FINITE-TIME ANALYSIS OF KERNELISED CONTEXTUAL BANDITS

MICHAL.VALKO@INRIA.FR, NATHANIEL.KORDA@INRIA.FR, RÉMI.MUNOS@INRIA.FR,
ILIAS.FLAOUNAS@BRISTOL.AC.UK, NELLO.CRISTIANINI@BRISTOL.AC.UK

Inria (informatics mathematics) — University of BRISTOL

MOTIVATION: NEWSFEEDS

- **Goal:** Recommendation of interesting articles from newsfeeds (RSS).
- **Challenges:** Too many newsfeeds to even check all of them once and way too many articles.
- **Context:** Every feed has a set of features gathered during the RSS crawling: URL, feed titles, anchor text, ...
- **Smoothness Assumption:** Feeds with similar contexts are interesting in a similar way (have similar rewards).
- **Kernels:** We want to extract a **non-linear** relationship between the contexts and rewards, only from similarity information between the contexts.
- **Bandit setting:** We only receive the reward for the newsfeed that we try.
- **Noise:** Moreover, we only receive a reward for a specific article, which is only a noisy estimate for the reward of the whole newsfeed.

SETTING: KERNEL BANDITS

We model the setting as contextual bandits.

- **Action space:** $\mathcal{A} := \{1, \dots, N\}$
- **Contexts:** For each a, there is a context: $\boldsymbol{x}_{a,t} \in \mathbb{R}^d$, **that can change with time** t
- **Protocol:** At time $t = 1 \dots T$:
  - receive contexts $\boldsymbol{x}_{a,t}$ for all a
  - choose our action a_t
  - obtain a reward r_t
- **Rewards** depend on the context non-linearly, i.e. they are linear in mapping to the corresponding *reproducing kernel Hilbert space* (RKHS) defined by a kernel k.

$$\mathbb{E}(r_{a,t} \mid \boldsymbol{x}_{a,t}) = \phi(\boldsymbol{x}_{a,t})^\intercal \theta^*$$

- **Best action**, a_t^* at time t is **context dependent**: $a_t^* := \arg\max_{a \in \mathcal{A}} \{\mathbb{E}(r_{a,t} \mid \boldsymbol{x}_{a,t})\}$.
- **Loss:** How well we do over time w.r.t. the best possible action — **contextual regret**:

$$R(T) := \sum_{t=1}^{T} \left[r_{a_t^*,t} - r_t \right]$$

CONTRIBUTIONS

The main challenge in lifting the known analysis for the contextual bandits where the reward is **linear in primal** to the case where the reward is **linear in dual** is that dual (RKHS) may be of **infinite** dimension.

We provide:

- **frequentist** analysis of kernelised bandits
- cumulative regret bound $\tilde{O}(\sqrt{T\tilde{d}})$
- **match** $\Omega(\sqrt{d})$ lower bound for the **linear case**
- **link with GP-UCB**
  - comparison between effective dimension $\tilde{d}$ and information gain $I(y_T; \theta^*)$
  - improved analysis for the **agnostic case**
  - *data-independent* worst case upper bounds

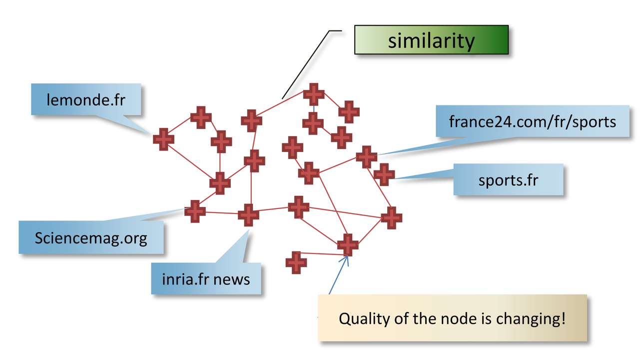
ACKNOWLEDGEMENTS AND CODE

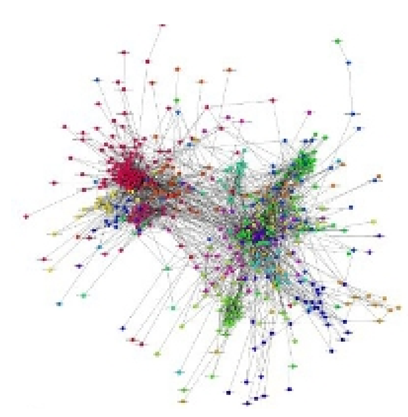
This research was funded by European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 270327 (project CompLACS).

CompLACS

Code at: HTTPS://SEQUEL.LILLE.INRIA.FR/SOFTWARE/KERNELUCB

NEWSFEEDS





KERNELUCB ALGORITHM

Input and initialisation:
 N the number of actions, T the number of pulls
 γ, η regularization and exploration parameters
 $k(\cdot,\cdot)$ kernel function
$u_0 \leftarrow [1, 0, ..., 0]^\intercal$ (at start, the first action is tried)
$y_0 \leftarrow \emptyset$
Run:
for $t = 1$ **to** T **do**
 Choose $a \leftarrow \arg\max u_{t-1}$ and get reward r_{t-1}
 Update $y_t \leftarrow [r_1, \dots, r_{t-1}]^\intercal$ and K_t
 for $a = 1$ **to** N **do**
  $\sigma_{a,t} \leftarrow \sqrt{k(x_{a,t}, x_{a,t}) - k_{x,t}^\intercal K_t^{-1} k_{x,t}}$
  $u_{a,t} \leftarrow \left(k_{x,t}^\intercal K_t^{-1} y_t + \frac{\eta}{\gamma^{1/2}} \sigma_{a,t} \right)$
 end for
end for

HOW IT WORKS?

- UCB algorithm with kernelised ridge regression:

$$u_{a,t} = \overbrace{\hat{\mu}_{a,t}}^{\text{estimator}} + \overbrace{\eta/\gamma^{1/2}\tilde{\sigma}_{a,t}}^{\text{confidence width}}.$$

- Widths in terms of the Mahalanobis distance of $\phi(x_{a,t})$ from the matrix Φ_t:

$$\hat{\sigma}_{a,t} := \sqrt{\phi(x_{a,t})^\intercal(\Phi_t^\intercal\Phi_t + \gamma I)^{-1}\phi(x_{a,t})}.$$

- $\hat{\sigma}_{a,t}$ can be also expressed using kernel trick:

$$\gamma^{-1/2}\sqrt{k(x_{a,t}, x_{a,t}) - k_{x_{a,t},t}^\intercal (K_t + \gamma I)^{-1} k_{x_{a,t},t}}$$

- In practice:
  - iterative matrix inversion for K_t^{-1}
  - lazy variance calculation for arg max

EFFECTIVE DIMENSION

- Known regret bounds for linear contextual bandits can be vacuous (dimension of the RKHS may be infinite).
- We give a bound in terms of a data dependent ***effective dimension*** $\tilde{d}$: Let $(\lambda_{i,t})_{i\geq 1}$ denote the eigenvalues of $C_t^\gamma = \Phi_t^\intercal\Phi_t + \gamma I$ in decreasing order and define:

$$\tilde{d} := \min\{j : j\gamma \ln T \geq \Lambda_{T,j}\} \text{ where } \Lambda_{T,j} := \sum_{i>j} \lambda_{i,T} - \gamma.$$

- We call $\tilde{d}$ the effective dimension because it gives a proxy for the number of principle directions over which the projection of the data in the RKHS is spread.
- If the data all fall within a subspace of $\mathcal{H}$ of dimension d', then $\Lambda_{T,d'} = 0$ and $\tilde{d} \leq d'$.
- More generally $\tilde{d}$ can be thought of as a measure of how quickly the eigenvalues of $\Phi_t^\intercal\Phi_t$ are decreasing.
- For example if the eigenvalues are only polynomially decreasing in i (i.e. $\lambda_i \leq Ci^{-\alpha}$ for some $\alpha > 1$ and some constant $C > 0$) then $\tilde{d} \leq 1 + (C/(\gamma \ln T))^{1/\alpha}$.
- When $\Phi \equiv \text{Id}$, $\tilde{d} \leq d$, the assumption that $\|\phi(x_{a,t})\| \leq 1$ becomes the assumption that the contexts are normalised in the primal, and we recover exactly the result for linear bandits which matches the lower bound for this setting.

MAIN RESULT

Theorem 1. *Assume that $\|\phi(x_{a,t})\| \leq 1$ and $|r_{a,t}| \in [0,1]$ for all $a \in A$ and $t \geq 1$, and set $\eta = \sqrt{2\ln 2TN/\delta}$. Then with probability $1-\delta$, SupKernelUCB satisfies:*

$$R(T) \leq \left[2 + 2\left(1 + \sqrt{\frac{\gamma}{2\ln(2TN(1+\ln T)/\delta)}}\right)\|\theta^*\| + \right.$$
$$+ 8\sqrt{\left(12 + \frac{15}{\gamma}\right)\max\left\{\ln\left(\frac{T}{\tilde{d}\gamma} + 1\right), \ln T\right\}^3} \times$$
$$\left. \times \sqrt{\left(2\ln\frac{2TN(1+\ln T)}{\delta}\right)} \right] \sqrt{\tilde{d}T}$$

Remark 1. *Theorem 1 suggests that if we know that $\|\theta^*\| \leq L$, for some L, we should set γ to be of the order of L^{-1} so that we obtain $\tilde{O}(\sqrt{L\tilde{d}T})$ regret. If we do not have such knowledge, just setting γ to a constant (e.g., found by a cross-validation) will incur $\tilde{O}(\|\theta^*\|\sqrt{\tilde{d}T})$ regret.*

Remark 2. *The proof uses a technique of Auer [1] in order to deal with dependent $\hat{\mu}_{a,t}$. This technique builds mutually exclusive subsets of "time steps". In this way, the Azuma-Hoeffding inequality can be applied on each subset to get a regret bound. Furthermore, although $\Phi_t^\intercal\Phi_t$ may be of infinite dimension, we show that only $\tilde{d}$ dimensions matter.*

COMPARISON

	Bayesian	Frequentist
regression	GP-Regression	Kernel Ridge Regression
bandits	GP-UCB	**KernelUCB** this work

Bayesian and frequentist approaches to kernelized regression and contextual bandits

COMPARISON TO GP-UCB

- GP-UCB is a special case of KernelUCB when γ is set to the model (GP) noise.
- Our analysis improves upon that of GP-UCB for the agnostic case: when context-to-reward mapping θ^* is not from GP.
- From the GP-UCB analysis for the agnostic case, the cumulative regret is bounded as:

$$O\left(\left(I(y_A;\theta^*) + \|\theta^*\|^2\sqrt{I(y_A;\theta^*)}\right)\sqrt{T}\right), \quad (1)$$

 where $I(y_T;\theta^*)$ is the mutual information between θ^* and the vector of (noisy) observations y_T.

- Both $I(y_T;\theta^*)$ and $\tilde{d}$ are data dependent quantities.
- Since the eigenvalues of $\Phi_T^\intercal\Phi_T$ are the same as the eigenvalues of $\Phi_T\Phi_T^\intercal$, we can show that:

$$I(y_T; f) \geq \Omega(\tilde{d}\ln\ln T)$$

- This shows that $\tilde{d}$ is at least as good as $I(y_T;\theta^*)$, and comparing our Theorem 1 with (1), our regret bound only scales as $O(\sqrt{\tilde{d}})$, while the dependence of the regret bound (1) is linear in $I(y_T;\theta^*)$.
- As a consequence of the link between $I(y_T;\theta^*)$, γ_T and $\tilde{d}$, we may also express our bounds in terms of γ_T and obtain data-independent worst case upper bounds for certain kernels: e.g. for RBF kernel, our bound scales with $O(\ln T)^{d/2}$ in place of $O(\ln T)^d$.

REFERENCES

[1] Auer P. Using confidence bounds for exploitation-exploration trade-offs. *JMLR*, 2002.

[2] Chu L., Li L., Reyzin L., and Schapire R. E. Contextual Bandits with Linear Payoff Functions. *AISTATS*, 2011.

[3] Srinivas N., Krause A., Kakade S., and Seeger M. Gaussian Process Optimization in the Bandit Setting. *ICML*, 2010.