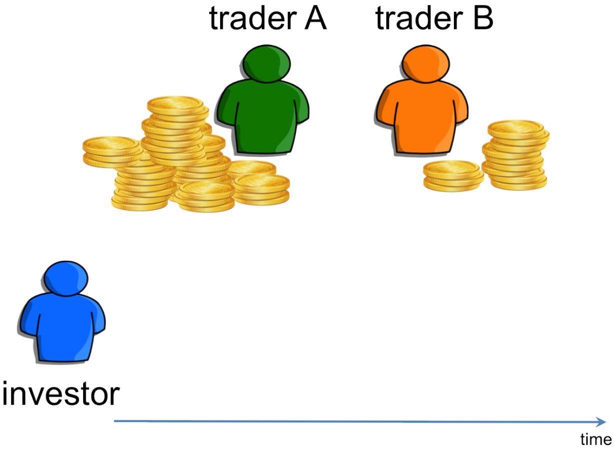
Optimism and Randomness in Linear Multi-Armed Bandit

Alessandro LAZARIC (*Inria-Lille*)

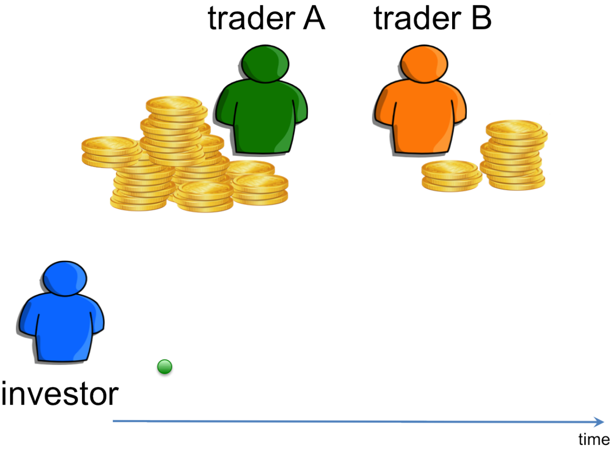
International Conference on Monte-Carlo Techniques

Paris, July 5-8, 2016

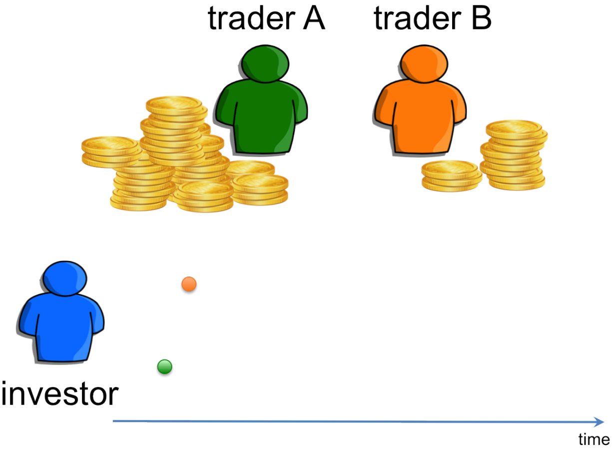
SequeL – Inria Lille



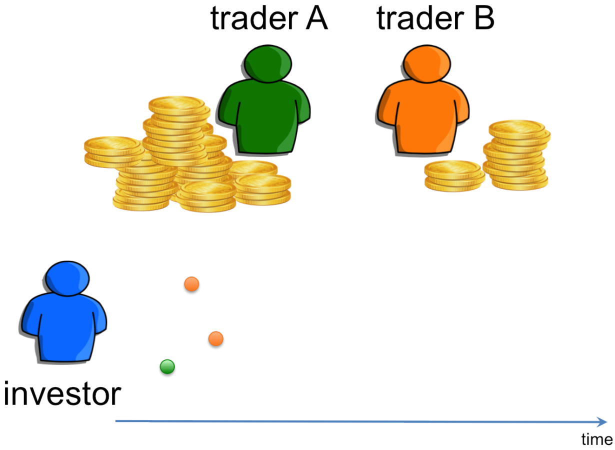
Example from *[Lamberton, Pagès, Tarrès, 2004]*



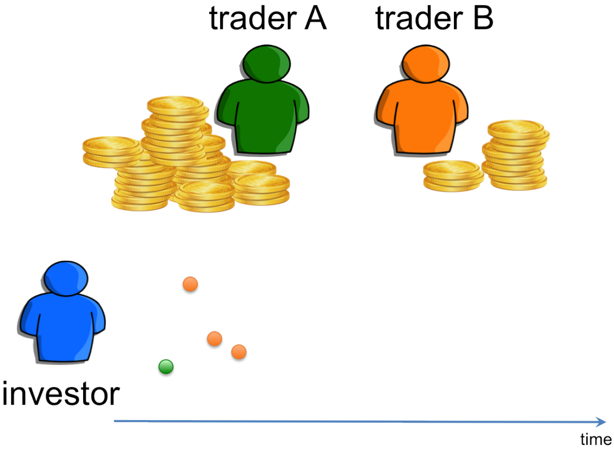
Example from *[Lamberton, Pagès, Tarrès, 2004]*



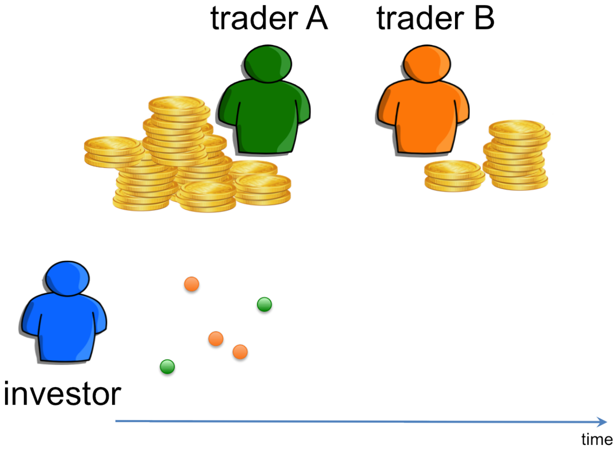
Example from *[Lamberton, Pagès, Tarrès, 2004]*



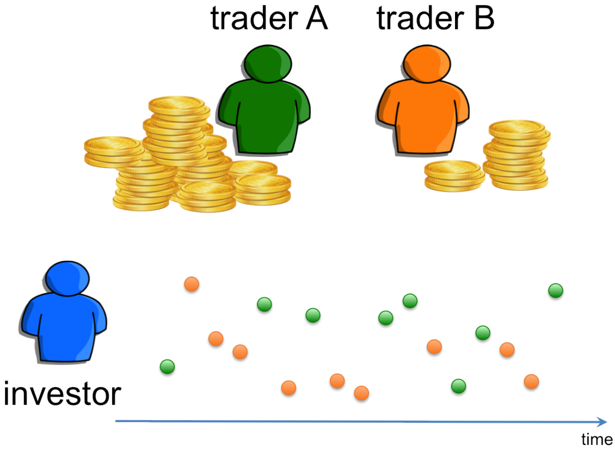
Example from *[Lamberton, Pagès, Tarrès, 2004]*



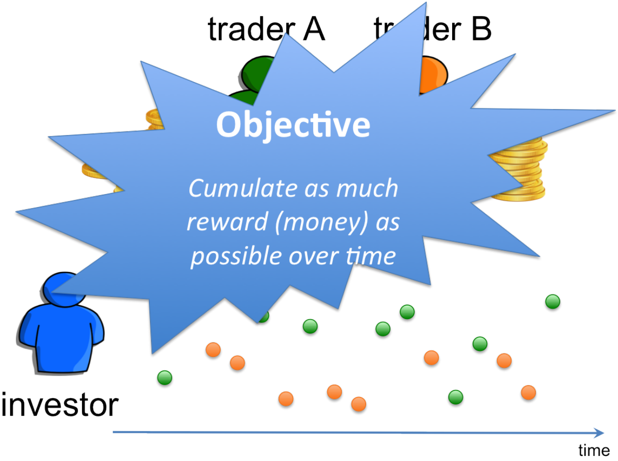
Example from *[Lamberton, Pagès, Tarrès, 2004]*



Example from *[Lamberton, Pagès, Tarrès, 2004]*



Example from *[Lamberton, Pagès, Tarrès, 2004]*



Example from *[Lamberton, Pagès, Tarrès, 2004]*

Applications of Multi-armed Bandit

- Recommendation systems
- Clinical trials
- Packet routing, cognitive radios
- Trading
- Education
- ...

Outline

The Linear Bandit Framework

The setting:

- Set of arms $\mathcal{X} \subset \mathbb{R}^d$

The Linear Bandit Framework

The setting:

- Set of arms $\mathcal{X} \subset \mathbb{R}^d$
- Reward of arm $x \in \mathcal{X}$

$$r(x) = x^\mathsf{T}\theta^\star + \xi \qquad \textit{(standard linear regression model)}$$

with $\theta^\star \in \mathbb{R}^d$ unknown and ξ a zero-mean, sub-Gaussian noise

The Linear Bandit Framework

The setting:

- Set of arms $\mathcal{X} \subset \mathbb{R}^d$
- Reward of arm $x \in \mathcal{X}$

$$r(x) = x^\mathsf{T}\theta^\star + \xi \qquad \textit{(standard linear regression model)}$$

with $\theta^\star \in \mathbb{R}^d$ unknown and ξ a zero-mean, sub-Gaussian noise

- Best arm and best value for any parameter θ

$$x^\star(\theta) = \arg\max_{x\in\mathcal{X}} x^\mathsf{T}\theta; \quad J(\theta) = \max_{x\in\mathcal{X}} x^\mathsf{T}\theta$$

The Linear Bandit Framework

The setting:

- Set of arms $\mathcal{X} \subset \mathbb{R}^d$
- Reward of arm $x \in \mathcal{X}$

$$r(x) = x^\mathsf{T}\theta^\star + \xi \qquad \textit{(standard linear regression model)}$$

with $\theta^\star \in \mathbb{R}^d$ unknown and ξ a zero-mean, sub-Gaussian noise

- Best arm and best value for any parameter θ

$$x^\star(\theta) = \arg\max_{x\in\mathcal{X}} x^\mathsf{T}\theta; \quad J(\theta) = \max_{x\in\mathcal{X}} x^\mathsf{T}\theta$$

- *Optimal strategy*: select arm $x^\star(\theta^*)$ *(constrained linear optimization)*

The Linear Bandit Framework

The learning problem:

- Finite horizon T

The Linear Bandit Framework

The learning problem:

- Finite horizon T
- Select an arm x_t at each step $t = 1, \ldots, T$

The Linear Bandit Framework

The learning problem:

- Finite horizon T
- Select an arm x_t at each step $t = 1, \ldots, T$
- Cumulate as much reward as possible

$$\sum_{t=1}^{T} x_t^{\mathsf{T}} \theta^* \qquad \textit{(explore-exploit trade-off)}$$

The Linear Bandit Framework

The learning problem:

- Finite horizon T
- Select an arm x_t at each step $t = 1, \ldots, T$
- Cumulate as much reward as possible

$$\sum_{t=1}^{T} x_t^\mathsf{T} \theta^* \qquad \textit{(explore-exploit trade-off)}$$

- *Equivalently*: minimize the regret

$$R(T) = \sum_{t=1}^{T} \left(x^\star (\theta^\star)^\mathsf{T} \theta^\star - x_t^\mathsf{T} \theta^\star \right)$$

The Linear Bandit Framework

The learning problem:

- Finite horizon T
- Select an arm x_t at each step $t = 1, \ldots, T$
- Cumulate as much reward as possible

$$\sum_{t=1}^{T} x_t^{\mathsf{T}} \theta^* \qquad \textit{(explore-exploit trade-off)}$$

- *Equivalently*: minimize the regret

$$R(T) = \sum_{t=1}^{T} \left(x^\star (\theta^\star)^{\mathsf{T}} \theta^\star - x_t^{\mathsf{T}} \theta^\star \right)$$

$\Rightarrow$ a good learning algorithm should have $o(T)$ regret!

The Linear Bandit Framework

The core ingredient: *regularized least-squares estimator*

- Given samples $\{(x_1, r_1), (x_2, r_2), \ldots, (x_{t-1}, r_{t-1})\}$ compute

$$\widehat{\theta}_t = \arg\min_{\theta \in \mathbb{R}^d} \sum_{s=1}^{t-1} \big(r_s - x_s^{\mathsf{T}}\theta\big)^2 + \lambda \|\theta\| \qquad \textit{(λ regularization parameter)}$$

The Linear Bandit Framework

The core ingredient: *regularized least-squares estimator*

- Given samples $\{(x_1, r_1), (x_2, r_2), \ldots, (x_{t-1}, r_{t-1})\}$ compute

$$\widehat{\theta}_t = \arg\min_{\theta \in \mathbb{R}^d} \sum_{s=1}^{t-1} \left(r_s - x_s^{\mathsf{T}}\theta\right)^2 + \lambda \|\theta\| \qquad \textit{(λ regularization parameter)}$$

- In closed form

$$V_t = \lambda I + \sum_{s=1}^{t-1} x_s x_s^{\mathsf{T}} \quad \textit{(design matrix)} \qquad \widehat{\theta}_t = V_t^{-1} \sum_{s=1}^{t-1} x_s r_s \quad \textit{(RLS estimator)}$$

The Linear Bandit Framework

The core ingredient: *regularized least-squares estimator*

- Given samples $\{(x_1, r_1), (x_2, r_2), \ldots, (x_{t-1}, r_{t-1})\}$ compute

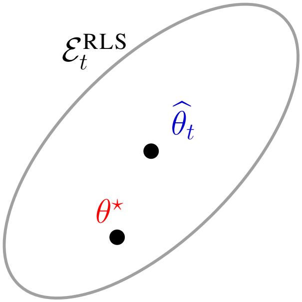
$$\widehat{\theta}_t = \arg\min_{\theta \in \mathbb{R}^d} \sum_{s=1}^{t-1} \big(r_s - x_s^{\mathsf{T}}\theta\big)^2 + \lambda \|\theta\| \qquad (\lambda \textit{ regularization parameter})$$

- In closed form

$$V_t = \lambda I + \sum_{s=1}^{t-1} x_s x_s^{\mathsf{T}} \quad \textit{(design matrix)} \qquad \widehat{\theta}_t = V_t^{-1} \sum_{s=1}^{t-1} x_s r_s \quad \textit{(RLS estimator)}$$

- Guarantees (w.h.p.) *(Gauss-Markov confidence interval for martinagales)*
  - *(estimation)* $\|\widehat{\theta}_t - \theta^\star\|_{V_t} \leq \sqrt{d \log(t/\delta)}$
  - *(prediction)* $|x^{\mathsf{T}}(\widehat{\theta}_t - \theta^\star)| \leq \|x\|_{V_t^{-1}} \sqrt{d \log(t/\delta)}$

Confidence Ellipsoid



$$\mathcal{E}_t^{\mathrm{RLS}} = \left\{\theta \in \mathbb{R}^d \mid ||\theta - \widehat{\theta}_t||_{V_t} \leq \sqrt{d \log(1/\delta)}\right\}$$

Outline

Optimism in Face of Uncertainty

- *Exploit*: given past observations, compute $\widehat{\theta}_t$ and confidence ellipsoid $\mathcal{E}_t^{\mathrm{RLS}}$

Optimism in Face of Uncertainty

- *Exploit*: given past observations, compute $\widehat{\theta}_t$ and confidence ellipsoid $\mathcal{E}_t^{\mathrm{RLS}}$
- *Explore*: given past observations, any $\theta \in \mathcal{E}_t^{\mathrm{RLS}}$ could be $\theta^\star$

Optimism in Face of Uncertainty

- *Exploit*: given past observations, compute $\widehat{\theta}_t$ and confidence ellipsoid $\mathcal{E}_t^{\mathrm{RLS}}$
- *Explore*: given past observations, any $\theta \in \mathcal{E}_t^{\mathrm{RLS}}$ could be $\theta^\star$
- *Optimism*: trade-off exploration and exploitation by taking the most "optimistic" θ *compatible* with current estimates

$$\widetilde{\theta}_t = \arg\max_{\theta \in \mathcal{E}_t^{\mathrm{RLS}}} J(\theta) = \arg\max_{\theta \in \mathcal{E}_t^{\mathrm{RLS}}} \max_{x \in \mathcal{X}} x^\mathsf{T}\theta$$

Optimism in Face of Uncertainty

- *Exploit*: given past observations, compute $\widehat{\theta}_t$ and confidence ellipsoid $\mathcal{E}_t^{\mathrm{RLS}}$
- *Explore*: given past observations, any $\theta \in \mathcal{E}_t^{\mathrm{RLS}}$ could be $\theta^\star$
- *Optimism*: trade-off exploration and exploitation by taking the most "optimistic" θ *compatible* with current estimates

$$\widetilde{\theta}_t = \arg\max_{\theta \in \mathcal{E}_t^{\mathrm{RLS}}} J(\theta) = \arg\max_{\theta \in \mathcal{E}_t^{\mathrm{RLS}}} \max_{x \in \mathcal{X}} x^\mathsf{T} \theta$$

- Act *as-if* $\widetilde{\theta}_t$ was the true parameter

$$x_t = \arg\max_{x \in \mathcal{X}} x^\mathsf{T} \widetilde{\theta}_t$$

Optimism in Face of Uncertainty

- *Exploit*: given past observations, compute $\widehat{\theta}_t$ and confidence ellipsoid $\mathcal{E}_t^{\mathrm{RLS}}$
- *Explore*: given past observations, any $\theta \in \mathcal{E}_t^{\mathrm{RLS}}$ could be $\theta^\star$
- *Optimism*: trade-off exploration and exploitation by taking the most "optimistic" θ *compatible* with current estimates

$$\widetilde{\theta}_t = \arg\max_{\theta \in \mathcal{E}_t^{\mathrm{RLS}}} J(\theta) = \arg\max_{\theta \in \mathcal{E}_t^{\mathrm{RLS}}} \max_{x \in \mathcal{X}} x^\mathsf{T} \theta$$

- Act *as-if* $\widetilde{\theta}_t$ was the true parameter

$$x_t = \arg\max_{x \in \mathcal{X}} x^\mathsf{T} \widetilde{\theta}_t$$

$\Rightarrow$ the resulting algorithm is called LinUCB or OFUL

How It Works

High-level intuition

- The arm choice can be written as *(by def. of $\mathcal{E}_t^{\mathrm{RLS}}$)*

$$x_t = \arg\max_{x \in \mathcal{X}} x^{\mathsf{T}} \widetilde{\theta}_t = \arg\max_{x \in \mathcal{X}} \Big(\underbrace{x^{\mathsf{T}} \widehat{\theta}_t}_{\textit{exploit}} + \underbrace{\|x\|_{V_t^{-1}} \sqrt{d \log(t/\delta)}}_{\textit{explore}} \Big)$$

How It Works

High-level intuition

- The arm choice can be written as *(by def. of $\mathcal{E}_t^{\mathrm{RLS}}$)*

$$x_t = \arg\max_{x \in \mathcal{X}} x^\mathsf{T} \widetilde{\theta}_t = \arg\max_{x \in \mathcal{X}} \Big(\underbrace{x^\mathsf{T} \widehat{\theta}_t}_{exploit} + \underbrace{\|x\|_{V_t^{-1}} \sqrt{d \log(t/\delta)}}_{explore} \Big)$$

- *Case 1*: $x_t = x^\star(\theta^\star) \Rightarrow$ *no regret*

How It Works

High-level intuition

- The arm choice can be written as *(by def. of $\mathcal{E}_t^{\mathrm{RLS}}$)*

$$x_t = \arg\max_{x \in \mathcal{X}} x^{\mathsf{T}} \widetilde{\theta}_t = \arg\max_{x \in \mathcal{X}} \Big(\underbrace{x^{\mathsf{T}} \widehat{\theta}_t}_{\text{exploit}} + \underbrace{\|x\|_{V_t^{-1}} \sqrt{d \log(t/\delta)}}_{\text{explore}} \Big)$$

- *Case 1*: $x_t = x^\star(\theta^\star) \Rightarrow$ *no regret*

- *Case 2*: $x_t \neq x^\star(\theta^\star) \Rightarrow$ *the confidence ellipsoid is tightened along the direction whose uncertainty had the largest impact in the decision of* x_t

How It Works

High-level intuition

- The arm choice can be written as *(by def. of $\mathcal{E}_t^{\mathrm{RLS}}$)*

$$x_t = \arg\max_{x \in \mathcal{X}} x^\mathsf{T} \widetilde{\theta}_t = \arg\max_{x \in \mathcal{X}} \Big(\underbrace{x^\mathsf{T} \widehat{\theta}_t}_{\text{exploit}} + \underbrace{\|x\|_{V_t^{-1}} \sqrt{d \log(t/\delta)}}_{\text{explore}} \Big)$$

- *Case 1*: $x_t = x^\star(\theta^\star) \Rightarrow$ *no regret*
- *Case 2*: $x_t \neq x^\star(\theta^\star) \Rightarrow$ *the confidence ellipsoid is tightened along the direction whose uncertainty had the largest impact in the decision of* x_t

$\Rightarrow$ either *instantaneous regret is small* or *useful information is obtained* and future regret will be small

How It Works

Proof sketch

- Regret decomposition

$$
\begin{aligned}
R(T) &= \sum_{t=1}^{T} \left(x^\star (\theta^\star)^\mathsf{T} \theta^\star - x_t^\mathsf{T} \theta^\star \right) \\
&= \sum_{t=1}^{T} \left(x^\star (\theta^\star)^\mathsf{T} \theta^\star - x_t^\mathsf{T} \widetilde{\theta}_t \right) + \sum_{t=1}^{T} \left(x_t^\mathsf{T} \widetilde{\theta}_t - x_t^\mathsf{T} \theta^\star \right) \\
&= \underbrace{\sum_{t=1}^{T} \left(J(\theta^\star) - J(\widetilde{\theta}_t) \right)}_{R_1(T)} + \underbrace{\sum_{t=1}^{T} \left(x_t^\mathsf{T} \widetilde{\theta}_t - x_t^\mathsf{T} \theta^\star \right)}_{R_2(T)}
\end{aligned}
$$

How It Works

Proof sketch

- Regret decomposition

$$R(T) = \sum_{t=1}^{T} \left(x^\star (\theta^\star)^\mathsf{T} \theta^\star - x_t^\mathsf{T} \theta^\star \right)$$
$$= \sum_{t=1}^{T} \left(x^\star (\theta^\star)^\mathsf{T} \theta^\star - x_t^\mathsf{T} \widetilde{\theta}_t \right) + \sum_{t=1}^{T} \left(x_t^\mathsf{T} \widetilde{\theta}_t - x_t^\mathsf{T} \theta^\star \right)$$
$$= \underbrace{\sum_{t=1}^{T} \left(J(\theta^\star) - J(\widetilde{\theta}_t) \right)}_{R_1(T)} + \underbrace{\sum_{t=1}^{T} \left(x_t^\mathsf{T} \widetilde{\theta}_t - x_t^\mathsf{T} \theta^\star \right)}_{R_2(T)}$$

- $R_1(T) \leq 0$ by construction *(recall $\widetilde{\theta}_t = \arg\max_{\theta \in \mathcal{E}_t^{\mathrm{RLS}}} J(\theta)$)*

How It Works

Proof sketch

- Regret decomposition

$$
\begin{aligned}
R(T) &= \sum_{t=1}^{T} \left(x^\star(\theta^\star)^\mathsf{T}\theta^\star - x_t^\mathsf{T}\theta^\star \right) \\
&= \sum_{t=1}^{T} \left(x^\star(\theta^\star)^\mathsf{T}\theta^\star - x_t^\mathsf{T}\widetilde{\theta}_t \right) + \sum_{t=1}^{T} \left(x_t^\mathsf{T}\widetilde{\theta}_t - x_t^\mathsf{T}\theta^\star \right) \\
&= \underbrace{\sum_{t=1}^{T} \left(J(\theta^\star) - J(\widetilde{\theta}_t) \right)}_{R_1(T)} + \underbrace{\sum_{t=1}^{T} \left(x_t^\mathsf{T}\widetilde{\theta}_t - x_t^\mathsf{T}\theta^\star \right)}_{R_2(T)}
\end{aligned}
$$

- $R_1(T) \leq 0$ by construction *(recall $\widetilde{\theta}_t = \arg\max_{\theta \in \mathcal{E}_t^{\mathrm{RLS}}} J(\theta)$)*
- $R_2(T)$ is the prediction error on points x_t used to estimate $\widehat{\theta}_t$ and $\widetilde{\theta}_t$ is in $\mathcal{E}_t^{\mathrm{RLS}}$
  $\Rightarrow$ cumulatively small

How It Works

Theorem (Abbasi-Yadkori *et al.*, 2011)

If OFUL *is run over* T *steps on arms in* $\mathcal{X} \subset \mathbb{R}^d$*, then it suffers a cumulative regret*

$$R(T) = \widetilde{O}\big(d\sqrt{T}\big)$$

with high probability.

Main Issue

Computing $\widetilde{\theta}_t$ requires solving a doubly-linear optimization problem

$$\widetilde{\theta}_t = \arg \max_{\theta \in \mathcal{E}_t^{\mathrm{RLS}}} J(\theta) = \arg \max_{\theta \in \mathcal{E}_t^{\mathrm{RLS}}} \max_{x \in \mathcal{X}} x^\mathsf{T}\theta$$

Main Issue

Computing $\widetilde{\theta}_t$ requires solving a doubly-linear optimization problem

$$\widetilde{\theta}_t = \arg\max_{\theta\in\mathcal{E}_t^{\mathrm{RLS}}} J(\theta) = \arg\max_{\theta\in\mathcal{E}_t^{\mathrm{RLS}}} \max_{x\in\mathcal{X}} x^\mathsf{T}\theta$$

$\Rightarrow$ ***computational expensive*** for non-trivial arm sets $\mathcal{X}$

Outline

Thompson Sampling

A Bayesian algorithm *(dating back to [Thompson, 1933])*

- Define a *prior* on parameter $p(\theta)$ *(e.g., $\theta \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$)*

Thompson Sampling

A Bayesian algorithm *(dating back to [Thompson, 1933])*

- Define a *prior* on parameter $p(\theta)$ *(e.g.,* $\theta \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$*)*
- At each step $t = 1, \ldots, T$
  - Draw $\widetilde{\theta}_t$ from posterior $p(\theta | x_1, r_1, \ldots, x_{t-1}, r_{t-1})$ *(e.g.,* $\theta \sim \mathcal{N}(\widehat{\theta}_t, V_t^{-1})$*)*

Thompson Sampling

A Bayesian algorithm *(dating back to [Thompson, 1933])*

- Define a *prior* on parameter $p(\theta)$ *(e.g., $\theta \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$)*
- At each step $t = 1, \dots, T$
  - Draw $\widetilde{\theta}_t$ from posterior $p(\theta | x_1, r_1, \dots, x_{t-1}, r_{t-1})$ *(e.g., $\theta \sim \mathcal{N}(\widehat{\theta}_t, V_t^{-1})$)*
  - Select arm $x_t = \arg\max_{x \in \mathcal{X}} x^\mathsf{T} \widetilde{\theta}_t$

Thompson Sampling

A Bayesian algorithm *(dating back to [Thompson, 1933])*

- Define a *prior* on parameter $p(\theta)$ *(e.g., $\theta \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$)*
- At each step $t = 1, \ldots, T$
  - Draw $\widetilde{\theta}_t$ from posterior $p(\theta | x_1, r_1, \ldots, x_{t-1}, r_{t-1})$ *(e.g., $\theta \sim \mathcal{N}(\widehat{\theta}_t, V_t^{-1})$)*
  - Select arm $x_t = \arg\max_{x \in \mathcal{X}} x^\mathsf{T} \widetilde{\theta}_t$

$\Rightarrow$ sampling $\widetilde{\theta}_t$ from the posterior implements an *exploration*-*exploitation* trade-off

How It Works

- Regret decomposition

$$
\begin{aligned}
R(T) &= \sum_{t=1}^{T} \big(x^\star(\theta^\star)^\mathsf{T}\theta^\star - x_t^\mathsf{T}\theta^\star \big) \\
&= \underbrace{\sum_{t=1}^{T} \big(J(\theta^\star) - J(\widetilde{\theta}_t) \big)}_{R_1(T)} + \underbrace{\sum_{t=1}^{T} \big(x_t^\mathsf{T}\widetilde{\theta}_t - x_t^\mathsf{T}\theta^\star \big)}_{R_2(T)}
\end{aligned}
$$

How It Works

- Regret decomposition

$$R(T) = \sum_{t=1}^{T} \left(x^\star(\theta^\star)^\mathsf{T} \theta^\star - x_t^\mathsf{T} \theta^\star \right)$$
$$= \underbrace{\sum_{t=1}^{T} \left(J(\theta^\star) - J(\widetilde{\theta}_t) \right)}_{R_1(T)} + \underbrace{\sum_{t=1}^{T} \left(x_t^\mathsf{T} \widetilde{\theta}_t - x_t^\mathsf{T} \theta^\star \right)}_{R_2(T)}$$

- $R_2(T)$ is the same as before

How It Works

- Regret decomposition

$$
\begin{aligned}
R(T) &= \sum_{t=1}^{T} \left(x^\star(\theta^\star)^\mathsf{T}\theta^\star - x_t^\mathsf{T}\theta^\star \right) \\
&= \underbrace{\sum_{t=1}^{T} \left(J(\theta^\star) - J(\widetilde{\theta}_t) \right)}_{R_1(T)} + \underbrace{\sum_{t=1}^{T} \left(x_t^\mathsf{T}\widetilde{\theta}_t - x_t^\mathsf{T}\theta^\star \right)}_{R_2(T)}
\end{aligned}
$$

- $R_2(T)$ is the same as before
- $\widetilde{\theta}_t^{\mathrm{OFUL}} = \arg\max_{\theta \in \mathcal{E}_t^{\mathrm{RLS}}} J(\theta) \quad$ vs $\quad \widetilde{\theta}_t^{\mathrm{TS}} \sim p(\theta | x_1, r_1, \ldots, x_{t-1}, r_{t-1})$

$\Rightarrow J(\widetilde{\theta}_t^{\mathrm{TS}}) = ??$

How It Works

- Regret decomposition

$$R(T) = \sum_{t=1}^{T} \left(x^\star(\theta^\star)^\mathsf{T}\theta^\star - x_t^\mathsf{T}\theta^\star \right)$$
$$= \underbrace{\sum_{t=1}^{T} \left(J(\theta^\star) - J(\widetilde{\theta}_t) \right)}_{R_1(T)} + \underbrace{\sum_{t=1}^{T} \left(x_t^\mathsf{T}\widetilde{\theta}_t - x_t^\mathsf{T}\theta^\star \right)}_{R_2(T)}$$

- $R_2(T)$ is the same as before
- $\widetilde{\theta}_t^{\mathrm{OFUL}} = \arg\max_{\theta \in \mathcal{E}_t^{\mathrm{RLS}}} J(\theta) \quad \text{vs} \quad \widetilde{\theta}_t^{\mathrm{TS}} \sim p(\theta | x_1, r_1, \ldots, x_{t-1}, r_{t-1})$

$\Rightarrow J(\widetilde{\theta}_t^{\mathrm{TS}}) =?? \quad \Rightarrow R_1(T) \nleq 0$

The Importance of Being Optimistic

Let $R_t^{\mathrm{TS}} = J(\theta^\star) - J(\widetilde{\theta}_t)$

- At step τ, $\widetilde{\theta}_\tau$ is optimistic (i.e., $J(\widetilde{\theta}_\tau) \geq J(\theta^\star)$), then $R_\tau^{\mathrm{TS}} \leq 0$

The Importance of Being Optimistic

Let $R_t^{\mathrm{TS}} = J(\theta^\star) - J(\widetilde{\theta}_t)$

- At step τ, $\widetilde{\theta}_\tau$ is optimistic (i.e., $J(\widetilde{\theta}_\tau) \geq J(\theta^\star)$), then $R_\tau^{\mathrm{TS}} \leq 0$
- At any other subsequent (non-optimistic) step t

$$R_t^{\mathrm{TS}} \leq J(\widetilde{\theta}_\tau) - J(\widehat{\theta}_t) \qquad J(\widetilde{\theta}_\tau) \geq J(\theta^\star)$$

The Importance of Being Optimistic

Let $R_t^{\mathrm{TS}} = J(\theta^\star) - J(\widetilde{\theta}_t)$

- At step τ, $\widetilde{\theta}_\tau$ is optimistic (i.e., $J(\widetilde{\theta}_\tau) \geq J(\theta^\star)$), then $R_\tau^{\mathrm{TS}} \leq 0$
- At any other subsequent (non-optimistic) step t

$$R_t^{\mathrm{TS}} \leq J(\widetilde{\theta}_\tau) - J(\widehat{\theta}_t) \qquad J(\widetilde{\theta}_\tau) \geq J(\theta^\star)$$

$$\leq \nabla J(\widetilde{\theta}_\tau)^\mathsf{T}(\widetilde{\theta}_\tau - \widetilde{\theta}_t) \qquad J(\theta) \text{ is convex}$$

The Importance of Being Optimistic

Let $R_t^{\mathrm{TS}} = J(\theta^\star) - J(\widetilde{\theta}_t)$

- At step τ, $\widetilde{\theta}_\tau$ is optimistic (i.e., $J(\widetilde{\theta}_\tau) \geq J(\theta^\star)$), then $R_\tau^{\mathrm{TS}} \leq 0$
- At any other subsequent (non-optimistic) step t

$$
\begin{aligned}
R_t^{\mathrm{TS}} &\leq J(\widetilde{\theta}_\tau) - J(\widehat{\theta}_t) && J(\widetilde{\theta}_\tau) \geq J(\theta^\star) \\
&\leq \nabla J(\widetilde{\theta}_\tau)^\mathsf{T}(\widetilde{\theta}_\tau - \widetilde{\theta}_t) && J(\theta) \textit{ is convex} \\
&\leq {x_\tau}^\mathsf{T}(\widetilde{\theta}_\tau - \widetilde{\theta}_t) && \nabla J(\theta) = x^\star(\theta) \textit{ by def.}
\end{aligned}
$$

The Importance of Being Optimistic

Let $R_t^{\mathrm{TS}} = J(\theta^\star) - J(\widetilde{\theta}_t)$

- At step τ, $\widetilde{\theta}_\tau$ is optimistic (i.e., $J(\widetilde{\theta}_\tau) \geq J(\theta^\star)$), then $R_\tau^{\mathrm{TS}} \leq 0$
- At any other subsequent (non-optimistic) step t

$$
\begin{aligned}
R_t^{\mathrm{TS}} &\leq J(\widetilde{\theta}_\tau) - J(\widehat{\theta}_t) && J(\widetilde{\theta}_\tau) \geq J(\theta^\star) \\
&\leq \nabla J(\widetilde{\theta}_\tau)^{\mathsf{T}}(\widetilde{\theta}_\tau - \widetilde{\theta}_t) && J(\theta) \textit{ is convex} \\
&\leq {x_\tau}^{\mathsf{T}}(\widetilde{\theta}_\tau - \widetilde{\theta}_t) && \nabla J(\theta) = x^\star(\theta) \textit{ by def.} \\
&\leq \|x_\tau\|_{V_\tau^{-1}} \|\widetilde{\theta}_\tau - \widetilde{\theta}_t\|_{V_t} && \textit{by Cauchy-Schwarz}
\end{aligned}
$$

The Importance of Being Optimistic

- Summing up *(ν_k time between any two optimistic choices, τ_k optimistic times)*

$$\sum_{t=1}^{T} R_t^{\mathrm{TS}} \leq \sqrt{dT} \sum_{k=1}^{K} \nu_k \|x_{\tau_k}\|_{V_{\tau_k}^{-1}}$$

The Importance of Being Optimistic

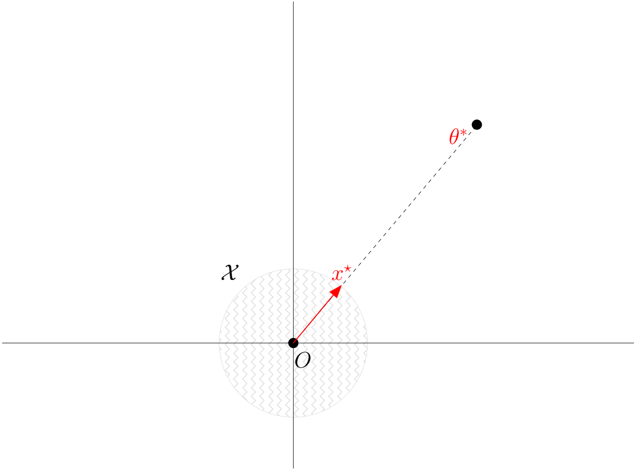
- Summing up *(ν_k time between any two optimistic choices, τ_k optimistic times)*

$$\sum_{t=1}^{T} R_t^{\mathrm{TS}} \leq \sqrt{dT} \sum_{k=1}^{K} \nu_k \|x_{\tau_k}\|_{V_{\tau_k}^{-1}}$$

- If $\widetilde{\theta}_t$ is optimistic with probability p, then $\mathbb{E}[\nu_k] = 1/p$ and

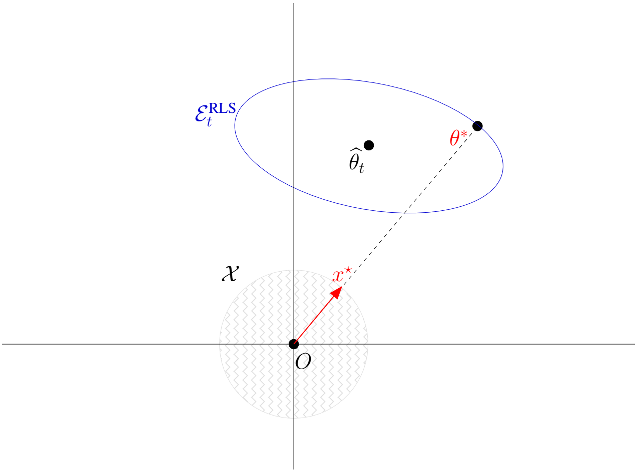
$$R(T) \leq \widetilde{O}\big(d/p\sqrt{T}\big)$$

How It Works



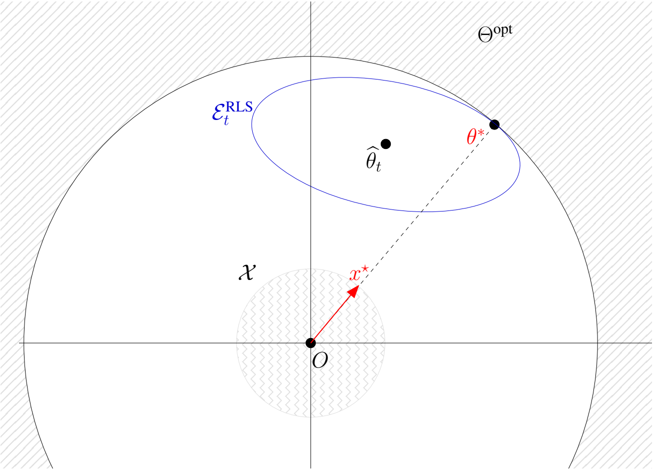
Consider $\mathcal{X} = \mathcal{B}(0,1)$, then $x^\star(\theta) = \theta/\|\theta\|$ and $J(\theta) = \|\theta\|$

How It Works



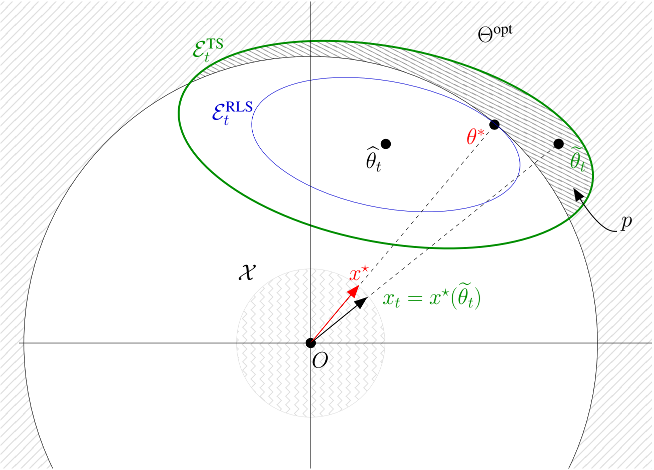
Consider $\mathcal{X} = \mathcal{B}(0,1)$, then $x^\star(\theta) = \theta/\|\theta\|$ and $J(\theta) = \|\theta\|$

How It Works



Consider $\mathcal{X} = \mathcal{B}(0,1)$, then $x^\star(\theta) = \theta/\|\theta\|$ and $J(\theta) = \|\theta\|$

How It Works



Consider $\mathcal{X} = \mathcal{B}(0,1)$, then $x^\star(\theta) = \theta/\|\theta\|$ and $J(\theta) = \|\theta\|$

How It Works

Theorem (Agrawal & Goyal, 2012; Abeille & L., 2016)

If TS *is run over* T *steps on arms in* $\mathcal{X} \subset \mathbb{R}^d$*, then it suffers a cumulative regret*

$$R(T) = \widetilde{O}\big(d^{3/2}\sqrt{T}\big)$$

with high probability.

Discussion

+ TS is computationally faster than OFUL
+ TS often performs better than OFUL

Discussion

+ TS is computationally faster than OFUL
+ TS often performs better than OFUL
– the need for optimism worsens the bound by $\sqrt{d}$
– the Bayesian design requires choosing appropriate priors

Outline

Thompson Sampling as a Stochastic Algorithm?

Probability matching algorithm

- Define a *prior* on parameter $p(\theta)$

Thompson Sampling as a Stochastic Algorithm?

Probability matching algorithm

- Define a *prior* on parameter $p(\theta)$
- At each step $t = 1, \ldots, T$
  - For any arm $x \in \mathcal{X}$, compute

$$\pi_t(x) = \mathbb{P}(x = x^\star | x_1, r_1, \ldots, x_{t-1}, r_{t-1})$$

  - Select arm $x_t \sim \pi_t$

Thompson Sampling as a Stochastic Algorithm?

Probability matching algorithm

- Define a *prior* on parameter $p(\theta)$
- At each step $t = 1, \ldots, T$
  - For any arm $x \in \mathcal{X}$, compute

$$\pi_t(x) = \mathbb{P}(x = x^\star | x_1, r_1, \ldots, x_{t-1}, r_{t-1})$$

  - Select arm $x_t \sim \pi_t$

$\Rightarrow$ TS is an *efficient implementation* of probability matching

Thompson Sampling as a Stochastic Algorithm?

- In some cases we can explicitly write the Bayesian update as ($\boldsymbol{\pi}_t$ being a distribution over $\mathcal{X}$)

$$\boldsymbol{\pi}_{t+1} = \boldsymbol{\pi}_t + \boldsymbol{\Delta}_t$$

Thompson Sampling as a Stochastic Algorithm?

- In some cases we can explicitly write the Bayesian update as ($\boldsymbol{\pi}_t$ being a distribution over $\mathcal{X}$)

$$\boldsymbol{\pi}_{t+1} = \boldsymbol{\pi}_t + \boldsymbol{\Delta}_t$$

- The *Narendra-Shapiro* algorithm is a *stochastic algorithm* updating a distribution over arms as

$$\boldsymbol{\pi}_{t+1} = \boldsymbol{\pi}_t + \gamma_t \boldsymbol{\Delta}'_t$$

Thompson Sampling as a Stochastic Algorithm?

- In some cases we can explicitly write the Bayesian update as ($\boldsymbol{\pi}_t$ being a distribution over $\mathcal{X}$)
$$\boldsymbol{\pi}_{t+1} = \boldsymbol{\pi}_t + \boldsymbol{\Delta}_t$$
- The *Narendra-Shapiro* algorithm is a *stochastic algorithm* updating a distribution over arms as
$$\boldsymbol{\pi}_{t+1} = \boldsymbol{\pi}_t + \gamma_t \boldsymbol{\Delta}'_t$$
- The (over-penalized)-*Narendra-Shapiro* is shown *[Gadat et al., 2016]*
  - to converge to the stationary distribution of a *piecewise deterministic Markov process*
  - to suffer from a worst-case regret $\widetilde{O}(\sqrt{T})$ (in the 2-arm independent Bernoulli case)

Thompson Sampling as a Stochastic Algorithm?

- In some cases we can explicitly write the Bayesian update as ($\boldsymbol{\pi}_t$ being a distribution over $\mathcal{X}$)

$$\boldsymbol{\pi}_{t+1} = \boldsymbol{\pi}_t + \boldsymbol{\Delta}_t$$

- The *Narendra-Shapiro* algorithm is a *stochastic algorithm* updating a distribution over arms as

$$\boldsymbol{\pi}_{t+1} = \boldsymbol{\pi}_t + \gamma_t \boldsymbol{\Delta}'_t$$

- The (over-penalized)-*Narendra-Shapiro* is shown *[Gadat et al., 2016]*
  - to converge to the stationary distribution of a *piecewise deterministic Markov process*
  - to suffer from a worst-case regret $\widetilde{O}(\sqrt{T})$ (in the 2-arm independent Bernoulli case)

$\Rightarrow$ If TS can be seen as a stochastic algorithm, we could have a much better understanding of the dynamics and behavior of bandit algorithms

Thank you!

Alessandro Lazaric
alessandro.lazaric@inria.fr
sequel.lille.inria.fr