Metric Entropy and Gaussian Bandits

S. Grünewälder UCL

J.-Y. Audibert Université Paris-Est

M. Opper TU-Berlin

J. Shawe-Taylor UCL

Abstract

Metric entropy and generic chaining methods are powerful tools from probability theory that can be used to study pathwise properties of stochastic processes. Despite this fact they have largely been ignored in machine learning. We demonstrate their power in this work in applying them to a bandit problem with a Gaussian process prior. The difficulty of the setting lies in the fact that we are dealing with a continuous space of arms and we need to control the supremum of a reward process on the arms. We apply the so called Dudley integral to reduce the problem of controlling the supremum of a "difficult" stochastic process to the problem of bounding a canonical metric that is based solely on the covariance function (which is an analytical and thus "simple" object). We consider the scenario in which there is no noise in the observed reward. Our main result is to bound the regret experienced by algorithms relative to the a posteriori optimal strategy of playing the best arm throughout based on benign assumptions about the covariance function defining the Gaussian process.

1 Introduction

We apply the metric entropy technique in this work to Gaussian process (GP) bandits. Bandit models are used to define optimization problems where the objective function is unknown beforehand. For a given system the question arises how to set up such a model and how to encode similarities between different system states. GPs are powerful modelling tools and similarities can naturally be encoded in them through kernels. To make it more concrete, think of mobile phones for which you want to maximise throughput. The arms of the bandit would correspond to system settings. The throughput to the reward and a kernel could be used to model similarities between settings of the mobile.

The main tool we use to get the regret bounds is the Dudley integral [1] which bounds the expected supremum of a GP $r(x)$:

$$\mathbb{E} \sup_{x \in X} \{ r(x) - \mathbb{E} r(x) \} \leq 12 \int_0^\infty \sqrt{\log N(\epsilon, X, d)} d\epsilon,$$

The term $N(\epsilon, X, d)$ is the covering number of the space X (for the bandits it is the space of the arms) in the canonical metric $d(x,y)^2 = k(x,x) - 2k(x,y) + k(y,y)$, where $k(x,y)$ is the covariance of the process $r(x)$. Thus, to bound the supremum of a stochastic process, which is a complicated object, we "only" need to be able to control $d(x,y)$. Which is a much simpler task. The Dudley integral can also be used to read out other pathwise properties and extensions to other classes of processes exist, like infinitely divisible processes [3].

We apply the technique to a bandit problem, where the bandits are indexed by the unit cube $[0,1]^D$ and the reward function $r(x)$ is drawn once from a zero mean GP with specified covariance function k and can be observed without noise. The main result is an upper bound on the regret for GPs with covariance functions that are Hölder continuous with a coefficient α and a constant L_k (details are given in the next section). Assuming a zero mean GP prior, we know that after playing T arms in $[0,1]^D$ that the optimal strategy has a regret no bigger than

$$4\sqrt{\frac{L_k \log(2T)}{(2\tilde{T})^\alpha}} + 24\sqrt{\frac{DL_k}{\alpha \tilde{T}^\alpha}},$$

where $\tilde{T} := \lfloor T^{1/D} \rfloor$.

The problem is relevant as Gaussian process bandits are becoming popular in practice and theoretical underpinnings are missing.

2 The bandit game with Gaussian process prior

Central to our approach are two Hölder continuity assumptions for the mean and the covariance. The covariance function k and the mean function μ are assumed to satisfy the following mild conditions:

- **(A1)** For some $L_\mu \geq 0$, for any $x, y \in X$, $\big|\mu(x) - \mu(y)\big| \leq L_\mu |x-y|_\infty$.
- **(A2)** For some $L_k \geq 0$ and some $\alpha > 0$, for any $x, y \in X$, $\big|k(x,x) - k(x,y)\big| \leq L_k |x-y|_\infty^\alpha$.

We use the supremum norms as it leads to simpler bounds (compared to the Euclidean norm for instance). However, as X is a finite dimensional space all norms are equivalent, meaning the results apply to arbitrary norms up to an extra constant.

2.1 Game definition

The player knows the (prior) distribution of the Gaussian process (he knows the expectation function μ and the covariance function k). He also knows the number T of rounds to play. The game is the following.

For $t = 1, \ldots, T$

- the player chooses $\hat{x}_t \in X$
- the player observes $r(\hat{x}_t)$

At the end of these T rounds, the player receives the reward

$$\max\big(r(\hat{x}_1), \ldots, r(\hat{x}_T)\big).$$

A common alternative target encountered in the bandit litterature is the cumulative reward $\sum_{t=1}^T r(\hat{x}_t)$. In applications, it is often the case that the 'max' reward is more interesting than the cumulative one. The two problems differ substantially in the sense that there is no cost of exploring when we consider the 'max' reward.

2.2 The optimal strategy

It is then natural to define the optimal strategy as the one having the highest expected reward:

$$\mathbb{E} \max\big(r(\hat{x}_1), \ldots, r(\hat{x}_T)\big).$$

It occurs that the optimal strategy is the T-steps lookahead strategy in which at time t the player chooses

$$\hat{x}_t = \operatorname*{argmax}_{x_t \in X} \mathbb{E}\big[\max_{x_{t+1}\in X} \cdots \mathbb{E}\big[\max_{x_{T-1}\in X} \mathbb{E}\big[\max_{x_T \in X} \mathbb{E}\big[\max\big(r(x_1), \ldots, r(x_T)\big)\big| r(x_1), \ldots, r(x_{T-1})\big] \\ \big| r(x_1), \ldots, r(x_{T-2})\big] \cdots \big| r(x_1), \ldots, r(x_{t-1})\big]. \quad (1)$$

The formula is a bit awful, but the idea is simple. If we are at time T, the optimal action is

$$\hat{x}_T = \operatorname*{argmax}_{x_T \in X} \mathbb{E}\big[\max\big(r(x_1), \ldots, r(x_T)\big)\big| r(x_1), \ldots, r(x_{T-1})\big],$$

which corresponds to a one-step lookahead. From this, we know that at time $T-1$, if we choose x_{T-1}, our expected reward knowing the past will be

$$\mathbb{E}\big[\max_{x_T\in X}\mathbb{E}\big[\max\big(r(x_1),\dots,r(x_T)\big)\big|r(x_1),\dots,r(x_{T-1})\big]\big|r(x_1),\dots,r(x_{T-2})\big].$$

This leads to the optimal choice for x_{T-1} defined in (1). Repeated this argument again, we obtain the T-steps lookahead strategy. This policy is, in spirit, the same as the one proposed in [2, Section 3.2]. Our main result is to provide a regret bound for it. The difficulty of proving results on the algorithm relies on the understanding of the behaviours of the mean and covariance of the Gaussian process conditional to the past observations.

2.3 Guarantee on the performance of the optimal strategy

We call a guarantee on how the optimal strategy works a lower bound on the expected reward $\mathbb{E}\max\big(r(\hat{x}_1),\dots,r(\hat{x}_T)\big)$, or equivalently an upper bound on the expected regret $\mathbb{E}\big\{\sup_{x\in X} r(x) - \max\big(r(\hat{x}_1),\dots,r(\hat{x}_T)\big)\big\}$. We will use the following path to get the bound:

- define a simpler strategy: we will consider the naive grid strategy. We partition the space of arms X into $\tilde{T}^D \le T$ many cubes of side length $1/\tilde{T}$ and play the center of each of these cubes.
- lower bound the expected reward of this simpler strategy. This last point will rely on the use of concentration inequalities for the supremum of Gaussian and sub-Gaussian processes.

Theorem 2.1 *Under Assumptions (A1) and (A2), The optimal strategy satisfies*

$$\begin{aligned}
&\mathbb{E}\Big\{\sup_{x\in X} r(x) - \max\big(r(\hat{x}_1),\dots,r(\hat{x}_T)\big)\Big\}\\
\le\;& \mathbb{E}\Big\{\sup_{x\in X} r(x) - \max\big(r(x_1),\dots,r(x_T)\big)\Big\}\\
\le\;& 4\sqrt{\frac{L_k\log(2T)}{(2\tilde{T})^\alpha}} + 24\sqrt{\frac{DL_k}{\alpha\tilde{T}^\alpha}} + \frac{L_\mu}{2\tilde{T}}.
\end{aligned}$$

3 Future Work

We intend to extend this work to model of rewards incorporating a noise term so that the true reward would not then be directly observed. A more ambitious goal is to develop reward bounds for algorithms that merge the exploration and exploitation phases through for example selecting arms based on an upper confidence bound. It will also be interesting to see if more detail properties of the kernel (other than Hölder continuity) can be incorporated into the analysis to provide tighter bounds for special types of kernel, such as for example specific spectral properties.

We believe that the line of analysis developed here will provide important groundwork for these further studies and will at the same time help to inform improvements in practical Gaussian process bandit algorithms.

References

[1] P. Massart. *Concentration inequalities and model selection: Ecole d'été de Probabilités de Saint-Flour XXXIII - 2003.* Lecture Notes in Mathematics. Springer, 2003.

[2] M. A. Osborne, R. Garnett, and S. J. Roberts. Gaussian processes for global optimization. In *LION3*, 2009.

[3] M. Talagrand. *The Generic Chaining.* Springer, 2005.