Statistical Learning for Decision

Tamás Papp

January 2020

Abstract

This report briefly explores two areas of interest in statistical learning: multi-armed bandits and Bayesian optimisation. It expands on the latter, focusing on a recent method and discussing possible extensions of the overall methodology.

1 Introduction

The field of decision theory studies reasonable decision-makers in an uncertain environment. Multi-armed bandits, introduced by [1] in the context of clinical trials, are one of the most studied settings in the field: faced with a row of slot machines (or "one-armed bandits"), maximise the total payout.

Remarkably, advances in this area have also motivated strategies for efficient global optimisation. Major applications in science, technology and engineering require the optimisation of functions that have unknown structure and are expensive to evaluate. Bayesian optimisation methods offer a principled, sequential, decision-theoretic way of doing this.

2 Multi-armed Bandits

Consider the setting of an individual facing a row of K slot machines, with T rounds at their disposal. In any round t they have a choice $a_t \in \{1, \ldots, K\}$ of slot machine to play, which gives independent reward $X_{a_t,t}$ from an unknown distribution ν_{a_t}. They want a policy (or strategy) π that maximises overall expected reward, which is a difficult objective from a mathematical standpoint. Thankfully, the equivalent formulation of expected cumulative regret minimisation is more tractable. The expected cumulative regret is

$$\begin{aligned}\mathrm{Reg}_\pi(T) &:= \sum_{t=1}^{T} \left(\mu^* - \mathbb{E}_\pi\left[X_{a_t,t}\right]\right) \\ &= \sum_{a=1}^{K} \left(\mu^* - \mu_a\right) \mathbb{E}_\pi\left[\sum_{t=1}^{T} \mathbb{1}\left\{a_t = a\right\}\right],\end{aligned}$$

where $\mu_a = \mathbb{E}\left[X_{a,t}\right]$ and $\mu^* = \max_{1 \leqslant a \leqslant K} \mu_a$. Cumulative regret has been shown to be $\Omega(\log T)$ for any Bernoulli bandit [2] (later extended to more general situations [3]) and linear at the start of a policy [4]. If the cumulative regret is $o(T)$ for a given policy, that policy will converge to making the optimal decision as $T \to \infty$.

The reward distributions being initially unknown, a good policy will have to make exploratory moves in order to learn them, at the risk of playing suboptimal arms. At the same time, regret minimisation dictates greedily playing arms that seem best. This balance of exploration and exploitation is crucial to the success of a policy: exploring too much might mean good arms are not played enough, while the best arm might be missed altogether by being too greedy.

There have been several ideas proposed in the literature to achieve this balance. At any given step, these strategies compute indices for each arm and choose the arm with the highest index. The basic explore then commit algorithm plays all arms for an equal amount of time initially and then fully commits to the one with highest sample mean. The UCB algorithm [5] uses Hoeffding's inequality to upper bound the mean of each arm, setting said bound as its index. Thompson Sampling [1] is a Bayesian approach, its indices being draws from the posterior of each arm's mean. If more exploration is desired, an ϵ-greedy policy can be applied, where the original "greedy" policy is combined with playing an arm uniformly at random with small probability $\epsilon > 0$ at each step.

The main directions of research in this area lie in devising and adapting algorithms for more complex, realistic scenarios. These include contextual bandits that have rewards changing according to the context, for instance a particular user type of an online marketplace, time-varying bandits that allow the reward distributions to change over time and history-dependent bandits that consider the reward distribution altering once an arm has been played.

3 Bayesian Optimisation

In many science and engineering applications we are interested in globally optimising a function that has unknown structure and is expensive to evaluate. We restrict to the minimization problem

$$x^* = \operatorname*{argmin}_{x \in \mathcal{X}} f(x)$$

for $f : \mathcal{X} \to \mathbb{R}$ generally considered to be Lipschitz continuous and $\mathcal{X} \in \mathbb{R}^D$ compact. A grid or random search is prohibitively expensive and while metaheuristic methods such as genetic and generalised hillclimbing algorithms exist for this purpose, they are poorly understood. A particularly successful approach is instead attempting to learn f based on a statistical model of its structure. Such "Bayesian optimisation" methods impose a stochastic prior on f, iteratively searching for new evaluation points, evaluating f at them, and updating the posterior using Bayes' rule.

3.1 The Surrogate Model

To model our uncertainty about f, the prior is commonly chosen to be a Gaussian process (GP). We write $f \sim \mathcal{GP}(\mu, k)$ for a GP with mean and covariance (kernel) functions $\mu : \mathcal{X} \to \mathbb{R}$ and $k : \mathcal{X}^2 \to \mathbb{R}$ respectively. The model implies joint Gaussianity, that is for all $x_1, \ldots, x_n \in \mathcal{X}$

$$\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} \sim \mathcal{N} \left(\boldsymbol{\mu} = \begin{pmatrix} \mu(x_1) \\ \vdots \\ \mu(x_n) \end{pmatrix}, K = \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{pmatrix} \right).$$

GPs are attractive for Bayesian optimisation for two main reasons. Firstly, they have a posterior with closed form. Assume we've already collected data $\mathcal{D}_n = \{(x_i, y_i)_{i=1}^n\}$ from independent noisy observations $y_i \sim \mathcal{N}(f(x_i), \sigma^2)$. By a straightforward application of Bayes' rule we obtain

$$\begin{aligned} f|\mathcal{D}_n &\sim \mathcal{GP}(\mu_n, k_n), \\ \mu_n(x) &= \mu(x) + \boldsymbol{k}_n(x)^T \left(K + \sigma^2 I\right)^{-1} (\boldsymbol{y} - \boldsymbol{\mu}), \\ k_n(x, x') &= k(x, x') - \boldsymbol{k}_n(x)^T \left(K + \sigma^2 I\right)^{-1} \boldsymbol{k}_n(x'), \end{aligned}$$

where $\boldsymbol{k}_n^{(i)}(x) = k(x_i, x)$. We write $f(x)|\mathcal{D}_n \sim \mathcal{N}(\mu_n(x), \sigma_n^2(x))$ and assume this model in §3.2, §3.3. Of practical note here is the matrix inversion in the posterior updates. No matter how quick our search procedure is (even if the function evaluations were relatively cheap), not being query-efficient eventually incurs a cubic slowdown.

Secondly, GPs are able to accurately fit a wide class of functions. The kernel k uniquely defines a broad function space called a reproducing kernel Hilbert space $\mathcal{H}$, which intuitively contains all functions that are as smooth as a posterior mean of a GP with said kernel. The algorithms in the sequel converge to optimality when $f \in \mathcal{H}$.

The kernel warrants special attention in practical Bayesian optimisation threefold: it encodes the degree of smoothness of f, the underlying shape of f and it guides the GP updates. The popular squared exponential and Matérn classes of kernels

$$k^{\mathrm{sqe}}(x, x') := \exp\left(-\frac{r^2}{h}\right), \qquad\qquad k^{\mathrm{Mat}}(x, x') := \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{r\sqrt{2\nu}}{h}\right)^{\nu} B_\nu \left(\frac{r\sqrt{2\nu}}{h}\right)$$

where $\nu, h > 0$, $r := \|x - x'\|_2$ and B_ν is a modified Bessel function of the second kind, cover the whole spectrum of smoothness we would reasonably expect from the objective, from most to least. If the kernel is too coarse, spurious evaluations will be made, and if a kernel is too smooth, the search procedure may not recognise the potential for global minima. Another potential issue is the above kernels' stationarity: the covariance depending only on the distance between points is not a good fit for an objective that, say, curves downwards near the boundary of $\mathcal{X}$. To work around this, we can attempt to capture the shape of f in μ a priori, consider linear combinations of kernels with long and short scales h, or apply a warping transformation $k(x, x') \mapsto k(w(x), w(x'))$.

Student-t processes have also been suggested for Bayesian optimisation [6]. They are both tractable and more lenient than Gaussian processes with regard to misspecification due to their heavier tail, therefore having the potential to achieve faster convergence.

3.2 The Search

Guided by the model, Bayesian optimisation chooses to evaluate f at locations specified by a simpler inner optimisation problem. Reminiscent of a bandit policy, this must trade off exploration and exploitation by balancing areas of high uncertainty and areas of low mean and low uncertainty. In standard myopic form, where only one point is acquired at a time, the $n+1^{\mathrm{th}}$ location for evaluation of f is chosen as the maximum of acquisition function $\alpha_n : \mathcal{X} \to \mathbb{R}$

$$x_{n+1} = \operatorname*{argmax}_{x \in \mathcal{X}} \alpha_n(x).$$

Framing the acquisition function as an expected utility over the posterior of f, exploitation-focused heuristics such as the expected improvement (EI) [7] and the probability of improvement (PI) [8] from the current best $y_n^* = \min_{1 \leqslant i \leqslant n} y_i$ recover an efficiently gradient-optimisable closed form for the acquisition function in the noiseless evaluation case

$$\begin{aligned} \alpha_n^{\mathrm{EI}}(x) &= \mathbb{E}_{f(x)|\mathcal{D}_n}\left[\max(y_n^* - f(x), 0)\right] \\ &= (y_n^* - \mu_n(x))\Phi\left(u_n^*(x)\right) + \phi\left(u_n^*(x)\right), \end{aligned} \qquad \begin{aligned} \alpha_n^{\mathrm{PI}}(x) &= \mathbb{P}_{f(x)|\mathcal{D}_n}\left(y_n^* > f(x)\right) \\ &= \Phi\left(u_n^*(x)\right) \end{aligned}$$

where $u_n^*(x) = \left(y_n^* - \mu_n(x)\right)/\sigma_n(x)$ and Φ and ϕ denote the standard univariate normal cumulative distribution and probability density functions. In practice, subtracting a small offset $\xi > 0$ from y_n^*

to encourage an improvement of at least $\xi > 0$ may speed up the procedure. In the noisy case, the "improvement" is more difficult to define, as the current best is now an inexact evaluation of f, so care should be taken when using these methods in such situations. All methods that follow handle the noisy evaluation case without issue.

Near-optimal theoretical rates of convergence for ϵ-greedy EI have been showed in [9]. EI also performs well in practice, perhaps most notably being used for the tuning the neural network in Deepmind's AlphaGo [10]. However, it can be quite sensitive to the choice of initial points of function evaluations and can also too greedy for functions that are sufficiently irregular.

Taking inspiration from well-established multi-arm bandit policies, confidence bound (GP-UCB) [11] and Thompson sampling [12] approaches have been proposed for Bayesian optimisation, alongside a bandit setting. Both are provably convergent, the latter being optimally so under stringent conditions. Nonetheless, their focus on cumulative regret minimization make these methods unsatisfactory. In a bandit problem, each evaluation gives a direct reward which accrues as the routine progresses. In a typical optimisation problem, the sole interest is optimising f itself. We are in essence trying to solve a more stringent problem by treating Bayesian optimisation as an infinite-armed bandit, and by relaxing this constraint we can obtain faster convergence.

A line of work based on information theory aims to both offer a more pleasing criterion than greedy local heuristics and solve the intrinsic issues of regret-based methods. Entropy search (ES) [13] and its refinement predictive entropy search (PES) [14] focus on effective exploration, choosing the acquisition function to be the mutual information between the location of the minimum x^* and a new point x under the model

$$\begin{aligned}\alpha_n^{\mathrm{ES}}(x) &= I(f(x); x^* | \mathcal{D}_n) \\ &= H(x^*|\mathcal{D}_n) - \mathbb{E}_{f(x)|\mathcal{D}_n}[H(x^*|\mathcal{D}_n, f(x))] \\ &= H(f(x)|\mathcal{D}_n) - \mathbb{E}_{x^*|\mathcal{D}_n}[H(f(x)|\mathcal{D}_n, x^*)],\end{aligned}$$

with ES using the second formulation and PES the third. Strong performance is shown empirically for both methods and a bound for simple regret is also established for PES. These methods rely on approximations for either the intractable distribution $p(x^*|\mathcal{D}_n)$ or its entropy, so while competitive performance is achieved, the computational complexity of each acquisition step is greatly increased compared to the heuristic or bandit methods. Philosophically pleasing as they might be, ES and PES are thus only suitable for functions that are very expensive to evaluate.

3.3 Max-Value Entropy Search

Max-value entropy search (MES) [15] aims to remedy this issue. Crucially, by focusing on the information between the value of the minimum $y^* = f(x^*)$ and an evaluation point, the acquisition step is greatly simplified and can be computed in much shorter time. The acquisition function is

$$\begin{aligned}\alpha_n^{\mathrm{MES}}(x) &= I(f(x); y^* | \mathcal{D}_n) \\ &= H(f(x)|\mathcal{D}_n) - \mathbb{E}_{y^*|\mathcal{D}_n}[H(f(x)|\mathcal{D}_n, y^*)] \\ &= -\mathbb{E}_{y^*|\mathcal{D}_n}\left[\frac{u^*(x)\phi(u^*(x))}{2(1-\Phi(u^*(x)))} + \log(1-\Phi(u^*(x)))\right],\end{aligned}$$

where $u^*(x) = (y^* - \mu_n(x))/\sigma_n(x)$. Since $p(y^*|\mathcal{D}_n)$ is intractable, the expectation must be approximated by a Monte-Carlo method and can then be quickly gradient-optimised. As opposed to

$p(x^*|\mathcal{D}_n)$ however, the distribution $p(y^*|\mathcal{D}_n)$ is one-dimensional, so both a more efficient sampling scheme can be employed and the number of samples necessary for a good Monte-Carlo approximation can be greatly reduced.

MES offers competitive performance and is provably convergent — assuming that the true f is drawn from a Gaussian process, a bound for simple regret is established by connection to the authors' previous work in [16].

3.4 Further work in Bayesian optimisation

There are two main lines of current research in Bayesian optimisation. One focuses on efficient computation, the other on adapting the methods to search spaces of high dimension.

In practice we have limited serial computing power, but may have generous parallel computing capabilities. Leveraging this at the acquisition step, multi-point acquisition functions allow for function evaluations to be run in parallel, cutting down on time. While previous myopic methods can and have been adapted for this, there are challenges consisting in both the intractability of the acquisition functions (where Monte-Carlo approximations can be made) and the fact that such acquisition functions can be non-convex, rendering them difficult to optimise.

Multi-fidelity Bayesian optimisation offers a convincing framework for cutting down computation cost by considering less expensive approximations to f. Such lower fidelity information sources are often possible to obtain in practice, for example cheap computer simulations approximating the real behaviour of a robot. GP-UCB [17] and MES [18] have been adapted for this task, both at a significant speedup.

As search space dimensions increase, both the number of points needed for a well-fitting GP and the number of iterations of gradient methods used to optimise the acquisition function increase exponentially. In order to make Bayesian optimisation feasible in such a regime, a promising direction of work assumes an additive structure of lower-dimension GPs [19] for the model. This simplifies the problem while retaining generality, and only increases the computation time linearly in the number of dimensions compared to applying Bayesian optimisation to one lower-dimensional component.

There is potential for further research into the model used by Bayesian optimisation. Practical problems exist that cannot be adequately modelled by a GP. Novel statistical approaches could therefore broaden the scope of the methodology.

Finally, there is also a need for deeper theoretical results concerning Bayesian optimisation. While asymptotic results such as [9] exist, finite-time bounds are needed to explain the good empirical performance of Bayesian optimisation methods. These are particularly lacking in the multi-step acquisition case.

References

[1] W. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 1933.

[2] H. Robbins T. Lai. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 1985.

[3] A. Burnetas, M. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 1996.

[4] A. Garivier et al. Explore first, exploit next: the true shape of regret in bandit problems. *Mathematics of Operational Research*, 2019.

[5] P. Auer et al. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 2002.

[6] A. Shah et al. Bayesian optimization using Student-t processes. In *NIPS Workshop on Bayesian Optimisation*, 2013.

[7] J. Močkus. On Bayesian methods for seeking the extremum. In *Proceedings of the Optimization Techniques IFIP Technical Conference*, 1974.

[8] H. Kushner. A new method for locating the maximum point of an arbitrary multipeak curve in the presence of noise. *Journal of Basic Engineering*, 1964.

[9] A. Bull. Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research*, 2011.

[10] Y. Chen et al. Bayesian optimization in AlphaGo. *arXiv preprint*, 2018.

[11] N. Srinivas et al. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th ICML*, 2010.

[12] S. Ghosh K. Basu. Analysis of Thompson sampling for Gaussian process optimization in the bandit setting. *arXiv preprint*, 2018.

[13] P. Hennig, C.J. Schuler. Entropy search for information-efficient global optimization. *Journal of Machine Learning Research*, 2012.

[14] J. Hernández-Lobato et al. Predictive entropy search for efficient global optimization of black-box functions. In *Proceedings of the 27th NIPS*, 2014.

[15] Z. Wang, S. Jegelka. Max-value entropy search for efficient Bayesian optimization. In *Proceedings of the 34th ICML*, 2017.

[16] Z. Wang et al. Optimization as estimation with Gaussian processes in bandit settings. In *Proceedings of the 19th AISTATS*, 2015.

[17] K. Kandasamy et al. Gaussian process bandit optimisation with multi-fidelity evaluations. In *Proceedings of the 30th NIPS*, 2016.

[18] H. Moss. MUMBO: Multi-task max-value Bayesian optimisation. Presented at the *STOR-i Conference*, 2020.

[19] K. Kandasamy et al. High dimensional Bayesian optimisation and bandits via additive models. In *Proceedings of the 32nd ICML*, 2015.