Elementary Hyperparameter Tuning

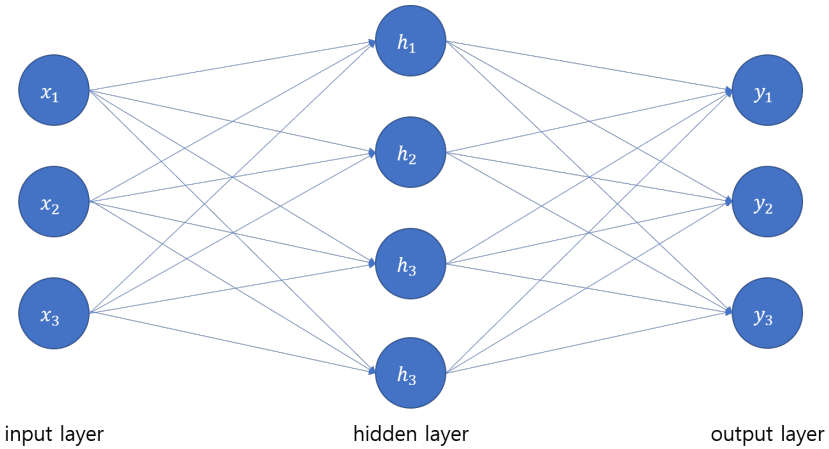
김선중

August 24, 2021

Table of Contents

Parameters

Consider a simple neural network, representing a function $f : \mathbb{R}^3 \to \mathbb{R}^3$.



With two matrices $W \in \mathbb{R}^{3\times4}$ and $W' \in \mathbb{R}^{4\times3}$, two vectors $b \in \mathbb{R}^4$ and $b' \in \mathbb{R}^3$ and the sigmoid function $\sigma(t) = (1 + e^{-t})^{-1}$, we have

$$f(x) = \sigma(W'\sigma(Wx + b) + b'), \tag{1}$$

where σ is evaluated elementwisely.

Parameters

For a dataset $D = \{(x^n, y^n) : x^n, y^n \in \mathbb{R}^3, n = 1, \cdots, N\}$, the cost can be evaluated as

$$C = \sum_{n=1}^{N} ||f(x^n) - y^n||_2{}^2. \tag{2}$$

Denote

$$\begin{aligned}\Theta &= (W, b, W', b') \\ &= (w_{11}, \cdots, w_{34}, b_1, b_2, b_3, b_4, w'_{11}, \cdots, w'_{43}, b'_1, b'_2, b'_3).\end{aligned}$$

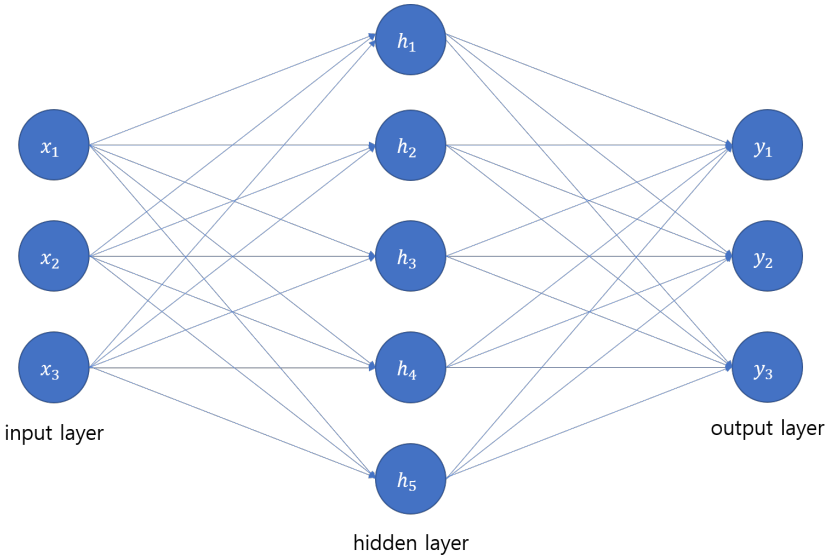
Since f depends on Θ, we may think of $C : \mathbb{R}^{31} \to \mathbb{R}$ as a function of Θ.

Parameters

- Using the gradient descent or the likes, we can find an approximated minimum of $C(\Theta)$.
- The optimal f^* is then obtained from the the minimum argument Θ^*.
- Components w_{ij}, b_k, w'_{ij}, b'_k of Θ are called the learnable parameters, or simply the parameters.

Hyperparameters

- We may change the structure of the network slightly to produce a different function f.
- Specifically, we can use different numbers of nodes in the hidden layer, say, 5, to enhance the performance.
- The number of hidden nodes is an example of hyperparameter.



Hyperparameters

- Beside the number of hidden nodes, the learning rate and the type of optimizers is also a hyperparameter.
- That is, we may think of an operator F with its arguements ϕ_1(the learning rate), ϕ_2(the number of hidden nodes) and ϕ_3 (the type of optimizer) where

$$F[\phi_1, \phi_2, \phi_3] = f,$$

Hyperparameters

- Suppose that each hyperparameters ϕ_i has its value in a set ;

$$\phi_1 \in [0.001, 0.01]$$
$$\phi_2 \in \{1, 2, 3, 4, 5, 6\}$$
$$\phi_3 \in \{\text{SGD}, \text{Adam}, \text{RMSprop}\}$$

- We may think of ϕ_i as random variables.
- The sample spaces of ϕ_i's are

$$S_1 = [0.001, 0.01]$$
$$S_2 = \{1, 2, 3, 4, 5, 6\}$$
$$S_3 = \{\text{SGD}, \text{Adam}, \text{RMSprop}\}$$

Hyperparameters

- For each $\Phi = (\phi_1, \phi_2, \phi_3)$, there corresponds a real value $C(\Theta^*)$.
- We are to find the optimal hyperparameters $\Phi = (\phi_1, \phi_2, \phi_3)$ that minimizes $C(\Theta^*)$.

The task of determining the set of hyperparameter to minimize the cost, given a dataset and the architecture of the model, is called the hyperparameter tuning, hyperparameter optimization, or hyperparameter selection.

Hyperparameters

Beside

- the learning rate
- the number of nodes in the hidden layers
- the type of optimizer,

the followings are also possible hyperparameters ;

- the size of minibatch
- the number of epochs
- the type of activation function
- the type of weight initialization
- the dropout parameter
- the regularization parameter

Manual Tuning

- As the most primitive and natural method, we may tune the hyperparameters by hands.
- We move from one point in the hyperparameter space to the other, each time evaluating the cost.
- There is no general rule for manual tuning.

$$(\phi_1, \phi_2, \phi_3) \longrightarrow (\phi_1 + \epsilon_1, \phi_2, \phi_3) \longrightarrow ({\phi_1}^*, \phi_2, \phi_3)$$
$$\longrightarrow ({\phi_1}^*, \phi_2 + \epsilon_2, \phi_3) \longrightarrow ({\phi_1}^*, {\phi_2}^*, \phi_3) \longrightarrow \cdots$$

Manual Tuning

Advantages

- We can learn the behavior of hyperparameters by heart.

Disadvantages

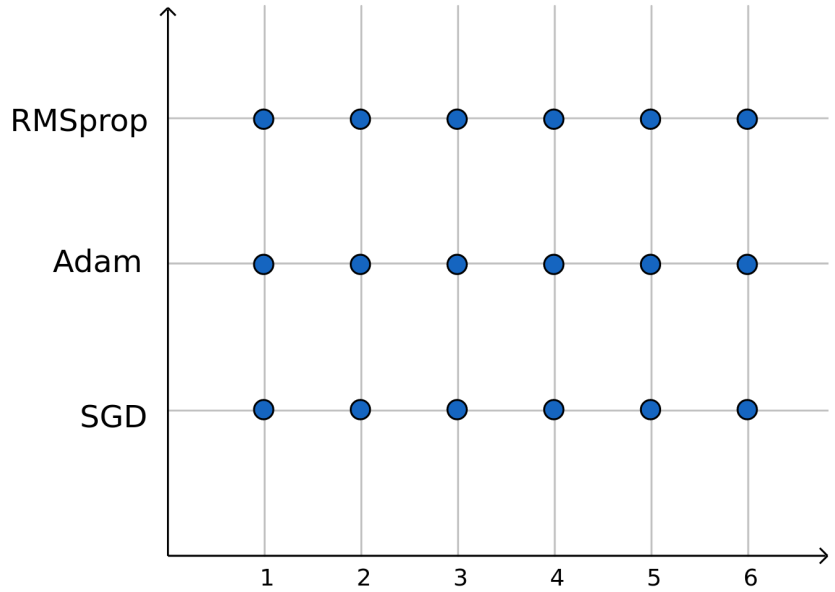
- Manual works are required.
- There is no general rule.

Grid Search

- Consider tuning ϕ_2 and ϕ_3 while ϕ_1 is fixed.

$$\phi_2 \in S_2 = \{1, 2, 3, 4, 5, 6\}$$
$$\phi_3 \in S_3 = \{\text{SGD}, \text{Adam}, \text{RMSprop}\}$$

- There are $|S_2| \times |S_3| = 6 \times 3 = 18$ possibilities that (ϕ_2, ϕ_3) can attain.



Grid Search

- Consider tuning ϕ_1, ϕ_2, ϕ_3 simultaneously.
- Since ϕ_1 lies in a (continuous) closed interval $S_1 = [0.001, 0.01]$ we need to discretize it ;

$$\phi_1 \in \{0.001, 0.005, 0.01\}.$$

- There are $3 \times 6 \times 3 = 54$ cases ;
  i.e. 54 points on the hyperparameter space $S_1 \times S_2 \times S_3$.
- For each cases, we train the model and test the performance independently.
- By comparing all the cost we get, we find the set of optimal hyperparameters and the corresponding cost value.

Grid Search

Advantages

- We can cover all possible prospective sets of hyperparameters.

Disadvantages

- It may take too long ; the running time increase exponentially as the number of types of hyperparameter increases.
- There exist holes between grids.

Grid Search : Code

1. Grid Search for Classification

In [1]:

```
%%time
# grid search logistic regression model on the sonar dataset
from pandas import read_csv
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import RepeatedStratifiedKFold, GridSearchCV
# load dataset
url = 'https://raw.githubusercontent.com/jbrownlee/Datasets/master/sonar.csv'
dataframe = read_csv(url, header=None)
# split into input and output elements
data = dataframe.values
X, y = data[:, :-1], data[:, -1]
# define model
model = LogisticRegression()
# define evaluation
cv = RepeatedStratifiedKFold(n_splits=10, n_repeats=3, random_state=1)
# define search space
space = dict()
space['solver'] = ['newton-cg', 'lbfgs', 'liblinear']
space['penalty'] = ['none', 'l1', 'l2', 'elasticnet']
space['C'] = [1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1, 10, 100]
# define search
search = GridSearchCV(model, space, scoring='accuracy', n_jobs=-1, cv=cv)
# execute search
result = search.fit(X, y)
# summarize result
print('Best Score: %s' % result.best_score_)
print('Best Hyperparameters: %s' % result.best_params_)
```

```
Best Score: 0.7828571428571429
Best Hyperparameters: {'C': 1, 'penalty': 'l2', 'solver': 'newton-cg'}
Wall time: 29.1 s
```

Random Search

- Think of the hyperparameters ϕ_i as random variables with adequate distribution.
- Sample from the joint distribution of $\Phi = (\phi_1, \phi_2, \phi_3)$, build a model and test the performance.
- We can repeat this procedure multiple times and select the best one as the optimal set of hyperparameters.

Random Search

Advantages

- Sampling can find unexpected points in the hyperparameter space with good performance.
- To get more accurate model, we simply increase the number of sampling.

Disadvantages

- If the hyperparameter space is big, then some hyperparameters might not be explored.
- We must specify the probability distribution of the sample space in advance.

Random Search : Code

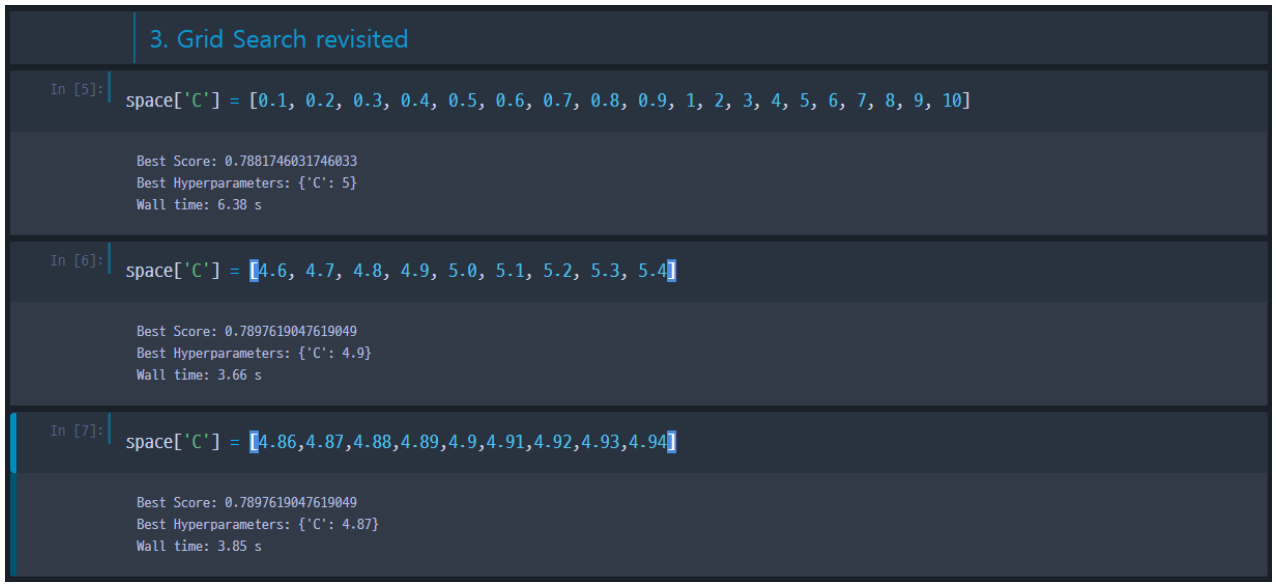
2. Random Search for Classification

In [2]:

```
%%time
# random search logistic regression model on the sonar dataset
from scipy.stats import loguniform
from pandas import read_csv
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import RepeatedStratifiedKFold, RandomizedSearchCV
# load dataset
url = 'https://raw.githubusercontent.com/jbrownlee/Datasets/master/sonar.csv'
dataframe = read_csv(url, header=None)
# split into input and output elements
data = dataframe.values
X, y = data[:, :-1], data[:, -1]
# define model
model = LogisticRegression()
# define evaluation
cv = RepeatedStratifiedKFold(n_splits=10, n_repeats=3, random_state=1)
# define search space
space = dict()
space['solver'] = ['newton-cg', 'lbfgs', 'liblinear']
space['penalty'] = ['none', 'l1', 'l2', 'elasticnet']
space['C'] = loguniform(1e-5, 100)
# define search
search = RandomizedSearchCV(model, space, n_iter=500, scoring='accuracy', n_jobs=-1, cv=cv, random_state=1)
# execute search
result = search.fit(X, y)
# summarize result
print('Best Score: %s' % result.best_score_)
print('Best Hyperparameters: %s' % result.best_params_)
```

```
Best Score: 0.7897619047619049
Best Hyperparameters: {'C': 4.878363034905756, 'penalty': 'l2', 'solver': 'newton-cg'}
Wall time: 1min 21s
```

Grid Search vs Random Search



- Grid method takes less time (29.1s, 6.38s, 3.66s, 3.85s) than random search(1min 21s).
- But random search requires just one step to implement and we don't have to set the grid carefully like we did in grid method.
- Morever, it produces quite an accurate result, compared to the random search.

Bayesian Optimization

- A drawback of random search : we must specify the probability distribution of the sample space in advance.
- In bayesian optimization, the probability distribution evolves from one to another.
- Specifically, given a prior distribution $P(f)$, we get the posterior distribution $(f|D)$ iteratively.
- Bayes Rule :

$$P(f|D) \propto P(D|f)P(f).$$

- More detailed explanation and the code will be covered in the final presentaion.

An Overview of Bayesian Optimizaiton

- Bayesian optimization is a method for maximizing (or minimizing) a function $f : A \to \mathbb{R}$ where $A \subset \mathbb{R}^d$. Sometimes we write

$$\max_{x \in A} f(x).$$

- The objective function f is unknown, we are to sample from the domain A using surrogate function and acquisition function to achieve our goal.

An Overview of Bayesian Optimizaiton

In our problem, we are assuming the followings;

- d, the dimension of the domain, is assumed to be less than 20.
- x, the independent variable, is assumed to be within a simple domain A like k-cube $\prod_{i=1}^{k}[a_i, b_i]$.
- f is assumed to be a good function in a sense that it is (Lipschitz) continuous. But it may not be differentiable. So it is impossible to approximate f using its derivatives or its second derivatives.
- f can be a really complicated function (it may lack special structure like concavity or linearity.), so one can not evaluate the value in low cost.

Overview of Bayesian Optimization

- It is called *Bayesian* because it uses the famous "Bayes theorem"
$$P(M|E) \propto P(E|M)P(M).$$
- The posterior probability of a model M given evidence E is proportional to the likelihood of E given M, multiplied by the prior probability of M.
- Let x_i be the ith sample. Consider the set
$$D_{1:t} = \{(x_i, f(x_i)) \,|1 \leq i \leq t\}$$
of ordered pair $(x_i, f(x_i)$, which plays a role of evidence. We have
$$P(f|D_{1:t}) \propto P(D_{1:t}|f)P(f).$$

Overview of Bayesian Optimization

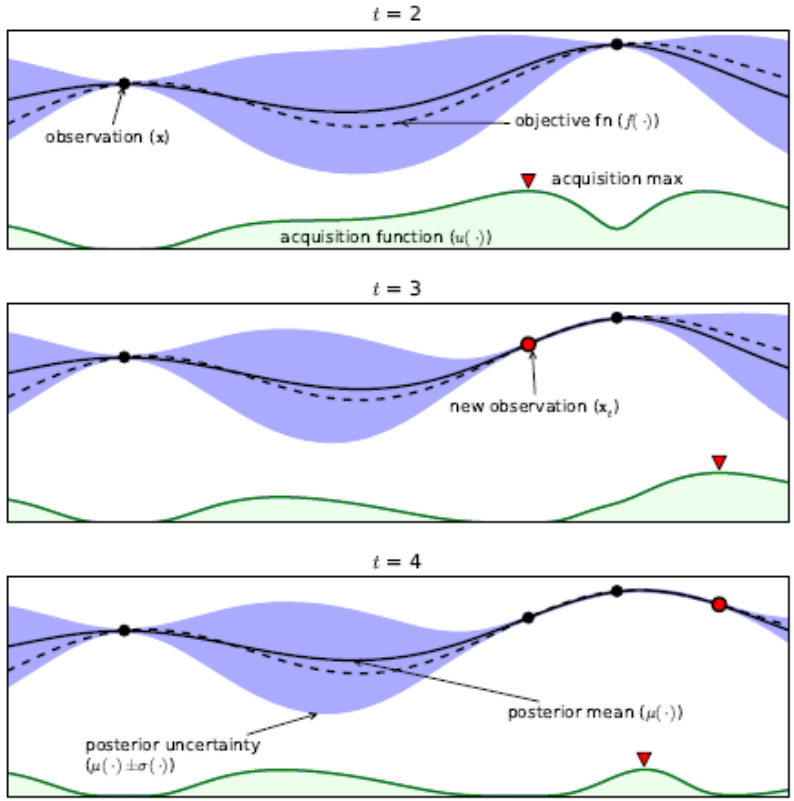


Figure 1: *An example of using Bayesian optimization on a toy 1D design problem. The figures show a Gaussian process (GP) approximation of the objective function over four iterations of sampled values of the objective function. The figure also shows the acquisition function in the lower shaded plots. The acquisition is high where the GP predicts a high objective (exploitation) and where the prediction uncertainty is high (exploration)—areas with both attributes are sampled first. Note that the area on the far left remains unsampled, as while it has high uncertainty, it is (correctly) predicted to offer little improvement over the highest observation.*

Priors

- We are to use Gaussian(normal) distribution for the prior, where we introduce the notion of *Gaussian process*.
- The collection $\{X_t : t \geq 0\}$ of random variables X_t where t is indexed in a continuous domain is called the random process.
- Gaussian process is a kind of random process, where each of the X_t for $t \geq 0$ is distributed normally.

Priors

- We are to approxiamte the function f.
- For each $x \in A$, the value $f(x)$ is not determined. So we may think of $f(x)$ as a random variable. That is, we may think of $\{f(x)|x \in A\}$ as a collection of random variable in a continuous domain A.
- Assuming $f(x)$ is distributed normally, we need only to specify its mean $m(x)$ for each $x \in A$ and covariance $k(x, x')$ for each $x, x' \in A$. We write

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')).$$

Priors

- Still, the function m and k are not specified yet.
- For convenience, let m be a zero function;

$$m(x) = 0. \quad (x \in A)$$

- We may choose k quite arbitrarily. But the most popular choice is to use squared exponential function;

$$k(x_i, x_j) = \exp\left(-\frac{1}{2}||x_i - x_j||^2\right).$$

- Note that the value $k(x_i, x_j)$ converges to 1 as $(x_i - x_j)$ approaches to 0.

Covariance Function

- The (popular) kernel function we have defined is quite a naive one. We add a hyperparameter θ so that

$$k(x_i, x_j) = \exp\left(-\frac{1}{2\theta}||x_i - x_j||^2\right).$$

- For anisotropic models, we may set

$$k(x_i, x_j) = \exp\left(-\frac{1}{2}(x_i - x_j)^T \text{diag}(\theta)^{-2}(x_i - x_j)\right),$$

where diag stands for a diagonal matrix of a vector θ.

Covariance Function

- Another important kernel for Bayesian Optimization is the Matérn kernel (Matérn, 1960 ; Stein, 1999)
- It incorporates a smoothness parameter ζ to permit great flexibility in modelling functions:

$$k(x_i, x_j) = \frac{1}{2^{\zeta-1}\Gamma(\zeta)} \left(2\sqrt{\zeta}||x_i - x_j||\right)^{\zeta} H_{\zeta}\left(2\sqrt{\zeta}||x_i - x_j||\right),$$

where $\Gamma(\cdot)$ is the gamma function and $H_{\zeta}(\cdot)$ is the Bessel function of order ζ.

Acquisition Function

- The acquisition function plays a role of guiding the search for the optimum.
- It is defined so that high acquisition is related to *potentially* high values of the objective function f.
- Denote u by the acquisition function.
- At timestep t, we need to sample another point x_{t+1} from the domain. We choose x_{t+1} as

$$x_{t+1} = \text{argmax}_{x \in A} u(x|D).$$

- There are several kinds of acquisition functions, which we'll illustrate without deep explanation.

Acquisition Function : PI

- (Kushner, 1964) suggested maximizing *probability of improvement* over the incumbent $f(x^+)$ where

$$x^+ = \text{argmax}_{1 \le i \le t} f(x_i),$$

so that

$$\begin{aligned} \text{PI}(x) &= P(f(x) \ge f(x^+)) \\ &= \Phi\left(\frac{\mu(x) - f(x^+)}{\sigma(x)}\right). \end{aligned}$$

Here, Φ stands for the normal cumulative distribution function.

- Kushner's approach is a natural, but greedy one. It pursuits exploitation without exploration.

Acquisition Function : EI

- More satisfing alternative acquisition function is the one which considers not only the probability of improvement, but also the magnitude of the improvement a point can potentially yield.
- (Močkus, 1978) Let I be a function (called *improvement function*) defined by

$$I(x) = \max\{0, f_{t+1}(x) - f(x^+)\}.$$

Define the new point x at which we sample, by

$$x = \text{argmax}_x \mathbb{E}\left(\max\{0, f_{t+1}(x) - f(x^+)\}|D_t\right).$$

Acquisition Function : Generalized EI

- EI, explained so far, can be evaluated analytically(Močkus, 1978 ; Jones, 1998) as follows;

$$\begin{aligned} \mathrm{EI}(\mathbf{x}) &= \begin{cases} (\mu(\mathbf{x}) - f(\mathbf{x}^+))\Phi(Z) + \sigma(\mathbf{x})\phi(Z) & \text{if } \sigma(\mathbf{x}) > 0 \\ 0 & \text{if } \sigma(\mathbf{x}) = 0 \end{cases} \\ Z &= \frac{\mu(\mathbf{x}) - f(\mathbf{x}^+)}{\sigma(\mathbf{x})} \end{aligned}$$

- (Lizotte, 2008) suggests new acquisition function that enables us to trade off between exploration and exploitation He adopted a parameter $\xi \geq 0$ such that

$$\mathrm{EI}(\mathbf{x}) = \begin{cases} (\mu(\mathbf{x}) - f(\mathbf{x}^+) - \xi)\Phi(Z) + \sigma(\mathbf{x})\phi(Z) & \text{if } \sigma(\mathbf{x}) > 0 \\ 0 & \text{if } \sigma(\mathbf{x}) = 0 \end{cases},$$

where

$$Z = \begin{cases} \frac{\mu(\mathbf{x}) - f(\mathbf{x}^+) - \xi}{\sigma(\mathbf{x})} & \text{if } \sigma(\mathbf{x}) > 0 \\ 0 & \text{if } \sigma(\mathbf{x}) = 0 \end{cases}.$$

References

1. James Bergestra, et al., 2011, "Algorithms for Hyper-Parameter Optimzation"
2. Jasper Snoek, et al., 2012, "Practical Bayesian Optimization of Machine Learning Algorithms"

References

3. Peter I. Frazier, 2018, "A Tutorial on Bayesian Optimization"
4. Peter I. Frazier, "TutORial : Bayesian Optimization", www.youtube.com/watch?v=c4KKvyWW_Xk&t=1924s
5. Eric Brochu, et al., 2010, TR-2009 "A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning"