Bayesian Optimization

Suppose we have a function $f\colon \mathcal{X} \to \mathbb{R}$ that we with to minimize on some domain $X \subseteq \mathcal{X}$. That is, we wish to find

$$x^* = \underset{x \in X}{\arg\min}\; f(x).$$

In numerical analysis, this problem is typically called (global) *optimization* and has been the subject of decades of study. We draw a distinction between global optimization, where we seek the absolute optimum in X, and local optimization, where we seek to find a local optimum in the neighborhood of a given initial point x_0.

A common approach to optimization problems is to make some assumptions about f. For example, when the objective function f is known to be convex and the domain X is also convex, the problem is known as *convex optimization* and has been widely studied. Convex optimization is a common tool used across machine learning.

If an exact functional form for f is not available (that is, f behaves as a "black box"), what can we do? *Bayesian optimization* proceeds by maintaining a probabilistic belief about f and designing a so-called *acquisition function* to determine where to evaluate the function next. Bayesian optimization is particularly well-suited to global optimization problems where f is an *expensive* black-box function; for example, evaluating f might require running an expensive simulation. Bayesian optimization has recently become popular for training expensive machine-learning models whose behavior depend in a complicated way on their parameters (e.g., convolutional neural networks). This is an example of the "AutoML" paradigm.

Although not strictly required, Bayesian optimization almost always reasons about f by choosing an appropriate Gaussian process prior:

$$p(f) = \mathcal{GP}(f; \mu, K).$$

Given observations $\mathcal{D} = (\mathbf{X}, \mathbf{f})$,[1] we can condition our distribution on $\mathcal{D}$ as usual:

$$p(f \mid \mathcal{D}) = \mathcal{GP}(f; \mu_{f|\mathcal{D}}, K_{f|\mathcal{D}}).$$

Given this set of observations, how do we select where to observe the function next? The meta-approach in Bayesian optimization is to design an acquisition function $a(x)$. The acquisition function is typically an inexpensive function that can be evaluated at a given point that is commensurate with how desirable evaluating f at x is expected to be for the minimization problem. We then optimize the acquisition function to select the location of the next observation. Of course, we have merely replaced our original optimization problem with another optimization problem, but on a much-cheaper function $a(x)$.

Acquisition Functions

Many acquisition functions can be interpreted in the framework of Bayesian decision theory as evaluating an expected loss associated with evaluating f at a point x. We then select the point with the lowest expected loss, as usual.

In the below we drop the $f \mid \mathcal{D}$ subscripts on the mean μ and covariance K functions for f.

[1] We will assume these observations to be noiseless here, but we could extend the methods here to the noisy case without difficulty.

Probability of improvement

Perhaps the first acquisition function designed for Bayesian optimization was *probability of improvement.* Suppose

$$f' = \min \mathbf{f}$$

is the minimal value of f observed so far. Probability of improvement evaluates f at the point most likely to improve upon this value. This corresponds to the following utility function[2] associated with evaluating f at a given point x:

$$u(x) = \begin{cases} 0 & f(x) > f' \\ 1 & f(x) \leq f'. \end{cases}$$

That is, we receive a unit reward if $f(x)$ turns out to be less than f', and no reward otherwise. The probability of improvement acquisition function is then the expected utility as a function of x:

$$\begin{aligned} a_{\text{PI}}(x) = \mathbb{E}\bigl[u(x) \mid x, \mathcal{D}\bigr] &= \int_{-\infty}^{f'} \mathcal{N}\bigl(f; \mu(x), K(x,x)\bigr)\,\mathrm{d}f \\ &= \Phi\bigl(f'; \mu(x), K(x,x)\bigr). \end{aligned}$$

The point with the highest probability of improvement (the maximal expected utility) is selected. This is the Bayes action under this loss.

Expected improvement

The loss function associated with probability of improvement is somewhat odd: we get a reward for improving upon the current minimum independent of the size of the improvement! This can sometimes lead to odd behavior, and in practice can get stuck in local optima and underexplore globally.

An alternative acquisition function that does account for the size of the improvement is *expected improvement.* Again suppose that f' is the minimal value of f observed so far. Expected improvement evaluates f at the point that, in expectation, improves upon f' the most. This corresponds to the following utility function:

$$u(x) = \max\bigl(0, f' - f(x)\bigr).$$

That is, we receive a reward equal to the "improvement" $f' - f(x)$ if $f(x)$ turns out to be less than f', and no reward otherwise. The expected improvement acquisition function is then the expected utility as a function of x:

$$\begin{aligned} a_{\text{EI}}(x) = \mathbb{E}\bigl[u(x) \mid x, \mathcal{D}\bigr] &= \int_{-\infty}^{f'} (f' - f)\,\mathcal{N}\bigl(f; \mu(x), K(x,x)\bigr)\,\mathrm{d}f \\ &= \bigl(f' - \mu(x)\bigr)\Phi\bigl(f'; \mu(x), K(x,x)\bigr) + K(x,x)\mathcal{N}\bigl(f'; \mu(x), K(x,x)\bigr). \end{aligned}$$

The point with the highest expected improvement (the maximal expected utility) is selected.

The expected improvement has two components. The first can be increased by reducing the mean function $\mu(x)$. The second can be increased by increasing the variance $K(x,x)$. These two terms can be interpreted as explicitly encoding a tradeoff between *exploitation* (evaluating at points with low mean) and *exploration* (evaluating at points with high uncertainty). The exploitation–exploration tradeoff is a classic consideration in such problems, and the expected improvement criterion *automatically* captures both as a result of the Bayesian decision theoretic treatment.

[2] Recall a utility function is simply a negative loss function.

Entropy search

A third alternative is *entropy search.* Here, we seek to minimize the uncertainty we have in the *location* of the optimal value

$$x^* = \operatorname*{arg\,min}_{x \in X} f(x).$$

Notice that our belief over f induces a distribution over x^*, $p(x^* \mid \mathcal{D})$. Unfortunately, there is no closed-form expression for this distribution.

Entropy search seeks to evaluate points so as to minimize the entropy of the induced distribution $p(x^* \mid \mathcal{D})$. Here the utility function is the reduction in this entropy given a new measurement at x, $\bigl(x, f(x)\bigr)$:

$$u(x) = H[x^* \mid \mathcal{D}] - H\bigl[x^* \mid \mathcal{D}, x, f(x)\bigr].$$

As in probability of improvement and expected improvement, we may build an acquisition function by evaluating the expected utility provided by evaluating f at a point x. Due to the nature of the distribution $p(x^* \mid \mathcal{D})$, this is somewhat complicated, and a series of approximations must be made.

Upper confidence bound

A final alternative acquisition function is typically known as GP-UCB, where UCB stands for *upper confidence bound.* GP-UCB is typically described in terms of maximizing f rather than minimizing f; however in the context of minimization, the acquisition function would take the form

$$a_{\text{UCB}}(x; \beta) = \mu(x) - \beta\sigma(x),$$

where $\beta > 0$ is a tradeoff parameter and $\sigma(x) = \sqrt{K(x,x)}$ is the marginal standard deviation of $f(x)$.[3]

Again, the GP-UCB acquisition function contains explicit exploitation ($\mu(x)$) and exploration ($\sigma(x)$) terms. Interestingly, the acquisition function cannot be interpreted as computing a natural expected utility function. Nonetheless, strong theoretical results are known for GP-UCB, namely, that under certain conditions, the iterative application of this acquisition function will converge to the true global minimum of f.

[3] In the context of minimization, this is better described as a *lower* confidence bound, but UCB is ingrained in the literature as a standard term.