CS2910 - Optimization Summer 2023

Lecture 10 — Bayesian Optimization

Lecturer: Yunwei Ren *Scribed by Zhidan Li*

Contents

1 Overview

Recall that, in the gradient descent algorithm, we can efficiently solve the optimization problem when:

- The objective function f is convex.
- The dimension d is large.
- The gradient of f can be efficiently computed.

However, when the gradient cannot be computed or it is too expensive to compute it such as hyperparameter tuning (architecture search). To solve this issue, we introduce the method called *Bayesian optimization*.

2 Bayesian Optimization

Now we give the high-level idea of the Bayesian optimization. To implement the Bayesian optimization, the major problem is, how to choose the next point x_{t+1}, based on the current sequence $\{(x_i, f(x_i))\}_{i=1}^{t}$.

- Firstly we 'fit a function' $f_n : \mathbb{R}^d \to \mathbb{R}$ such that $f_n(x_i) = f(x_i)$ for all $i \in [t]$.
- Then we 'optimize' f_n to find the next x_{t+1}.
- After all, we observe $f(x_{t+1})$ and append $(x_{t+1}, f(x_{t+1}))$ to the sequence.

Then it naturally gives rise to the following two questions: How to parameterize the function? Which kinds of functions are good?

The solution to the above questions is, we pick a distribution over all 'nice' functions passing through $(x_i, f(x_i))$ for all $i \in [t]$. It makes sense that, we hope the function will be smooth. That is to say, when x is getting close to x_i, $f(x)$ is also closer to $f(x_i)$.

2.1 Gaussian process regression

Typically we will choose the *Gaussian process* to specify the distribution over all 'nice' functions.

Definition 1 (Gaussian process)**.** *An $\mathbb{R}^d$-indexed stochastic process F is said to be a* Gaussian process *if for all $n \in \mathbb{N}$, $x_1, \ldots, x_n \in \mathbb{R}^d$, there exists a function $\mu : \mathbb{R}^d \to \mathbb{R}$ and $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that*

$$(F(x_1), \ldots, F(x_n)) \sim \mathcal{N}\left([\mu(x_i)]_{i\in[n]}, [K(x_i, x_j)]_{i,j\in[n]}\right).$$

Remark 1. Usually we pick K as the **RBF kernel**, i.e., for some chosen $\sigma \geq 0$, $x, y \in \mathbb{R}^d$,

$$K(x, y) := \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right).$$

Note that, when $\|x - y\|$ is decreasing (x is getting close to y), $K(x, y)$ is increasing, which means $\mathrm{Cov}\,(F(x), F(y))$ becomes larger. This implies $F(x)$ is more correlated with $F(y)$. Also, when σ is decreasing, $K(x, y)$ will be increasing as well. Then $F(x)$ would be more correlated with $F(y)$. The simple observation intuitively tells us the RBF kernel 'truly' controls the smoothness.

Then, we show the law of any $f(x)$ based on $\{(x_i, f(x_i)\}_{i=1}^t$.

Lemma 2. *Let $Y_1 \in \mathbb{R}^{d_1}$, $Y_2 \in \mathbb{R}^{d_2}$ and*

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^\top & \Sigma_{22} \end{bmatrix}\right).$$

Then

$$Y_1 | Y_2 \sim \mathcal{N}\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(Y_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^\top\right).$$

We leave the proof of Lemma 2 as a simple probability exercise.

Definition 3 (Gaussian process regression)**.** *Let $\{(x_i, f(x_i))\}_{i\in[t]}$ be the observed data. The* Gaussian process regression *proceeds as: for every $x \in \mathbb{R}^d$, we set the prior distribution to be*

$$\begin{bmatrix} F(x) \\ F(x_1) \\ \vdots \\ F(x_t) \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(x,x) & [K(x,x_i)]_{i\in[t]} \\ [K(x,x_i)]_{i\in[t]}^\top & [K(x_i,x_j)]_{i,j\in[t]} \end{bmatrix}\right) =: \mathcal{N}\left(0, \begin{bmatrix} K_{xx} & K_{xX} \\ K_{xX}^\top & K_{XX} \end{bmatrix}\right).$$

Then by Lemma 2, the posterior distribution is given by

$$F(x) | [f(x_i)]_{i\in[t]} \sim \mathcal{N}\left(K_{xX}K_{XX}^{-1}[f(x_i)]_{i\in[t]}, K_{xx} - K_{xX}K_{XX}^{-1}K_{xX}^\top\right). \tag{1}$$

Now we do a sanity check. Let $x = x_k$ for $k \in [t]$. Then we know $K_{xX} = \mathbf{e}_k^\top K_{XX}$ and $K_{xx} = \mathbf{e}_k^\top K_{XX}\mathbf{e}_k$. Hence

$$K_{xX}K_{XX}^{-1}[f(x_i)]_{i\in[t]} = \mathbf{e}_k^\top K_{XX}K_{XX}^{-1}[f(x_i)]_{i\in[t]} = f(x_k)$$
$$K_{xx} - K_{xX}K_{XX}^{-1}K_{xX}^\top = K_{xx} - \mathbf{e}_k^\top K_{XX}K_{XX}^{-1}K_{XX}\mathbf{e}_k = 0.$$

2.2 Bayesian optimization via Gaussian process regression

Now we illustrate the strategy to choose the next point x_{t+1}. For convenience, we consider the maximization problem.

Definition 4 (acquisition function)**.** *Let* $\{(x_i, f(x_i))\}_{i\in[t]}$ *be the observed data and* P_t *the distribution over all functions by* (1)*. Put* $f_t^* = \max\{f(x_1), \ldots, f(x_t)\}$*. The* acquisition function $\mathrm{EI}_t : \mathbb{R}^d \to \mathbb{R}$ *is defined as*

$$\mathrm{EI}_t(x) = \mathbb{E}_{g\sim P_t}\left[\max\{0, g(x) - f_t^*\}\right].$$

Remark 2. The function EI means the expected improvement (actually decrease) of the minimum when we pick $x \in \mathbb{R}^d$ as our next point. When the dimension is low, we can efficiently optimize EI_t by grid search or gradient ascent.

If $P_t(x) = \mathcal{N}(\mu(x), \sigma(x)^2)$, by simple calculation, it holds

$$\mathrm{EI}_t(x) = \sigma(x)\left(\gamma(x)\Phi(\gamma(x)) + \varphi(\gamma(x))\right)$$

where $\gamma(x) := \frac{\mu(x) - f_t^*}{\sigma(x)}$ and φ, Φ are the PDF and CDF of the standard normal distribution respectively.

Now we describe the algorithm formally. At each iteration t, assume that the observed dataset is $\{(x_i, f(x_i))\}_{i\in[t]}$.

- Firstly we compute P_t as (1) and EI_t.
- Find maximizer x_{t+1} with respect to EI_t (using grid search or gradient ascent).
- Evaluate $f(x_{t+1})$ and append the point $(x_{t+1}, f(x_{t+1}))$ to the dataset.
- Go to the next iteration.

Note that, it is not easy to give any theoretical guarantee or convergence rate for Bayesian optimization. Despite the unknown convergence result, it is still a widely used optimization algorithm.