Acquisition Functions

Previously, we introduced two relatively simple acquisition functions for Bayesian optimization: Probability of Improvement (PI) and Expected Improvement (EI). These acquisition functions are commonly used in practice, in part because they permit closed-form, easy-to-optimize expressions when the underlying probabilistic model is a Gaussian process. Below, we will detail a few alternatives, some of which have provable theoretical guarantees, some which result in intractable expressions but have been shown empirically to perform well under various approximate inference techniques. Ultimately, the choice of acquisition function is simply a model hyperparameter that can be tuned to fit the specific task/goals of the practitioner.

For the following discussion, assume that we model the (black-box) objective function as a *noiseless* GP, $f \mid \mathcal{D} \sim \mathcal{GP}(f; \mu_{\mathcal{D}}, k_{\mathcal{D}})$, where $\mathcal{D} = (\mathbf{X}, \mathbf{f})$ is a previously gathered set of observations of f and $\mu_{\mathcal{D}}$, $k_{\mathcal{D}}$ are the posterior mean and covariance functions respectively.

Knowledge Gradient

An alternative interpretation of EI is that it greedily maximizes the following utility function:

$$u'(\mathcal{D}) = \max_{\mathbf{x} \in \mathcal{D}} \mu_{\mathcal{D}}(\mathbf{x}).$$

In our noiseless setting, $\mu_{\mathcal{D}}(\mathbf{x}) = f(\mathbf{x}) \; \forall \; \mathbf{x} \in \mathcal{D}$ so for a fixed dataset, $u'(\mathcal{D}) = \max \mathbf{f} := f'$.

Using this utility function, we can then recreate the point-wise utility we had previously defined as

$$u(\mathbf{x}) = u'\Big(\mathcal{D} \; \cup \; \big(\mathbf{x}, f(\mathbf{x})\big)\Big) - u'(\mathcal{D}) = \max\big(0, f(\mathbf{x}) - f'\big).$$

Under this interpretation of EI, one might notice a bit of an oddity: why should our utility be defined over only points we have observed, $\mathbf{x} \in \mathcal{D}$? Intuitively, this would be like saying that at the end of our budget, we are limited to returning a point in our observed dataset as our "guess" at the optimum of the function.

We can extend EI by instead allowing us to return any arbitrary location in the domain after our final guess, including ones that we never observed during experimentation. This gives rise to the **knowledge gradient** (KG) acquisition function. The implied utility function is

$$u(\mathbf{x}) = u''\Big(\mathcal{D} \; \cup \; \big(\mathbf{x}, f(\mathbf{x})\big)\Big) - u''(\mathcal{D}) \text{ where } u''(\mathcal{D}) = \max_{\mathbf{x} \in \mathcal{X}} \mu_{\mathcal{D}}(\mathbf{x}).$$

Notably, the dataset utility, u'', is defined over the entire domain, $\mathbf{x} \in \mathcal{X}$. Let $\mu^*_{\mathcal{D}} = \max_{\mathbf{x} \in \mathcal{X}} \mu_{\mathcal{D}}(\mathbf{x})$.

The resulting acquisition function is

$$\begin{aligned} a_{\text{KG}}(\mathbf{x}) = \mathbb{E}\big[u(\mathbf{x}) \mid \mathbf{x}, \mathcal{D}\big] &= \int \left(\max_{\mathbf{v} \in \mathcal{X}} \mu_{\mathcal{D} \cup \left(\mathbf{x}, f(\mathbf{x})\right)}(\mathbf{v})\right) p\big(f(\mathbf{x}) \mid \mathbf{x}, \mathcal{D}\big) df(\mathbf{x}) - \max_{\mathbf{v} \in \mathcal{X}} \mu_{\mathcal{D}}(\mathbf{v}) \\ &= \int \left(\max_{\mathbf{v} \in \mathcal{X}} \mu_{\mathcal{D} \cup \left(\mathbf{x}, f(\mathbf{x})\right)}(\mathbf{v})\right) \mathcal{N}\big(f(\mathbf{x}); \mu_{\mathcal{D}}(\mathbf{x}), k_{\mathcal{D}}(\mathbf{x}, \mathbf{x})\big) df(\mathbf{x}) - \mu^*_{\mathcal{D}}. \end{aligned}$$

Intuitively, this integral averages all possible updates to the posterior mean that could occur if we were to observe the function at location $\mathbf{x}$. Unfortunately, this integral is intractable in all but a handful of degenerate cases (e.g., when the domain is discrete or one-dimensional and specific covariance function is used). As such, numerical integration techniques must be used to approximate

this integral; we will discuss many of these at length shortly but for the sake of example, a simple approximation would be the Monte Carlo estimate

$$\begin{aligned} a_{\text{KG}}(\mathbf{x}) &= \int \left(\max_{\mathbf{v} \in \mathcal{X}} \mu_{\mathcal{D} \cup \left(\mathbf{x}, f(\mathbf{x})\right)}(\mathbf{v}) \right) \mathcal{N}\big(f(\mathbf{x}); \mu_{\mathcal{D}}(\mathbf{x}), k_{\mathcal{D}}(\mathbf{x}, \mathbf{x})\big) df(\mathbf{x}) - \mu^*_{\mathcal{D}} \\ &\approx \left(\frac{1}{S} \sum_{s=1}^{S} \left(\max_{\mathbf{v} \in \mathcal{X}} \mu_{\mathcal{D} \cup \left(\mathbf{x}, f_s\right)}(\mathbf{v}) \right) \right) - \mu^*_{\mathcal{D}} \text{ where } f_s \sim \mathcal{N}\big(f(\mathbf{x}); \mu_{\mathcal{D}}(\mathbf{x}), k_{\mathcal{D}}(\mathbf{x}, \mathbf{x})\big), \end{aligned}$$

which simply draws (finitely many) samples from the posterior belief about $f(\mathbf{x})$, treats each sample as if it were the true observed function value and computes what the maximum of the posterior mean would be under each sample.

Empirically, it has been found that the knowledge gradient acquisition function outperforms EI, particularly in settings with noise (which admittedly, we have explicitly excluded from this discussion).

Upper confidence bound

The Bayesian optimization literature draws heavily from the work done in multi-armed bandit settings, which can be viewed as a noisy, discrete-domain formulation of the Bayesian optimization problem. One relatively famous acquisition function inspired by this line of research is known as the UCB or **upper confidence bound** acquisition function. This acquisition function is somewhat difficult to define in terms of an implied utility function for Bayesian decision theory (although in certain cases, it is possible to construct one); instead, we will simply consider its functional form:

$$a_{\text{UCB}}(\mathbf{x}; \beta) = \mu_{\mathcal{D}}(\mathbf{x}) + \beta \sqrt{k_{\mathcal{D}}(\mathbf{x}, \mathbf{x})}.$$

We refer to this (somewhat erroneously) as an upper confidence bound as it corresponds to an upper bound on the α-credible interval of our posterior belief about the function's value at $\mathbf{x}$, where $\alpha = \Phi(\beta)$.

Similar to the ε parameter for PI, $\beta > 0$ is a parameter that controls the exploration-exploitation tradeoff; like EI, that tradeoff can be explicitly seen in the functional form of $a_{\text{UCB}}(\mathbf{x}; \beta)$, where the $\mu(x)$ term favors exploitation and the $\sqrt{k_{\mathcal{D}}(\mathbf{x}, \mathbf{x})}$ prioritizes exploration.

Figure 1 shows the effect of the tradeoff parameter (in terms of α rather than β): as α increases, UCB tends to prefer exploration over exploitation and prioritizes observations farther from previously observed locations. We can see this in the sample run also depicted in Figure 1: relative to EI, this acquisition function tends to space out its observations more evenly over the domain (at least for this setting of α), even going so far as to depart from the neighborhood of the true optimal value after discovering it (the two tick marks circled in red in Figure 1).

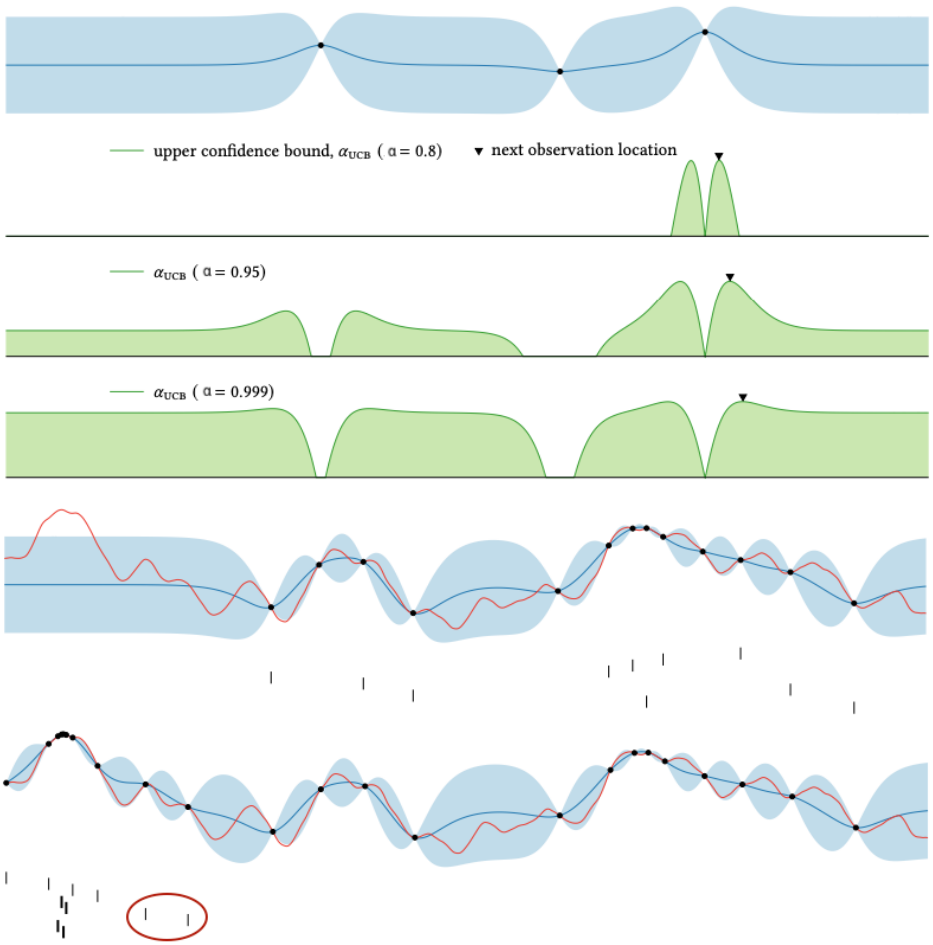

Figure 1: The upper confidence bound acquisition function evaluated on the running example from the previous notes (top), presented in terms of three different values of α, as well as the first ten observation locations selected by this policy when $\alpha = 0.999$ (middle) and the next ten observation locations (bottom). The height of the tick marks indicates the relative order that points are selected to be observed in, with higher tick marks being selected earlier; the thicker tick marks indicate observations with 0.2 units of the true optimum.

Regret Bounds

One major advantage of this acquisition function is that it enjoys strong theoretical guarantees under certain assumptions about the quality of the GP's fit to the underlying objective function. These theoretical guarantees are defined in terms of the *regret* of an acquisition function:

$$R_T = \sum_{t=1}^{T} f(\mathbf{x}^*) - f(\mathbf{x}_t)$$

where T is our budget of observations, $f(\mathbf{x}^*)$ is the optimal objective function value, and $\mathbf{x}_t$ is the t^{th} observation selected by the acquisition function.

Srinivas et al. (2010) proved that the regret of the UCB acquisition function is bounded according to

$$p\left(R_T \leq O^*\left(\sqrt{T\beta_T\gamma_T}\right)\right) \geq 1 - \delta$$

for some $\delta \in (0, 1)$, where O^* is equivalent to O-notation with polylogarithmic factors suppressed. The bound above holds if the following conditions are satisfied; the conditions are somewhat technical if rigorously defined so we provide an intuitive description of the conditions instead:

- the domain $\mathcal{X}$ is compact and convex,
- the objective function is "well-modeled" by the GP belief,
- β_t follows a schedule that increases as t increases and/or δ decreases,
- γ_t is some measure of how "informative" point-wise observations of the function are about the function as a whole.

Thompson Sampling

Another acquisition function with its roots in the multi-armed bandit literature is known as **Thompson sampling** (TS). Unlike the other acquisition functions we will consider, TS is an inherently stochastic policy.[1]

We can define the TS acquisition function as

$$a_{\text{TS}}(\mathbf{x}) = g(\mathbf{x}) \text{ where } g \sim \mathcal{GP}(f; \mu_{\mathcal{D}}, k_{\mathcal{D}}).$$

In words, Thompson sampling simply operates by drawing a sample path from the posterior GP belief on the objective function and observing the location where the sample is maximized.

While seemingly straightforward, exactly optimizing a sample path from a GP belief is non-trivial. Two common approaches are described below.

"Brute-force" or Exhaustive Sampling

For sufficiently low-dimensional (or ideally, discrete) domains, a reasonable approximation is to sample from the GP belief at a dense grid of locations, typically evenly spaced according to some low-discrepancy sequence, and then observe the location with the highest sampled function value. This amounts to simply sampling from a multivariate Gaussian distribution, which can be done

[1] Some of the approximations we will introduce for the intractable acquisition functions introduce randomness in the approximation but the functions we are approximately are themselves deterministic quantities.

efficiently for a small-ish number of locations (i.e., $< 100,000$) but high-dimensional domains or objective functions with very short relative length-scales, this may not give sufficient coverage of the possible inputs to accurately estimate the sample path's maximum.

Figure 2 shows the (approximate) distribution of the optimal location to observe,

$$p(x^* \mid \mathcal{D}) \text{ where } x^* = \arg\max_{x \in \mathcal{X}} f(x),$$

on our 1-dimensional running example.

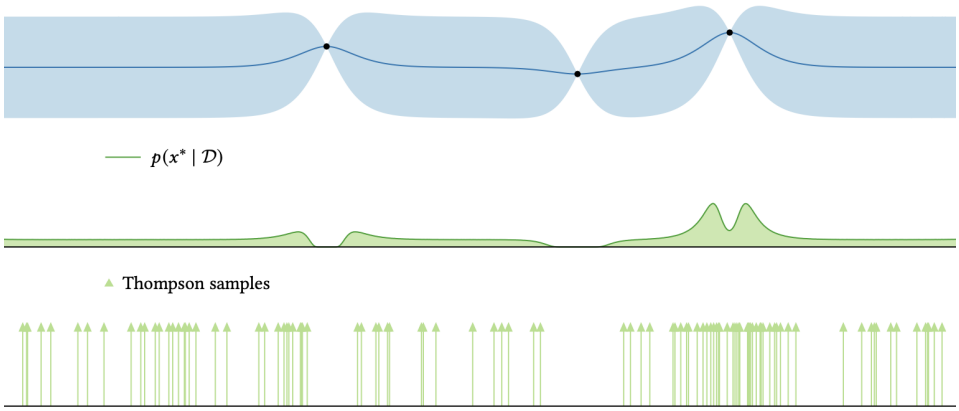

Figure 2: Thompson sampling used to approximate the distribution of the objective function's arg max: to generate the distribution (middle), 100 sample paths were drawn from the posterior GP belief (top) over a grid of 1000 evenly spaced locations; the location of the arg max of each sample path is shown as well (bottom).

On-demand Sampling

If exhaustive sampling is prohibitively expensive, we can rely on iterative optimization algorithms (e.g., gradient descent) which only query a single point from our sample path at a time. Whenever our optimizer requires a new observation from our sample path, we simply draw a univariate sample from the Gaussian posterior belief at the requested location, conditioned on all the previous samples. It turns out that with our GP belief, it is even possible to sample a gradient of our sample path at any given location!

We will discuss this property and its implications at a later point but one key consideration is that incorporating gradient samples can significantly increase computational costs as we are effectively going from a single Gaussian sample in each iteration to $d+1$ samples, where d is the dimensionality of the domain (recall that the gradient contains an element for each dimension of the input).

Entropy search

Finally, a recently proposed class of acquisitions functions leverages the notion of *mutual information* to guide the selection of observations; these are typically called some variant of the phrase **entropy search.** Entropy search acquisition functions seek to minimize the uncertainty we have about some quantity of interest. In the context of Bayesian optimization, there are two obvious quantities that we might wish to gain information about:

1. the optimal objective function value $f^* = \max_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$
2. the location of the optimal objective function value $\mathbf{x}^* = \arg\max_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$

Entropy search treats these quantities as random variables and seeks to evaluate points so as to maximize the mutual information between them and the selected observations. Recall that the mutual information between two random variables, X and Y, is given by

$$I(X;Y) = H(X) - \mathbb{E}_Y\left[H(X \mid Y)\right]$$

where $H(X) = \mathbb{E}_X[\log(X)]$ is the entropy of X. One key property of mutual information that we exploit shortly is symmetry: $I(X;Y) = I(Y;X)$.

Let ω be either of the two quantities of interest above. The implied utility function of these acquisition functions is then

$$u'(\mathcal{D}) = H(\omega) - H(\omega \mid \mathcal{D})$$

which gives rise to the point-wise utility function

$$\begin{aligned} u(\mathbf{x}) &= \mathbb{E}_{f(\mathbf{x})}\Big[u'(\mathcal{D} \cup (\mathbf{x}, f(\mathbf{x}))\Big] - u'(\mathcal{D}) \\ &= H(\omega \mid \mathcal{D}) - \mathbb{E}_{f(\mathbf{x})}\Big[H\Big(\omega \mid \mathcal{D} \cup (\mathbf{x}, f(\mathbf{x}))\Big)\Big] = I\big(\omega; f(\mathbf{x}) \mid \mathbf{x}, \mathcal{D}\big). \end{aligned}$$

Figure 3 compares these two acquisition functions for both $\omega = \mathbf{x}^*$ and $\omega = f^*$.

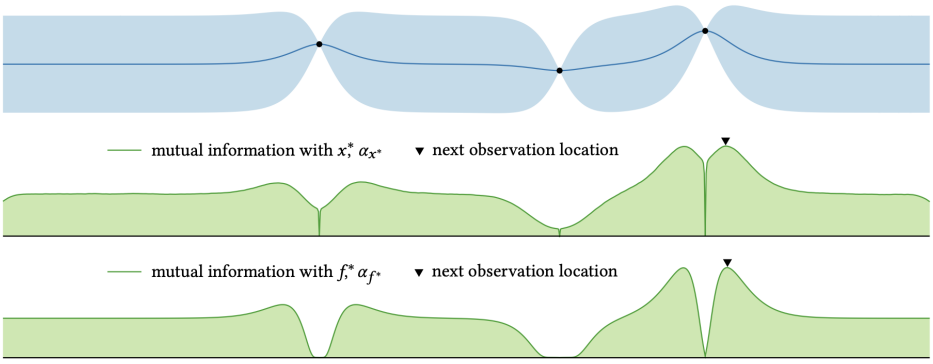

Figure 3: Entropy search for mutual information with respect to $\mathbf{x}^*$ (middle) and f^* (bottom). Note the dips in the $\text{ES}_{\mathbf{x}^*}$ acquisition function at the boundary: these dips are absent in all of the acquisition functions we've explored to date. This indicates that $\text{ES}_{\mathbf{x}^*}$ is less interested in observing the boundary because it can only decrease our uncertainty about the optimal location "in one direction"; observations further from the boundary reduce the uncertainty over a larger portion of the domain.

Unfortunately, this acquisition function is intractable for both $\mathbf{x}^*$ and f^*; instead, a somewhat complicated sequence of approximations must be made in either case to estimate these acquisition functions. For brevity, we will present just an (abridged) approximation for $\text{ES}_{\mathbf{x}^*}$; many corresponding approximations for ES_{f^*} follow a similar structure to the one presented below.

Mutual Information with $\mathbf{x}^*$

Instead of computing $I\big(\mathbf{x}^*; f(\mathbf{x}), \mathcal{D}\big)$, we will instead consider approximating the equivalent

$$I\big(f(\mathbf{x}); \mathbf{x}^* \mid \mathbf{x}, \mathcal{D}\big) = H\big(f(\mathbf{x}) \mid \mathbf{x}, \mathcal{D}\big) - \mathbb{E}_{\mathbf{x}^*}\Big[H\big(f(\mathbf{x}) \mid \mathbf{x}^*, \mathbf{x}, \mathcal{D}\big)\Big].$$

Luckily, the first term is simple under the GP belief:

$$H\big(f(\mathbf{x}) \mid \mathbf{x}, \mathcal{D}\big) = \frac{1}{2}\log\left(2\pi e\, k_{\mathcal{D}}(\mathbf{x}, \mathbf{x})\right).$$

The second term is harder to approximate. We begin by approximating the expectation via Monte Carlo integration:

$$\mathbb{E}_{\mathbf{x}^*}\Big[H\big(f(\mathbf{x}) \mid \mathbf{x}^*, \mathbf{x}, \mathcal{D}\big)\Big] \approx \frac{1}{S}\sum_{s=1}^{S} H\big(f(\mathbf{x}) \mid \mathbf{x}_s^*, \mathbf{x}, \mathcal{D}\big) \text{ where } \mathbf{x}_s^* \sim p(\mathbf{x}^* \mid \mathcal{D}).$$

The necessary samples of $\mathbf{x}^*$ can be drawn using Thompson sampling as detailed above.

The distribution $p\big(f(\mathbf{x}) \mid \mathbf{x}^*, \mathbf{x}, \mathcal{D}\big)$ is a strange one: intuitively, this distribution poses the question "how would knowing the *location* of the objective function's optimum affect my belief about the functions value at some other location?"

If we could approximate this distribution to be Gaussian,

$$p\big(f(\mathbf{x}) \mid \mathbf{x}^*, \mathbf{x}, \mathcal{D}\big) \approx \mathcal{N}(f(\mathbf{x}); m^*, \sigma^{2*}),$$

then we would once again have a simple expression for the entropy in question:

$$H\big(f(\mathbf{x}) \mid \mathbf{x}_s^*, \mathbf{x}, \mathcal{D}\big) \approx \frac{1}{2}\log\left(2\pi e\, \sigma_s^{2*}\right)$$

Putting this together with the previous term gives the final approximation for the $\text{ES}_{\mathbf{x}^*}$ acquisition function:

$$a_{\text{ES}} \approx \frac{1}{2}\log\left(2\pi e\, k_{\mathcal{D}}(\mathbf{x}, \mathbf{x})\right) - \frac{1}{S}\sum_{s=1}^{S}\frac{1}{2}\log\left(2\pi e\, \sigma_s^{2*}\right) = \log\left(\sqrt{k_{\mathcal{D}}(\mathbf{x}, \mathbf{x})}\right) - \frac{1}{S}\sum_{s=1}^{S}\log\left(\sigma_s^*\right).$$

Of course, the key unanswered question is how to approximate $p\big(f(\mathbf{x}) \mid \mathbf{x}^*, \mathbf{x}, \mathcal{D}\big)$ as a Gaussian. Hernández-Lobato et al. (2014) proposed using expectation propagation (EP) to do so, which gives rise to the **predictive entropy search** or PES acquisition function.

At a high level, their proposed EP approximation begins with $p\big(f(\mathbf{x}) \mid \mathbf{x}, \mathcal{D}\big)$ as the "prior", $p_0(f(\mathbf{x}))$. They then multiply this prior by a sequence of "likelihood" terms which correspond to enforcing that $\mathbf{x}^*$ satisfies certain properties of a global maximum. At a high level, there is

- a term that enforces the gradient at $\mathbf{x}^*$ is $\mathbf{0}$,
- a term that enforces the Hessian at $\mathbf{x}^*$ is negative semi-definite,
- a term that enforces $f^* > f'$ (recall that f' is the maximum observed function value in $\mathcal{D}$), and
- a term that enforces $f^* > f(\mathbf{x})$.

The first two terms correspond to $\mathbf{x}^*$ being a *local* maximum whereas the second two terms loosely enforce that $\mathbf{x}^*$ is globally optimal: they are necessary but not sufficient conditions although given that this is already an approximation, the somewhat loose correspondence can be forgiven.

Figure 4 shows the EP approximated posterior of f using this set of implied constraints: the effect is minimal to non-existent at locations far from $\mathbf{x}^*$ but the (approximate) conditioning on $\mathbf{x}^*$ has a clear impact on the posterior belief of nearby function values.

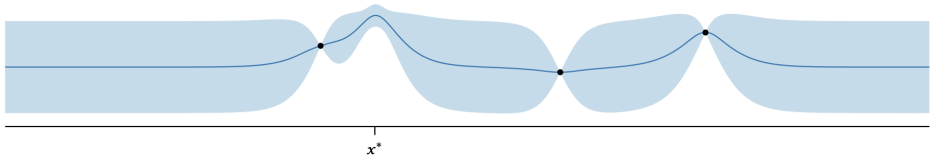

Figure 4: The approximate posterior belief on the objective function after using the PES EP approximation. The location being conditioned on was sampled using Thompson sampling.