F.

M.

SETTING: KERNEL BANDITS

We model the setting as contextual bandits.

- Action space: \( A := \{1, \ldots, N\} \)
- Contexts: For each \( a \), there is a context: \( x_{a,i} \in \mathbb{R}^d \), that can change with time \( t \)
- Protocol: At time \( t = 1, \ldots, T \):
  - receive contexts \( x_{a,i} \) for all \( a \)
  - choose our action \( a_t \)
  - obtain a reward \( r_{t,a} \)
- 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 \( K \).

\[
E(x_{a,i} | x_{a,i}) = \phi(x_{a,i})^T \theta^*
\]

Best action, \( a_t^* \) at time \( t \) is context dependent: \( a_t^* := \arg \max_{a \in A} E(x_{a,i} | x_{a,i}), E(x_{a,i} | x_{a,i}) \)

Loss: How well do we do over time w.r.t. the best possible action — contextual regret:

\[
R(T) := \sum_{t=1}^{T} [r_{a_t} - r_{a_t^*}]
\]

HOW IT WORKS?

- UCB algorithm with kernelised ridge regression:

\[
x_{a,t} := \arg \max_{a \in A} E(x_{a,i} | x_{a,i}) + \sqrt{\frac{\gamma^2}{2} K_{a,a_t}^{-1}}.
\]

- Widths in terms of the Mahalanobis distance of \( \phi(x_{a,i}) \) from the matrix \( \Phi^* \):

\[
\phi(x_{a,i})^T (\Phi^* \Phi^* + \gamma I)^{-1} \phi(x_{a,i}).
\]

- \( \gamma \) can be also expressed using kernel trick:

\[
\gamma^{-1/2} \sqrt{E(x_{a,i}^2) - K_{a,a_t}^{-1}} (\Phi^* + \gamma I)^{-1} K_{a,a_t}.
\]

In practice:

- iterative matrix inversion for \( K_{a,a_t}^{-1} \)
- 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 \( d \).

\[
\hat{d} := \min \{ j : j \geq \ln T \geq \lambda_j \}
\]

where \( \lambda_j := \sum_{j \geq i} \gamma_i (j - \gamma) \).

- We call \( \hat{d} \) 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 \( \lambda_{d'} = 0 \) and \( d = d' \).
- More generally \( d \) can be thought of as a measure of how quickly the eigenvalues of \( \Phi^* \) are decreasing.
- For example if the eigenvalues are only polynomially decreasing in \( i \) (i.e. \( \lambda_i \leq C_i \lambda_{i-1} \) for some \( \alpha > 1 \) and some constant \( C > 0 \) then \( \hat{d} \leq 1 + (C/(\alpha - 1))/\alpha \).

- When \( \Phi = 1d, \hat{d} \leq d \), the assumption that \( ||\phi(x_{a,i})|| \leq 1 \) 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 \( ||\phi(x_{a,i})|| \leq 1 \) and \( (r_{a,i}, \theta) \in \{0,1\} \) for all \( a \in A \) and \( t > 1 \), and set \( \gamma = \sqrt{2\ln TN/T} \).

Then with probability \( 1 - \delta \), SEQUEL satisfies:

\[
R(T) \leq \left[ 2 \sqrt{\frac{2\ln TN(1 + \ln T)/\delta}{\gamma^2}} + \sqrt{\frac{2\ln TN(1 + \ln T)/\delta}{\gamma^2}} \right] \sum_{a \in A} (r_{a,t} - \hat{d})
\]

Remark 1. Theorem 1 suggests that if we know that \( ||\theta^*|| \leq L \), for some \( L \), we should set \( \gamma \) to be the order of \( 1/L^2 \) so that we obtain \( O((\ln T)/L^2) \) regret. If we do not have such knowledge, just setting \( \gamma \) to a constant (e.g., found by a cross-validation) will incur \( O(||\theta^*||^2 / \gamma) \) regret.

Remark 2. The proof uses a technique of Auer [1] in order to deal with dependent \( \mu_{a,t} \). 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 \( \Phi^* \Phi^* \) may be infinite dimension, we show that only \( d \) dimensions matter.

COMPARISON

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

\[
O \left( (T r_{a,t} \log \frac{1}{\delta})^p + ||\theta^*||^2 / \sqrt{T (\ln T \gamma^2)} \right) \Delta T.
\]

where \( I(\gamma^2; \theta^*) \) is the mutual information between \( \theta^* \) and the vector of (noisy) observations \( y_t \).
- Both \( I(\gamma^2; \theta^*) \) and \( \Delta T \) are data dependent quantities.
- Since the eigenvalues of \( \Phi^* \Phi^* \) are the same as the eigenvalues of \( \Phi^* \Phi^* \), we can show that:

\[
I(\gamma^2; f) \geq \Omega(\Delta T \log T)
\]

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

REFERENCES