Bayesian Optimization with Linear Combination Observations

Firstname1 Lastname1 [*] [1] **Firstname2 Lastname2** [*] [1] [2] **Firstname3 Lastname3** [2]

Abstract

1. Setup

Assume a discrete (and finite) candidate set $\mathcal{X}$ with $|\mathcal{X}| = N$. Our goal is to choose $\arg\max_{\boldsymbol{x} \in \mathcal{X}} f^*(\boldsymbol{x})$, where $f : \mathcal{X} \to \mathbb{R}$ is some unknown latent function. The vector $\boldsymbol{f} \in \mathbb{R}^N$ will be the vector of responses from applying $f(\cdot)$ to all inputs $\boldsymbol{x} \in \mathcal{X}$.

Surrogate model. We will use a zero-mean Gaussian process as a surrogate model for optimization. Let $k_{\boldsymbol{\theta}}(\cdot, \cdot)$ be the covariance function with hyperparameters $\boldsymbol{\theta}$. (Whenever obvious, we will drop the subscript $\boldsymbol{\theta}$.) $\boldsymbol{K} \in \mathbb{R}^{N \times N}$ will refer to the Gram matrix formed from applying $k(\cdot, \cdot)$ to all pairs of inputs $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$.

Observation model. At iteration step i, we observe a noisy linear combination of the response variables. In other words, we select an *action*—a set of linear combination weights $\boldsymbol{s}_i \in \mathbb{R}$—and then we observe $\alpha_i = \boldsymbol{s}_i^\top \boldsymbol{f} + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2_{\text{obs}})$ is i.i.d. observational noise.[1] Without loss of generality, we assume $\|\boldsymbol{s}_i\|_2 = 1$ for all i. Letting $\boldsymbol{\alpha}_i = \begin{bmatrix} \alpha_1 & \cdots & \alpha_i \end{bmatrix}$, we have that

$$\boldsymbol{\alpha}_i \sim \mathcal{N}\left(\boldsymbol{0},\ \boldsymbol{S}_i^\top \boldsymbol{K} \boldsymbol{S}_i + \sigma^2 \boldsymbol{I}\right), \tag{1}$$

where $\boldsymbol{S}_i \in \mathbb{R}^{N \times i}$ is the concatenation of all of the linear combination action vectors.

Note that this setup reduces to the standard Bayesian optimization setup if the $\boldsymbol{s}_i$ are restricted to be unit vectors $\boldsymbol{e}_1$, ..., $\boldsymbol{e}_N$. However, by allowing $\boldsymbol{s}_i$ to be the space of all possible linear combinations, we hope to achieve faster convergence.

2. Information Theoretic Acquisition Functions for Linear Combinations

First we will reframe our goal through a general decision theoretic lens. Given a latent random variable $\boldsymbol{\omega}$, assume that we wish to maximize the following information theoretic utility function after T steps:

$$\mathcal{U}(\mathcal{D}_T) = \mathbb{H}\left[\boldsymbol{\omega}\right] - \mathbb{H}\left[\boldsymbol{\omega} \mid \mathcal{D}_T\right], \tag{2}$$

where $\mathcal{D}_i = \{(\boldsymbol{s}_i, \alpha_i)\}_i$ is the set of all linear combination actions and associated noisy observations.

At any given timestep, assume that we optimize the entropy search acquisition function (Wang & Jegelka, 2017) which corresponds to the one-step lookahead optimal action according to Eq. (2):

$$\begin{aligned} a_{\boldsymbol{\omega}}(\boldsymbol{s}_{i+1}, \mathcal{D}_i) &:= \mathbb{H}\left[\alpha_{i+1} \mid \boldsymbol{s}_{i+1}, \mathcal{D}_i\right] \\ &\quad - \mathop{\mathbb{E}}_{\boldsymbol{\omega} \mid \mathcal{D}_i}\left[\mathbb{H}\left[\alpha_{i+1} \mid \boldsymbol{s}_{i+1}, \mathcal{D}_i, \boldsymbol{\omega}\right]\right]. \end{aligned} \tag{3}$$

Following the analysis of Wenger et al. (2022, Sec. 3), the distribution $\alpha_{i+1} \mid \boldsymbol{s}_{i+1}, \mathcal{D}_i$ is $\mathcal{N}(\mu_{|\mathcal{D}_i}, \sigma^2_{|\mathcal{D}_i})$, where

$$\begin{aligned} \mu_{|\mathcal{D}_i}(\boldsymbol{s}_{i+1}) &:= \boldsymbol{s}_{i+1}^\top \boldsymbol{K} \boldsymbol{C}_i^\top \boldsymbol{S}_i^{\dagger\top} \boldsymbol{\alpha}_i \\ \sigma^2_{|\mathcal{D}_i}(\boldsymbol{s}_{i+1}) &:= \boldsymbol{s}_{i+1}^\top \left(\boldsymbol{K} - \boldsymbol{K} \boldsymbol{C}_i \boldsymbol{K}\right) \boldsymbol{s}_{i+1} + \sigma^2_{\text{obs}}. \end{aligned} \tag{4}$$

where $\boldsymbol{S}_i^\dagger = (\boldsymbol{S}_i^\top \boldsymbol{S}_i)^{-1} \boldsymbol{S}_i^\top$ is the pseudoinverse of $\boldsymbol{S}$ and $\boldsymbol{K}_{|\mathcal{D}_i}$ is the posterior covariance of $\boldsymbol{f}$, given by

$$\begin{aligned} \boldsymbol{K}_{|\mathcal{D}_i} &:= \boldsymbol{K} - \boldsymbol{K} \boldsymbol{C}_i \boldsymbol{K}, \\ \boldsymbol{C}_i &:= \boldsymbol{S}_i \left(\boldsymbol{S}_i^\top \boldsymbol{K} \boldsymbol{S}_i + \sigma^2_{\text{obs}} \boldsymbol{I}\right)^{-1} \boldsymbol{S}_i. \end{aligned} \tag{5}$$

Applying standard Gaussian conditioning rules, $\boldsymbol{K}_{|\mathcal{D}_i}$ obeys the following recursive formula:

$$\boldsymbol{K}_{|\mathcal{D}_{i+1}} = \boldsymbol{K}_{|\mathcal{D}_i} - \frac{\boldsymbol{K} \boldsymbol{s}_{i+1} \boldsymbol{s}_{i+1}^\top \boldsymbol{K}}{\boldsymbol{s}_{i+1}^\top \boldsymbol{K} \boldsymbol{s}_{i+1} + \sigma^2_{\text{obs}}} \tag{6}$$

2.1. Active Learning

In an active learning setting, a reasonable utility function to optimize is $\mathcal{U}(\mathcal{D}_T) = \mathbb{H}\left[\boldsymbol{f}\right] - \mathbb{H}\left[\boldsymbol{f} \mid \mathcal{D}_T\right]$—i.e. we wish to minimize the entropy (uncertainty) over our function. Given our observation model, we have that $\alpha_{i+1} \mid \boldsymbol{s}_{i+1}, \mathcal{D}_i, \boldsymbol{f} \sim \mathcal{N}(0, \sigma^2_{\text{obs}})$, and thus $\mathbb{H}[\alpha_{i+1} \mid \boldsymbol{s}_{i+1}, \mathcal{D}_i, \boldsymbol{f}]$,

[*] Equal contribution [1] Department of XXX, University of YYY, Location, Country [2] Company Name, Location, Country. Correspondence to: Firstname1 Lastname1 <first1.last1@xxx.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1] Alternatively, we could assume that we observe all $f(\boldsymbol{x})$ with i.i.d. observational noise $\mathcal{N}(0, \sigma^2_{\text{obs}})$. These two assumptions are equivalent by linear Gaussian identities.

is a constant for all $\boldsymbol{s}_{i+1}$. Therefore, maximizing Eq. (3) corresponds to:

$$\begin{aligned}\max_{\boldsymbol{s}_{i+1}} a_{\boldsymbol{f}}(\boldsymbol{s}_{i+1}, \mathcal{D}_i) &= \max_{\boldsymbol{s}_{i+1}} \mathbb{H}\left[\alpha_{i+1} \mid \boldsymbol{s}_{i+1}, \mathcal{D}_i\right] \\ &= \max_{\boldsymbol{s}_{i+1}} \sigma^2_{|\mathcal{D}_i}(\boldsymbol{s}_{i+1}) \\ &= \max_{\boldsymbol{s}_{i+1}} \boldsymbol{s}_{i+1}^\top \left(\boldsymbol{K}_{|\mathcal{D}_i}\right) \boldsymbol{s}_{i+1} \qquad (7)\end{aligned}$$

Recognizing Eq. (7) as a Raleigh quotient, the optimal $\boldsymbol{s}_{i+1}$ is the top eigenvector of $\boldsymbol{K}_{|\mathcal{D}_i}$.

At the subsequent iteration, the optimal $\boldsymbol{s}_{i+1}$ will be equal to the top eigenvector of $\boldsymbol{K}_{|\mathcal{D}_{i+1}}$. Following the recursive formula of Eq. (6), we can rewrite Eq. (7) at iteration $i+2$ as:

$$\begin{aligned}\max_{\boldsymbol{s}_{i+2}} a_{\boldsymbol{f}}(\boldsymbol{s}_{i+2}, \mathcal{D}_{i+1}) &= \max_{\boldsymbol{s}_{i+2}} \boldsymbol{s}_{i+2}^\top \left(\boldsymbol{K}_{|\mathcal{D}_{i+1}}\right) \boldsymbol{s}_{i+2} \\ &= \max_{\boldsymbol{s}_{i+2}^\top \boldsymbol{K} \boldsymbol{s}_{i+1} = 0} \boldsymbol{s}_{i+2}^\top \left(\boldsymbol{K}_{|\mathcal{D}_i}\right) \boldsymbol{s}_{i+2},\end{aligned}$$

which corresponds to maximizing the Raleigh quotient while being K-orthogonal to $\boldsymbol{s}_{i+1}$. If $\boldsymbol{s}_{i+1}$ is the top eigenvector of $\boldsymbol{K}_{|\mathcal{D}_i}$, then $\boldsymbol{s}_{i+2}$ will be the second largest eigenvector of $\boldsymbol{K}_{|\mathcal{D}_i}$.

Optimality of one-step lookahead. Continuing this argument via induction will show that $\max_{\boldsymbol{s}_{i+j}} a_{\boldsymbol{f}}(\boldsymbol{s}_{i+j}, \mathcal{D}_{i+j-1})$ is the j^{th} top eigenvector of $\boldsymbol{K}_{|\mathcal{D}_i}$. Note that this is the non-myopic maximum as well: if we were to select the actions $\boldsymbol{s}_{i+1}, \ldots, \boldsymbol{s}_{i+j}$ all at once, Eq. (7) would correspond to:

$$\begin{aligned}&\max_{\boldsymbol{S}_{i+1:i+j}} a_{\boldsymbol{f}}(\boldsymbol{S}_{i+1:i+j}, \mathcal{D}_i) \\ &= \max_{\boldsymbol{S}_{i+1:i+j}} \boldsymbol{S}_{i+1:i+j}^\top \left(\boldsymbol{K}_{|\mathcal{D}_i}\right) \boldsymbol{S}_{i+1:i+j},\end{aligned}$$

which would also correspond to choosing the top-j eigenvectors of $\boldsymbol{K}_{|\mathcal{D}_i}$.

Conjugate gradient actions as an approximation. Optimizing Eq. (7) may in many instances be too computationally expensive. Using a single power iteration [GP: (I think)] as a rough approximation would correspond to the $\boldsymbol{s}_{i+1}$ vectors equaling the steps taken by the conjugate gradients algorithm to solve $(\boldsymbol{K} + \sigma^2_{\text{obs}}\boldsymbol{I})^{-1}(\boldsymbol{f} + \boldsymbol{\epsilon})$, where the entries of $\boldsymbol{\epsilon} \in \mathbb{R}^N$ are i.i.d. $\mathcal{N}(0, \sigma^2_{\text{obs}})$.

2.2. Optimization with Max-Value Entropy Search

For optimization purposes, we might care about the utility $\mathcal{U}(\mathcal{D}_T) = \mathbb{H}[f^*] - \mathbb{H}[f^* \mid \mathcal{D}_T]$, where $f^* = \max_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x})$ is the maximum value of f.

Given a sample of f, let $\boldsymbol{e}^*$ correspond to the unit vector where $\boldsymbol{e}^{*\top}\boldsymbol{f} = f^*$. Note that

$$\mathbb{V}\left[\alpha_{i+1} \mid \boldsymbol{s}_{i+1}, \mathcal{D}_i, f^*\right] = \sigma^2_{|\mathcal{D}_i}(\boldsymbol{s}_{i+1}) - \frac{\left(\boldsymbol{s}_{i+1}^\top \boldsymbol{K}_{|\mathcal{D}_i} \boldsymbol{e}^*\right)^2}{\sigma^2_{|\mathcal{D}_i}(\boldsymbol{e}^*)}.$$

Thus, maximizing Eq. (3) corresponds to the following problem:

$$\begin{aligned}&\max_{\boldsymbol{s}_{i+1}} a_{f^*}(\boldsymbol{s}_{i+1}, \mathcal{D}_i) \\ &= \max_{\boldsymbol{s}_{i+1}} \mathop{\mathbb{E}}_{f^* | \mathcal{D}_i} \left[\log\left(\frac{\sigma^2_{|\mathcal{D}_i}(\boldsymbol{s}_{i+1})}{\mathbb{V}[\alpha_{i+1} | \boldsymbol{s}_{i+1}, \mathcal{D}_i, f^*]}\right)\right] \\ &= \min_{\boldsymbol{s}_{i+1}} \mathop{\mathbb{E}}_{f^* | \mathcal{D}_i} \left[\log\left(\frac{\mathbb{V}[\alpha_{i+1} | \boldsymbol{s}_{i+1}, \mathcal{D}_i, f^*]}{\sigma^2_{|\mathcal{D}_i}(\boldsymbol{s}_{i+1})}\right)\right] \\ &= \min_{\boldsymbol{s}_{i+1}} \mathop{\mathbb{E}}_{f^* | \mathcal{D}_i} \left[\log\left(\frac{\sigma^2_{|\mathcal{D}_i}(\boldsymbol{s}_{i+1}) - \left(\boldsymbol{s}_{i+1}^\top \boldsymbol{K}_{|\mathcal{D}_i} \boldsymbol{e}^*\right)^2 / \sigma^2_{|\mathcal{D}_i}(\boldsymbol{e}^*)}{\sigma^2_{|\mathcal{D}_i}(\boldsymbol{s}_{i+1})}\right)\right] \\ &= \min_{\boldsymbol{s}_{i+1}} \mathop{\mathbb{E}}_{f^* | \mathcal{D}_i} \left[\log\left(\frac{\sigma^2_{|\mathcal{D}_i}(\boldsymbol{e}^*) - \left(\boldsymbol{s}_{i+1}^\top \boldsymbol{K}_{|\mathcal{D}_i} \boldsymbol{e}^*\right)^2 / \sigma^2_{|\mathcal{D}_i}(\boldsymbol{s}_{i+1})}{\sigma^2_{|\mathcal{D}_i}(\boldsymbol{e}^*)}\right)\right].\end{aligned} \qquad (8)$$

Connection to Thompson sampling. If approximate the expectation in Eq. (8) with a single Monte Carlo sample, then the solution to Eq. (8) will take the form

$$\begin{aligned}\boldsymbol{s}^*_{i+1} &= \operatorname*{arg\,min}_{\boldsymbol{s}_{i+1}} \left(\sigma^2_{|\mathcal{D}_i}(\boldsymbol{e}^*) - \frac{\left(\boldsymbol{s}_{i+1}^\top \boldsymbol{K}_{|\mathcal{D}_i} \boldsymbol{e}^*\right)^2}{\sigma^2_{|\mathcal{D}_i}(\boldsymbol{s}_{i+1})}\right) \qquad (9) \\ &= \operatorname*{arg\,min}_{\boldsymbol{s}_{i+1}} \mathbb{V}\left[f^* \mid \mathcal{D}_i, \boldsymbol{s}_{i+1}, \alpha_{i+1}\right].\end{aligned}$$

Choosing $\boldsymbol{s}_{i+1} = \boldsymbol{e}^*$ results in Eq. (9) to be zero. In other words, the optimal solution is the action that would also be proposed by Thompson sampling.

We will therefore only take advantage of the linear combination observations if we choose to approximate Eq. (8) with > 1 Monte Carlo samples.

3. (Potentially) Exponential Convergence for (Maybe 1D?) Bayesian Optimization

Can we use the findings above (or similar derivations with other utility/acquisition functions) to show that Bayesian optimization and/or active learning converges faster if you're allowed to use linear combination observations? [GP: TODO.]

References

Wang, Z. and Jegelka, S. Max-value entropy search for efficient Bayesian optimization. In *ICML*, 2017.

Wenger, J., Pleiss, G., Pförtner, M., Hennig, P., and Cunningham, J. P. Posterior and computational uncertainty in Gaussian processes. *NeurIPS*, 2022.