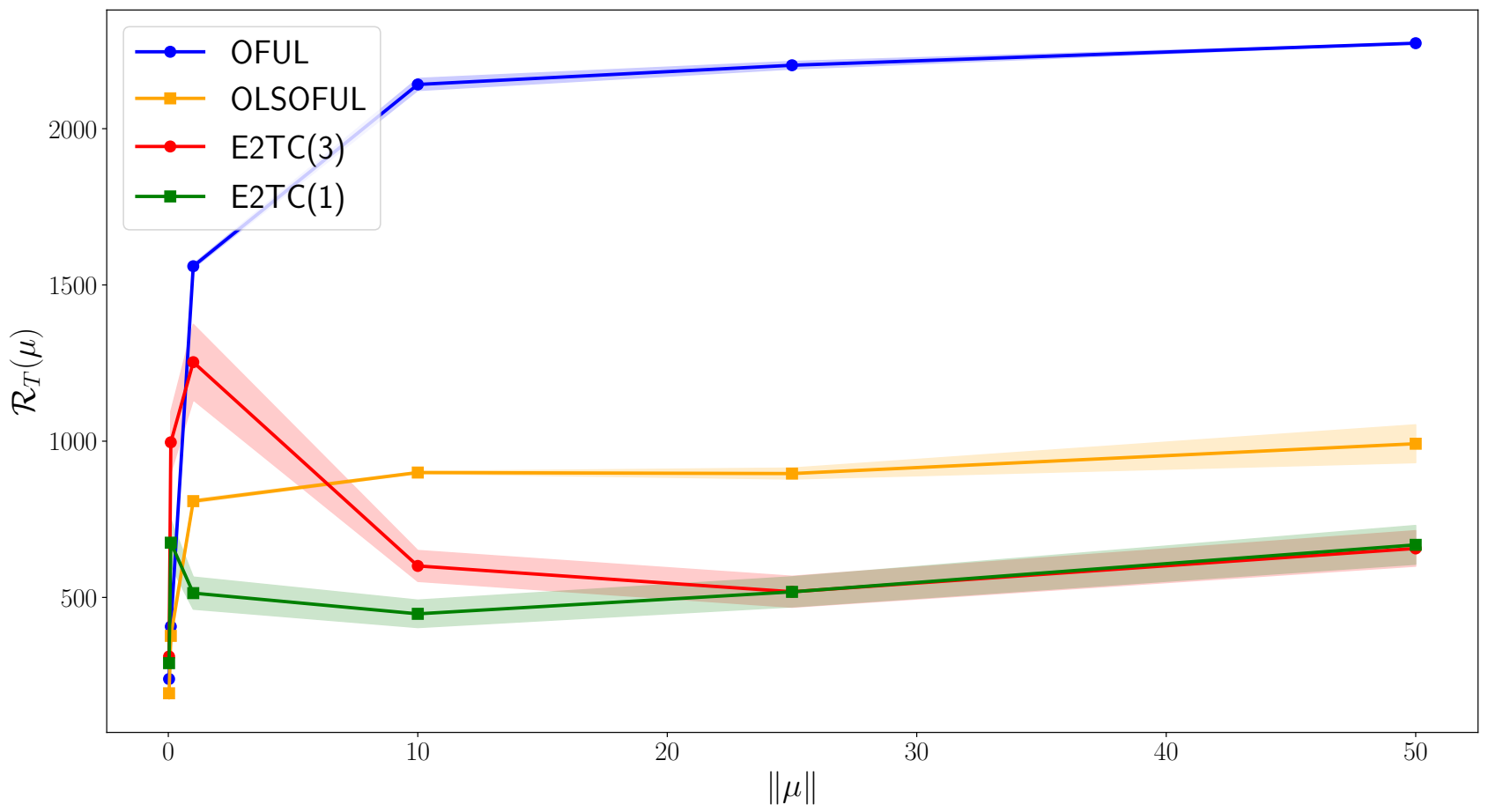


Figure 1: $T = 10000$, $\sigma = 1$, $M = I_d$, $d = 3$.

Regret and Small Norm

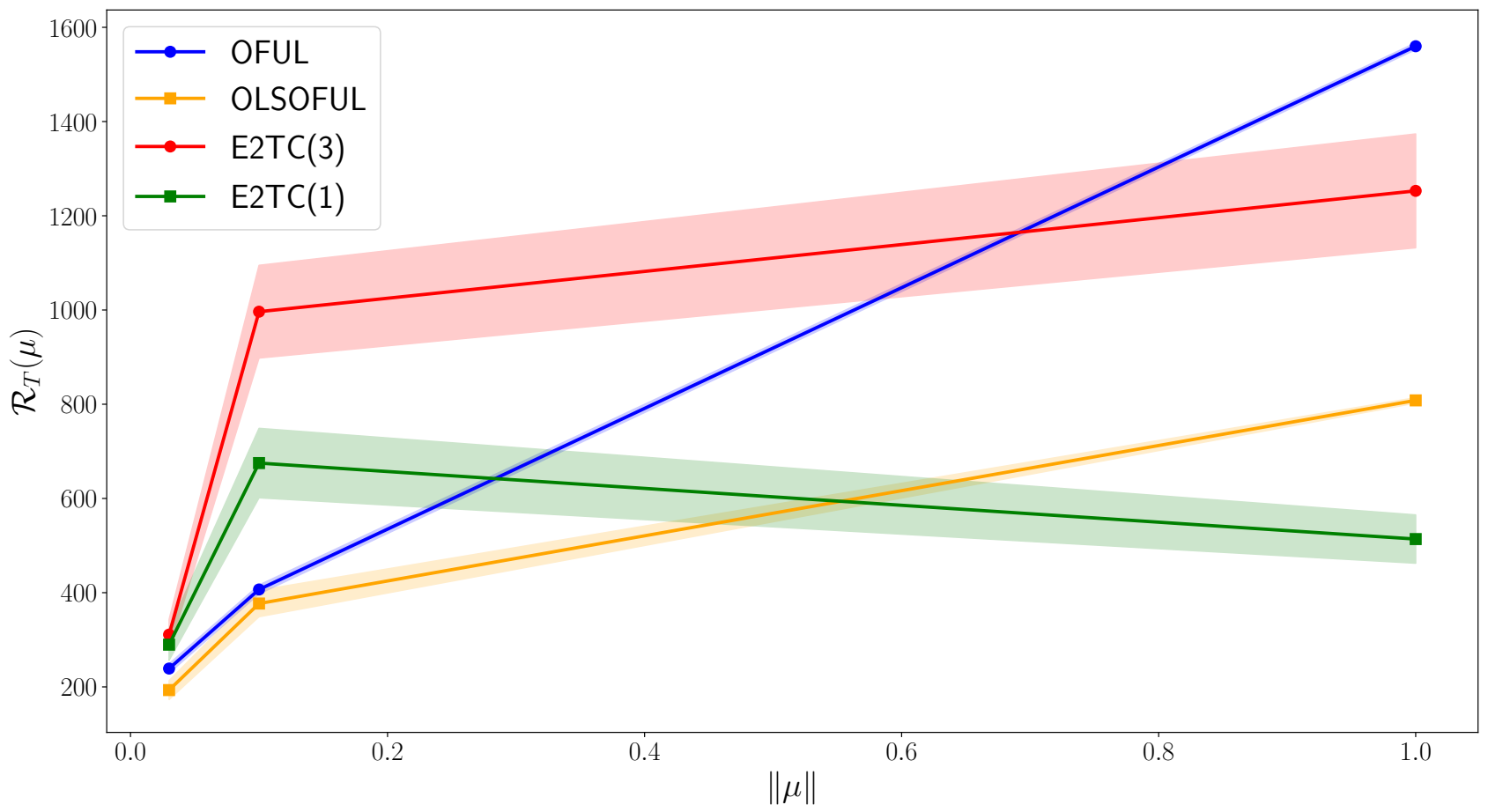


Figure 2: $T = 10000$, $\sigma = 1$, $M = I_d$, $d = 3$.

Regret and Dimension

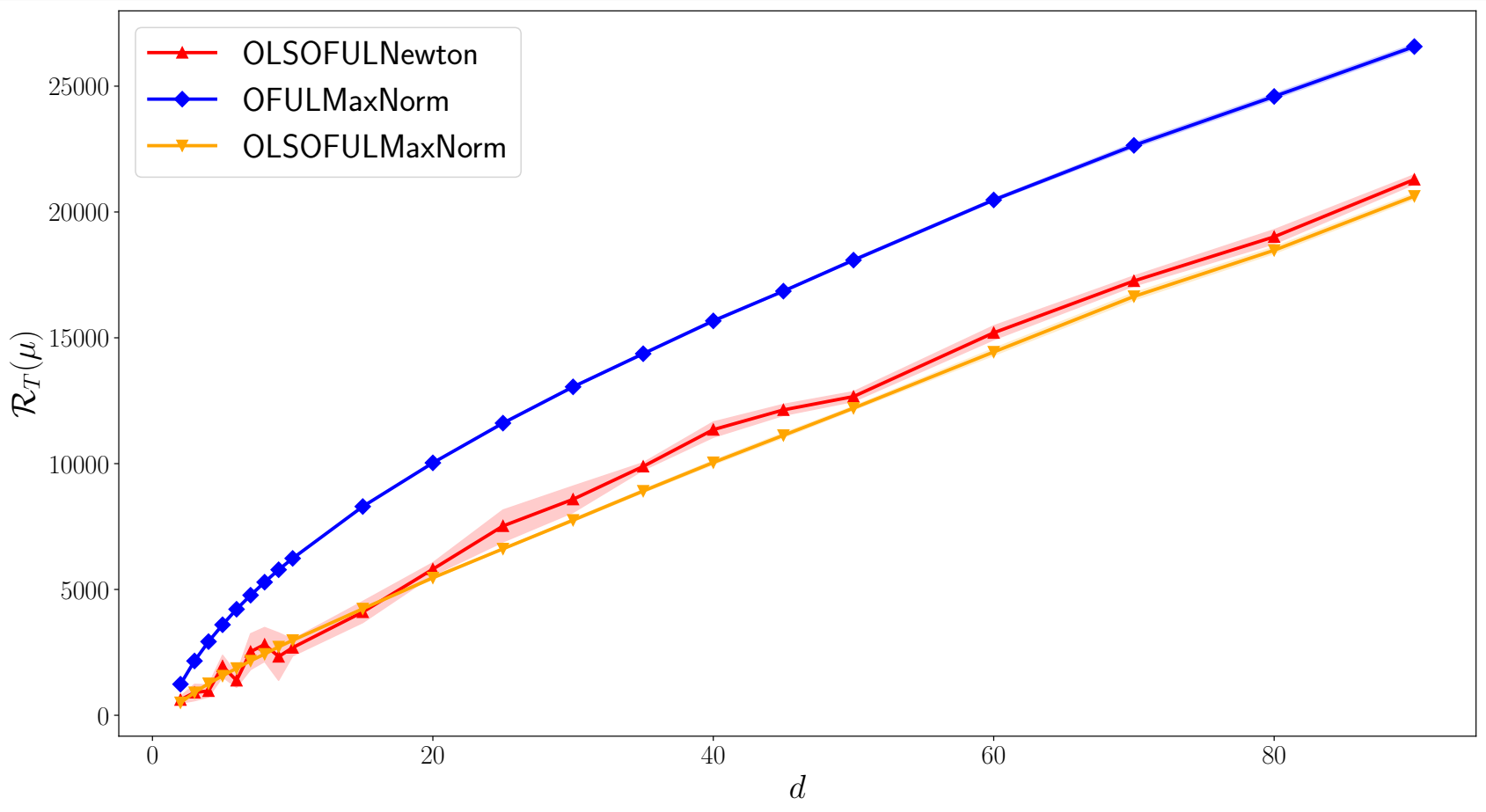


Figure 3: $T = 10000$, $\sigma = 1$, $M = I_d$, $d = 3$.

Regret and Dimension

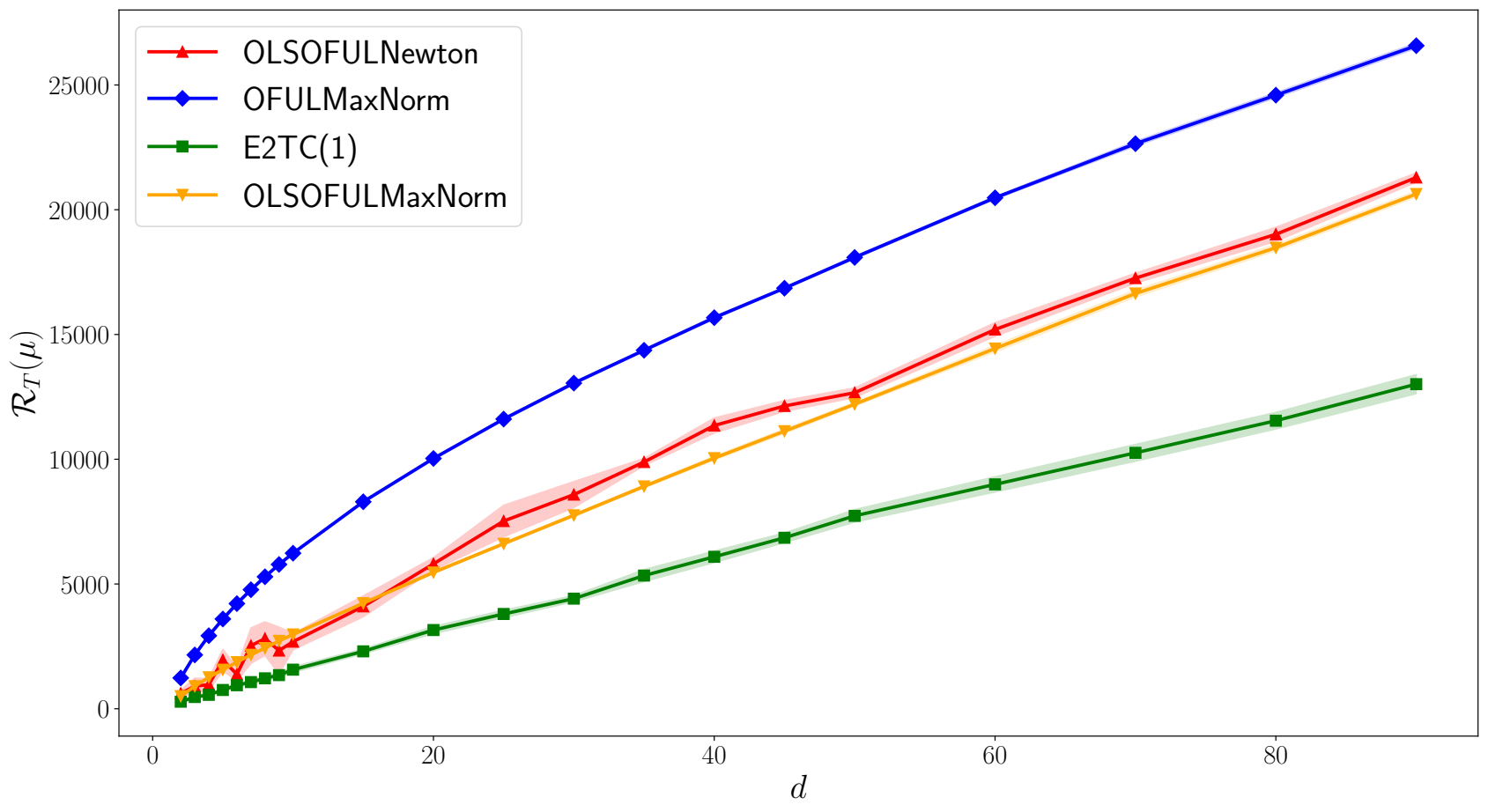
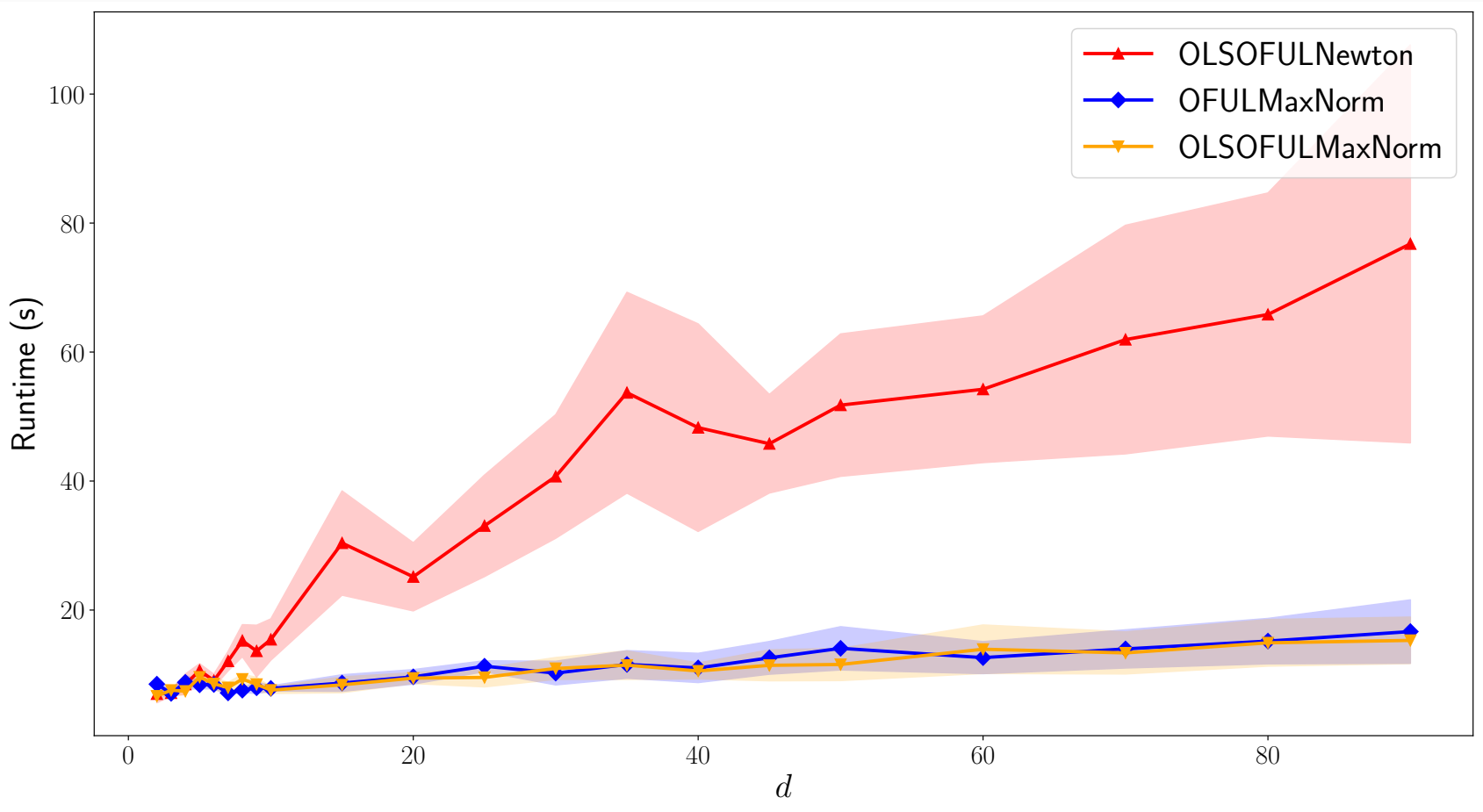
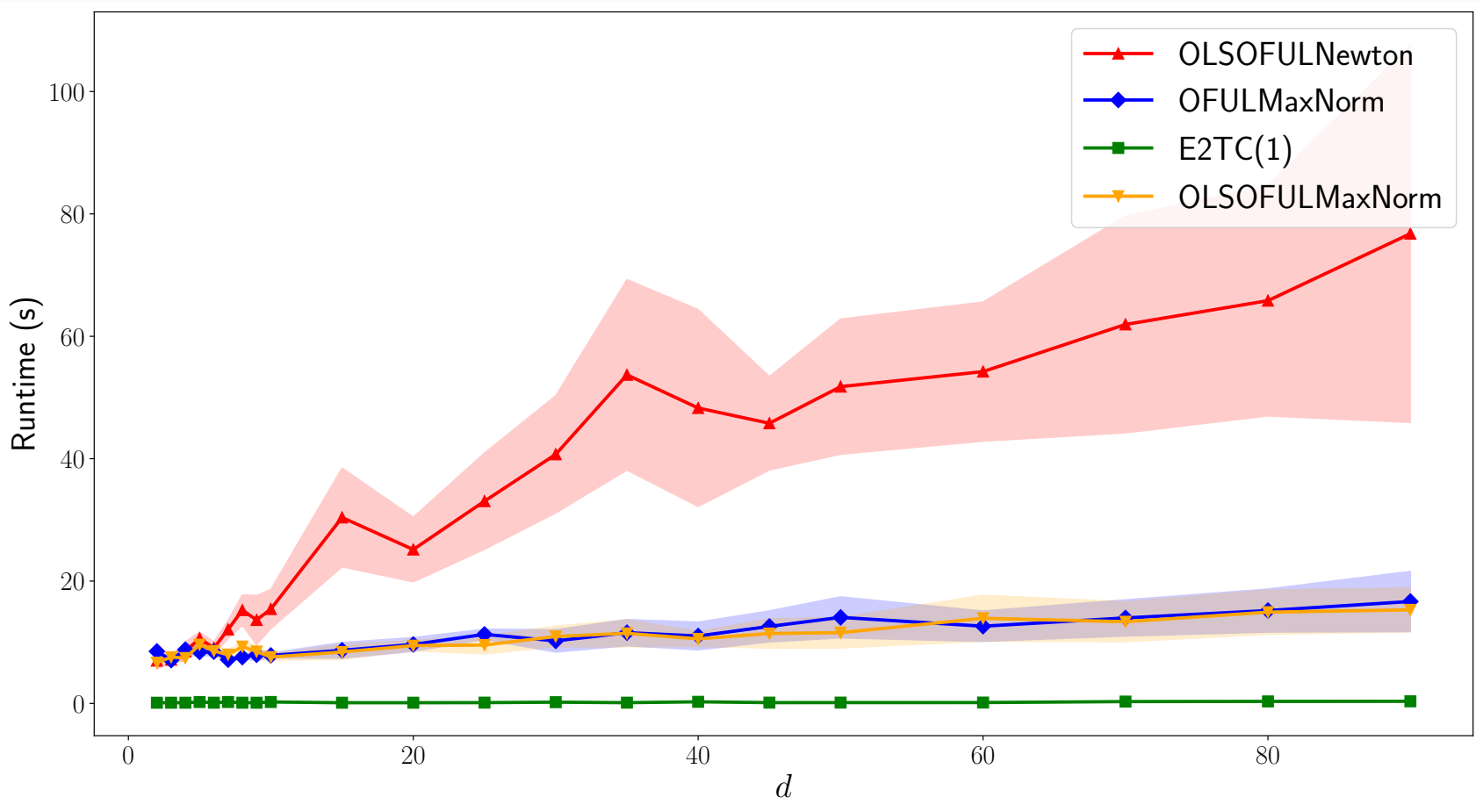


Figure 4: $T = 10000$, $\sigma = 1$, $M = I_d$, $d = 3$.

Running Time and Dimension



Running Time and Dimension



Take away Messages

On Ellipsoids :

- We provided the first locally asymptotic minimax lower bound.
- We implemented optimistic algorithms.
- Our `EETC` algorithm is the first to be locally asymptotic minimax optimal.
- `EETC` algorithm is simple and computationally efficient. It only requires a logarithmic number of matrix inversions.
- We provided a novel norm estimation procedure that can be of independent interest.

Combinatorial Semi-Bandits and Thompson Sampling

Combinatorial semi-bandits

- At time $t \in [T]$,
- A learner selects decision $A_t \in \mathcal{A} \subset \{0,1\}^d$
- Gets reward $Y_t = A_t^\top X_t$ and observes $A_t \odot X_t = (A_{t,i} X_{t,i})_{i \in [d]}$
- $(X_t)_t$ i.i.d. with mean $\mu \in \mathbb{R}^d$ and independent entries
- Goal: minimize regret

$$\mathcal{R}_T = T \underbrace{\max_{a \in \mathcal{A}} a^\top \mu}_{\text{oracle}} - \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} A_t^\top X_t\right]}_{\text{your algorithm}} .$$

- Size $m = \max_{a \in \mathcal{A}} \|a\|_1$, gap $\Delta_a = (\max_{a \in \mathcal{A}} a^\top \mu) - a^\top \mu$.

Thompson Sampling Algorithms

- Set a prior p_μ on μ
- Observations up to time t, $\mathcal{H}_{t-1} = (A_s \odot X_s, A_s)_{s<t}$,

$$A_t \in \arg\max_{a \in \mathcal{A}} a^\top \theta_t \text{ with } \theta_t \sim p_{\mu|\mathcal{H}_{t-1}}$$

Exemples of Thompson Sampling

At round t :

$$A_t \in \arg\max_{a \in \mathcal{A}} a^\top \theta_t$$

- Example 1: (`B-CTS`) Bernoulli rewards and uniform priors, then

$$\theta_t \sim \bigotimes_i^d \text{Beta}\left(N_{t-1,i}\widehat{\mu}_{t-1,i} + 1, N_{t-1,i}\left(1 - \widehat{\mu}_{t-1,i}\right) + 1\right)$$

- Example 2: (`G-CTS`) Gaussian rewards and gaussian priors, then

$$\theta_t \sim \mathcal{N}(\widehat{\mu}_{t-1}, 2\sigma^2 V_{t-1})$$
$$V_{t-1} = \text{diag}\left(\frac{1}{N_{t-1,1}}, ..., \frac{1}{N_{t-1,d}}\right)$$

Combinatorial Semi-Bandits and Thompson Sampling

Exponential regret of B-CTS

Exponential regret of B-CTS

Theorem (Zhang et al, 2021)

On some simple action set $\mathcal{A}$, for Bernoulli reward B-CTS have an exponential regret in the dimension d

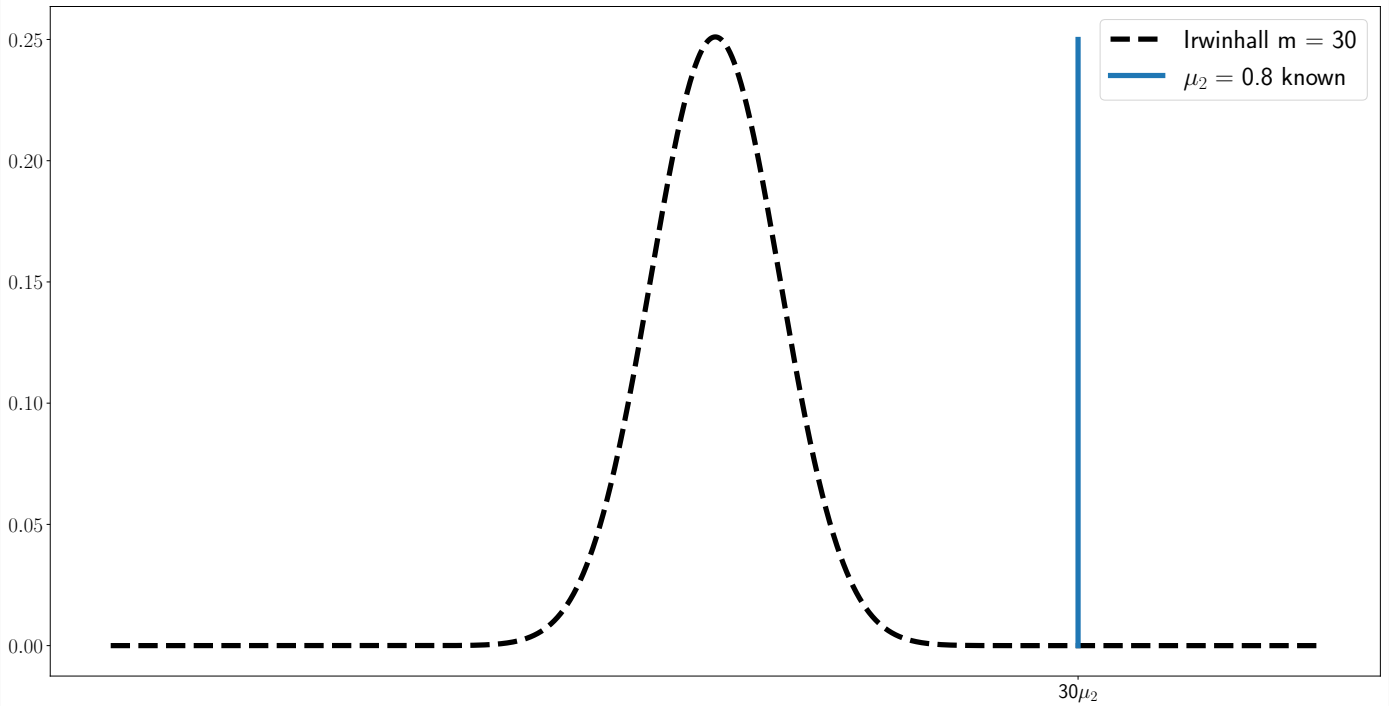
- B-CTS is too greedy, and can get "stuck" for exponentially long for time $T_0(d)$
- For $d = 20$, $T_0(d)$ is greater than the age of the universe (!)
- High dimensional phenomenon, when d is large enough, the sum of the posteriors is too concentrated around its mean.

Zhang and Combes, 2021, "On the suboptimality of thompson sampling in high dimensions"

The example

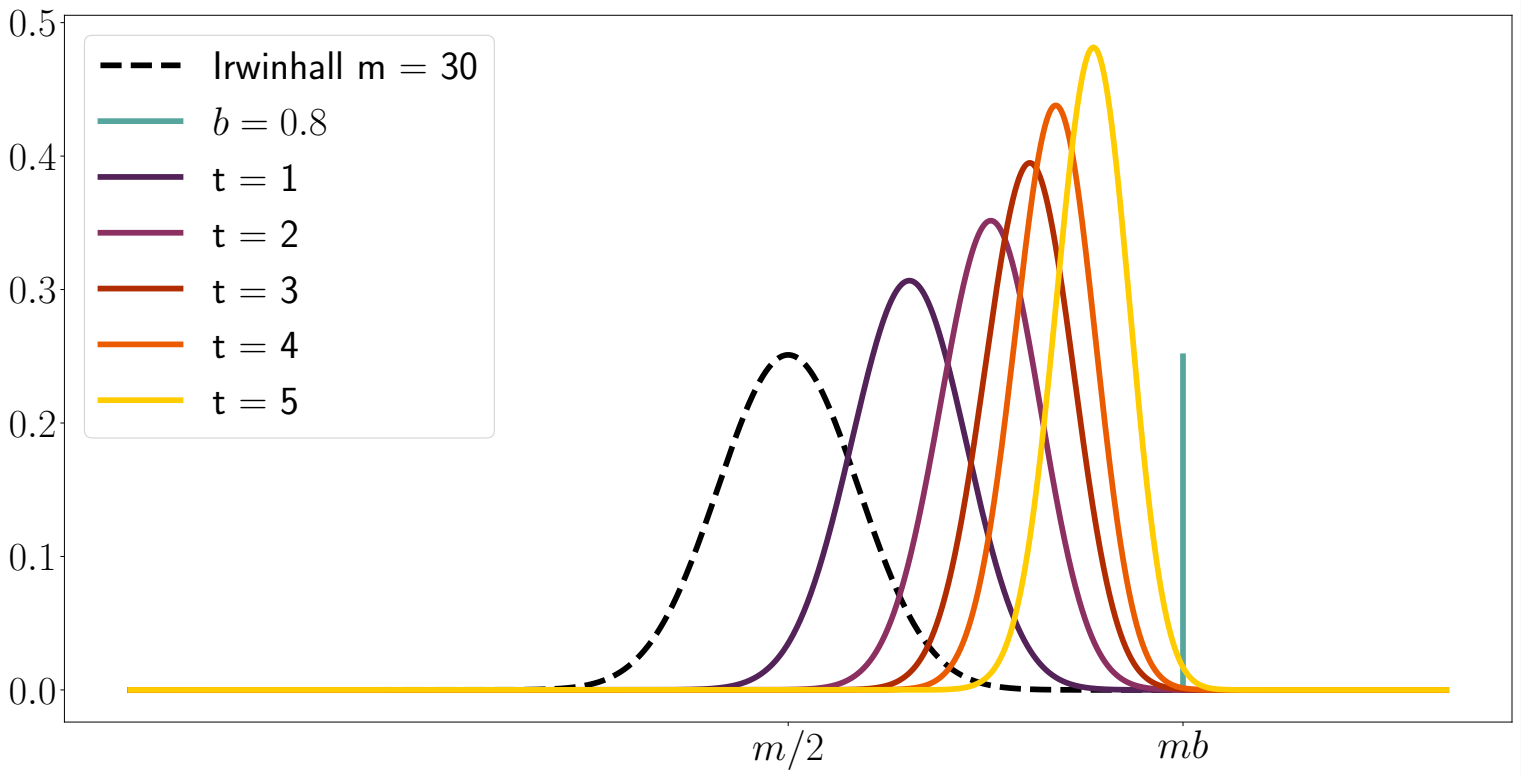
- $\mathcal{A} = \{a^1, a^2\}$ with action of size $m = d/2$
- $a^1 = (1, ..., 1, 0, ..., 0)$ and $a^2 = (0, ..., 0, 1, ..., 1)$
- $\mu = (\mu_1, ..., \mu_1, \mu_2, ..., \mu_2)$ with $\mu_1 \geqslant \mu_2 > 0.5$

Main idea : what if you start the algorithm knowing μ_2 perfectly ?



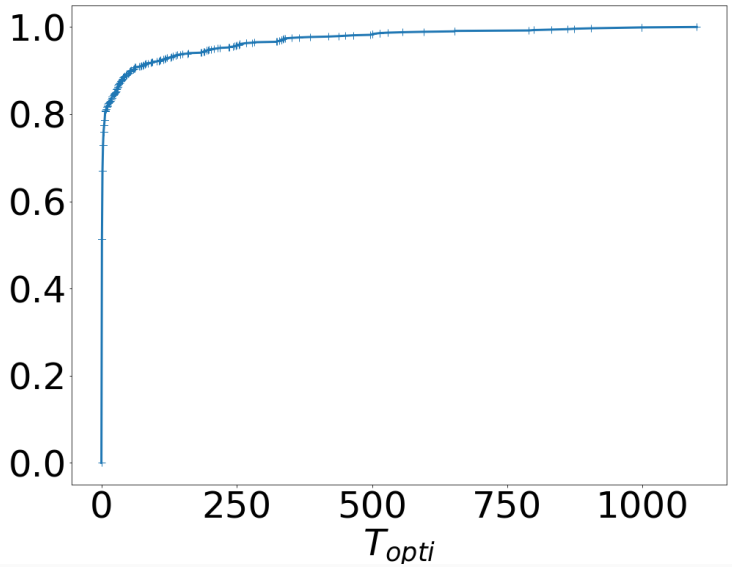
The example

The typical posterior evolution :

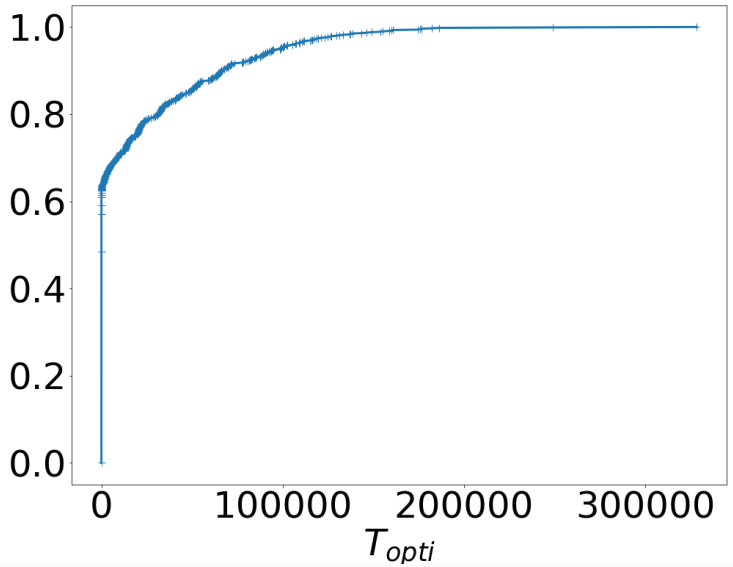


When need to make sure that the optimal action is played enough !

First Optimal play Cumulative Distribution Function



(a) $m = 6$



(b) $m = 14$

The optimal action can be not played enough for a long time when m is large.

Combinatorial Semi-Bandits and Thompson Sampling

The BG-CTS Algorithm

BG-CTS algorithm

- Sampling algorithm

$$A_t \in \arg\max a^\top \theta_t \text{ with } \theta_t \sim \mathcal{N}(\widehat{\mu}_{t-1}, 2\sigma^2 g_t V_{t-1})$$

$$V_{t-1} = \text{diag}\left(\frac{1}{N_{t-1,1}}, ..., \frac{1}{N_{t-1,d}}\right)$$

- Exploration boost

$$g(t) = (1+\lambda)\frac{\ln t + (m+2)\ln\ln t + (m/2)\ln(1+e/\lambda)}{\ln t}$$

- Similar to G-CTS for Gaussian rewards with a well chosen boost
- Implementable by linear programming over $\mathcal{A}$

Zhang and Combes, 2024, "Thompson sampling for combinatorial bandits: polynomial regret and mismatched sampling paradox"

Regret and complexity of BG-CTS

Theorem (4.1, Page 125)

Consider 1-subgaussian rewards. The regret of BG-CTS verifies

$$\mathcal{R}_T \leq C \left(d(\ln m) \ln T + d^2 m \ln \ln T \right) / \Delta_{\min} + P(m, d, 1/\Delta_{\min}) \,,$$

with C a universal constant and P a polynomial.

- Valid for Gaussian, Bernoulli, bounded etc.
- If linear programming over $\mathcal{A}$ is polynomial then polynomial complexity.
- Best known polynomial (complexity, regret) algorithm for asymptotic regret for general action set.

Rationale: self normalized concentration inequalities

- Why is the "correct" confidence boost g_t ?
- Self-normalized concentration inequality, choose g_t such that

$$\mathbb{P}\left(\sup_{s\leq t}\frac{|a^{\star\top}(\widehat{\mu}_s-\mu)|}{\sqrt{a^{\star\top}V_s a^\star}}\geq\sqrt{2\ln(t)g_t}\right)\leqslant\frac{1}{t(\ln t)^2}$$

- The proof relies on showing that with high probability :

$$\sum_s^t \mathbf{1}\left\{a^{\star\top}\theta_s\geqslant a^{\star\top}\mu\right\}>\lfloor ct^\alpha\rfloor$$

with a constants $\alpha>0$ and $c>0$ for all $t\in[T]$.

This ensures enough exploration of the optimal action !

Degenne and Perchet, 2016, "Combinatorial semi-bandit with known covariance"

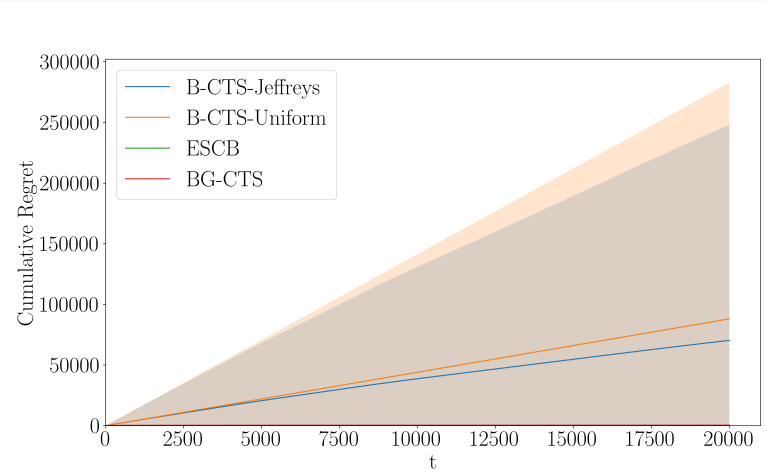
Mismatched sampling paradox

Consider a problem with Bernoulli rewards and parameters in $[0,1]^d$.

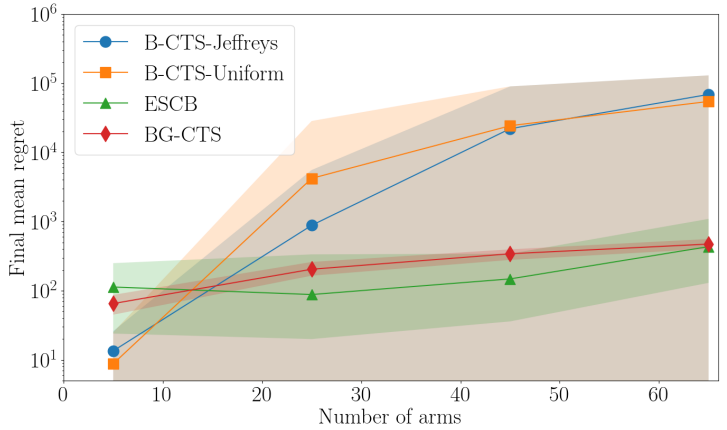
- Learner 1 knows the rewards distribution and the support $[0,1]^d$, uses a uniform (or Jeffreys) prior over $[0,1]^d$ and Bernoulli likelihood (B-CTS)
- Learner 2 does not know the rewards distribution and the support $[0,1]^d$, uses a Gaussian prior and Gaussian likelihood over $\mathbb{R}^d$ and a boost (BG-CTS)

Paradox: Learner 1 performs **exponentially worse** than Learner 2

Regret experiments of BG-CTS vs B-CTS



(a) Average regret over time



(b) Final regret as a function of m

Take away messages

- Sampling algorithms are fine, but posterior sampling sometimes does not work.
- Putting mass outside the parameter space can make things exponentially better !
- This Bayesian rationale of predicting using the posterior distribution is not universal for online problems.
- One may need to put prior on the reward of the action and not on μ.

Conclusion

- How to generalize EETC, LinUCB implementation to more complex action set ?
- Have a version of those algorithms for generalize linear bandits ?
- What is the relation between the regret of Thompson Sampling and the hardness of optimistic algorithms ?

Publications

- Zhang and Combes, "On the Suboptimality of Thompson Sampling in High Dimensions" (NeurIPS 2021)
- Zhang and Combes, "Thompson Sampling for Combinatorial Bandits: Polynomial Regret and Mismatched Sampling Paradox" (NeurIPS 2024)
- Zhang, Hadiji and Combes, "Linear Bandits on Ellipsoids: Minimax Optimal Algorithms" (COLT 2025)
- Zhang, Hadiji and Combes, "Tractable Instances of Bilinear Maximization: Implementing LinUCB on Ellipsoids" (Under review)

Thank you for your attention !