Bayesian Optimization: A Brief Review

Overview of Bayesian Optimization (BO)

Goal: optimize $f(\mathbf{x})$ over $\mathbf{x}$

- $f(\mathbf{x})$ is an expensive to evaluate function.
- $f(\mathbf{x})$ is a “black-box”.
- The first-order and/or second-order derivatives of $f(\mathbf{x})$ is not available.

References:

- P. Frazier, "A Tutorial on Bayesian Optimization" https://arxiv.org/abs/1807.02811
- Shahriari, Bobak, et al. "Taking the human out of the loop: A review of Bayesian optimization." Proceedings of the IEEE 104.1 (2015): 148-175.

Generic BO Algorithm

Elicit a prior distribution on the function f

while (budget is not exhausted) {

  Find $\mathbf{x}$ that maximizes Acquisition($\mathbf{x}$, prior)

  Evaluate $f(\mathbf{x})$ at $\mathbf{x}$

  Find the posterior distribution, and update the prior distribution.

}

Basic Concepts

How to update knowledge, as data is obtained?

- Prior distribution: what you know about parameter β, excluding the information in the data – denoted by $\pi(\beta)$.
- Likelihood: based on modeling assumptions, how [relatively] likely the data Y are if the truth is β – denoted $L(Y|\beta)$

So how to get a posterior distribution: stating what we know about β, combining the prior with the data denoted $p(\beta|Y)$.
Bayes Theorem used for inference tells us to multiply:

$$p(\beta|Y) \propto L(Y|\beta)\pi(\beta)$$

Essentially, Posterior $\propto$ Likelihood $\times$ Prior.

Generic Bayesian Update Algorithm

Given a prior distribution $\pi^{(0)}(\beta)$ for the target parameter β, and a model assumption $L(Y|\beta)$

For $t = 1, \ldots, N$ {

obtain data $Y^{(t)}$

find the posterior $p(\beta|Y^{(t)}) \propto L(Y^{(t)}|\beta)\pi^{(t-1)}(\beta)$

update $\pi^{(t)}(\beta) \leftarrow p(\beta|Y^{(t)})$

}

A Simple Example: Normal Prior with Known Variance

- Goal: learning parameter μ
- Prior: $\mu \sim N(\theta^{(0)}, \sigma^{(0),2})$
- Data: $Y|\mu \sim N(\mu, \lambda^2)$ where λ is known.
- Posterior: $p(\mu|Y) \propto L(Y|\mu)\pi(\mu)$ is also a normal distribution.

Acquisition functions

- Improvement-based policies: expected improvement, knowledge gradient,...
- Information-based policies: Thompson sampling
- ...

Getting into some details... with a simple example

- A collection of finite alternatives $\mathcal{X} = \{1, \ldots, M\}$

Getting into some details... with a simple example

- A collection of finite alternatives $\mathcal{X} = \{1, \ldots, M\}$
- Problem of Interests:

$$\max_{x \in \mathcal{X}} \mu_x$$

where μ_x is the unknown true performance of alternative x.

Getting into some details... with a simple example

- A collection of finite alternatives $\mathcal{X} = \{1, \ldots, M\}$
- Problem of Interests:

$$\max_{x \in \mathcal{X}} \mu_x$$

  where μ_x is the unknown true performance of alternative x.
- The true performance μ_x can not be directly measured, but can be estimated through observation:

$$y_x = \mu_x + \varepsilon_x,$$

  where $\varepsilon_x \sim N(0, \sigma^2)$.

Reference: A Knowledge-Gradient Policy for Sequential Information Collection P.I. Frazier, W.B. Powell & S. Dayanik. SIAM Journal on Control and Optimization, 2008.

A simple example

- Generating an output y_x is expensive.

A simple example

- Generating an output y_x is expensive.
- So we have a budget: a total number of N observations.

A simple example

- Generating an output y_x is expensive.
- So we have a budget: a total number of N observations.
- Research Question: how to split the N among M alternatives?

A simple example

- Generating an output y_x is expensive.
- So we have a budget: a total number of N observations.
- Research Question: how to split the N among M alternatives?
- Keep in mind:

$$\max_{x \in \mathcal{X}} \mu_x$$

A simple example: statistical modeling

- Setup the prior belief about μ_x

$$\mu_x \sim N\left(\theta_x^{(0)}, (\sigma_x^{(0)})^2\right)$$

independent with each other over $\mathcal{X}$.

A simple example: statistical modeling

- Setup the prior belief about μ_x

$$\mu_x \sim N\left(\theta_x^{(0)}, (\sigma_x^{(0)})^2\right)$$

  independent with each other over $\mathcal{X}$.
- Assume that we collect the outputs $y_{x^1}, \ldots, y_{x^N}$ are collected one by one.

A simple example: statistical modeling

- Setup the prior belief about μ_x

$$\mu_x \sim N\left(\theta_x^{(0)}, (\sigma_x^{(0)})^2\right)$$

independent with each other over $\mathcal{X}$.

- Assume that we collect the outputs $y_{x^1}, \ldots, y_{x^N}$ are collected one by one.
- When the new observation $y_{x^{(t)}}$ arrives, we find the posterior distribution of μ_x given $y_{x^{(t)}}$

$$\mu_x | y_{x^{(t)}} \sim N\left(\theta_x^{(t)}, (\sigma_x^{(t)})^2\right),$$

A simple example: statistical modeling

- Setup the prior belief about μ_x

$$\mu_x \sim N\left(\theta_x^{(0)}, (\sigma_x^{(0)})^2\right)$$

  independent with each other over $\mathcal{X}$.
- Assume that we collect the outputs $y_{x^1}, \ldots, y_{x^N}$ are collected one by one.
- When the new observation $y_{x^{(t)}}$ arrives, we find the posterior distribution of μ_x given $y_{x^{(t)}}$

$$\mu_x | y_{x^{(t)}} \sim N\left(\theta_x^{(t)}, (\sigma_x^{(t)})^2\right),$$

- In the end, find $\max_{x \in \mathcal{X}} \theta_x^{(N)}$

Knowledge gradient under a simple example

- Idea: choose x which provides the maximum expected "improvement" to the target problem:

$$\max_{x \in \mathcal{X}} \mu_x,$$

where $\mathcal{X}$ contains K elements.

Knowledge gradient under a simple example

- Idea: choose x which provides the maximum expected "improvement" to the target problem:

$$\max_{x \in \mathcal{X}} \mu_x,$$

where $\mathcal{X}$ contains K elements.

- The Knowledge gradient:

$$\mathrm{KG}^{(t)}(x) = \mathrm{E}[\max_{x' \in \mathcal{X}} \theta_{x'}^{(t+1)} - \max_{x' \in \mathcal{X}} \theta_{x'}^{(t)} | x^{(t+1)} = x],$$

where the expectation is taken with respect to the posterior predictive distribution of $Y_x^{(t+1)}$.

Knowledge gradient under a simple example

- Idea: choose x which provides the maximum expected "improvement" to the target problem:

$$\max_{x \in \mathcal{X}} \mu_x,$$

  where $\mathcal{X}$ contains K elements.
- The Knowledge gradient:

$$\mathrm{KG}^{(t)}(x) = \mathrm{E}[\max_{x' \in \mathcal{X}} \theta_{x'}^{(t+1)} - \max_{x' \in \mathcal{X}} \theta_{x'}^{(t)} | x^{(t+1)} = x],$$

  where the expectation is taken with respect to the posterior predictive distribution of $Y_x^{(t+1)}$.
- Maximize $\mathrm{KG}^{(t)}(x)$ over $\mathcal{X}$ to select the alternative for new experiment.

Knowledge gradient under a simple example

- K alternatives
- For $k = 1, \ldots, K$

$$\mu_k \sim N(\theta_k^{(0)}, \sigma_k^{(0),2})$$

$$Y_k | \mu_k \sim N(\mu_k, \lambda_k^2),$$

  where λ_k^2 is known.
- Independence between alternatives.
- Model update (if sample from the k-th alternative at step $t+1$):

$$\theta_k^{(t+1)} = \theta_k^{(t)} + \frac{\sigma_k^{(t),2}}{\lambda_k^2 + \sigma_k^{(t),2}} (Y_k^{(t+1)} - \theta_k^{(t)})$$

$$\sigma_k^{(t+1),2} = \frac{\lambda_k^2 \sigma_k^{(t),2}}{\lambda_k^2 + \sigma_k^{(t),2}}$$

Knowledge gradient under a simple example

Under the normal model with known variance, we have that

$$\mathrm{KG}^{(t)}(x) = \mathrm{E}[\max_{x' \in \mathcal{X}} \theta_{x'}^{(t+1)} - \max_{x' \in \mathcal{X}} \theta_{x'}^{(t)} | x^{(t+1)} = x]$$

$$= \tilde{\sigma}_k^{(t)} g(\xi_k^{(t)}),$$

where

- $\tilde{\sigma}_k^{(t)} = \frac{\sigma_k^{(t),2}}{\sqrt{\lambda_k^2 + \sigma_k^{(t),2}}}$
- $\xi_k^{(t)} = -\left| \frac{\max_{j \neq k} \theta_j^{(t)} - \theta_k^{(t)}}{\tilde{\sigma}_k^{(t)}} \right|$
- $g(u) = u\Phi(u) + \phi(u)$.

Expected Improvement

The expected improvement acquisition function is given by

$$\mathrm{EI}^{(t)}(x) = \mathrm{E}\left[\max\{\mu_x - \max_j \theta_j^{(t)}, 0\}\right]$$

Under the normal model,

$$\mu_x \sim N(\theta_x^{(t)}, \sigma_x^{(t),2})$$

$$Y_x | \mu_x \sim N(\mu_x, \lambda_x^2),$$

for $k = 1, \ldots, K$. We have that,

$$\mathrm{EI}^{(t)}(x) = \sigma_x^{(t)} g\left(-\frac{|\theta_x^{(t)} - \max_j \theta_j^{(t)}|}{\sigma_x^{(t)}}\right)$$

Efficient Global Optimization, Jones et al, 1998

- Model: Gaussian process
- Acquisition function: Expected improvement

Gaussian Process (GP)

- Assume

$$y(\mathbf{x}) = \mathbf{f}(\mathbf{x}_i)\beta + \epsilon(\mathbf{x}_i), \tag{1}$$

where $\mathbf{f}(\mathbf{x}_i) = \mathbf{f}_i$ is a pre-specified $1 \times p$ regressor, β is the vector of unknown regression parameters, $\epsilon(\mathbf{x}_i)$ is a stationary Gaussian process with mean zero and covariance

$$\text{cov}\left[\epsilon(\mathbf{x}_i), \epsilon(\mathbf{x}_j)\right] = \sigma^2 R(\mathbf{x}_i, \mathbf{x}_j), \text{ for } i \neq j, \tag{2}$$

and R is a correlation function.

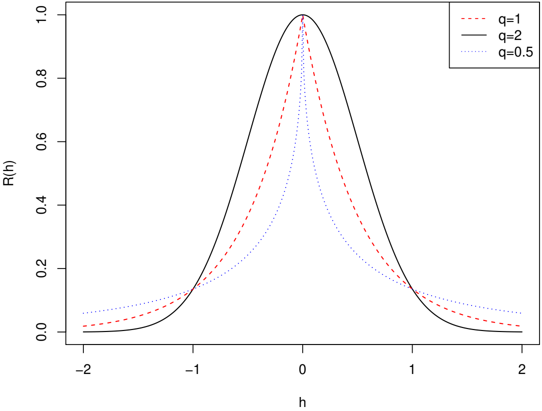
Correlation Functions

- The choice of R determines the smoothness of $\hat{y}(\mathbf{x})$.
- One popular example:

$$R(\mathbf{x}_i, \mathbf{x}_j) = R(|\mathbf{x}_i - \mathbf{x}_j|) = \exp\left(-\sum_{k=1}^{p} \theta_k |x_{ik} - x_{jk}|^{q_k}\right), \qquad (3)$$

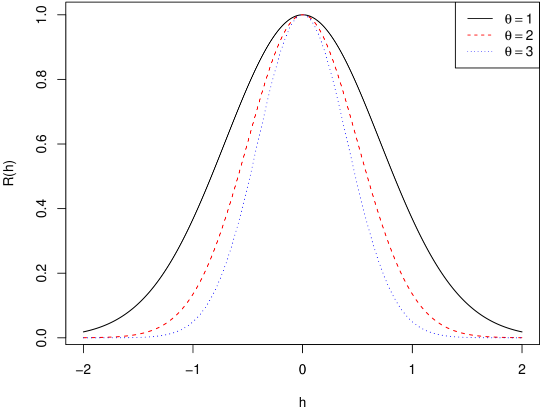
  where the subscript k denotes the kth dimension.
- Consider $R(\mathbf{h})$ for $\mathbf{h} \in \mathbb{R}^p$.

Comparison of Correlation Functions



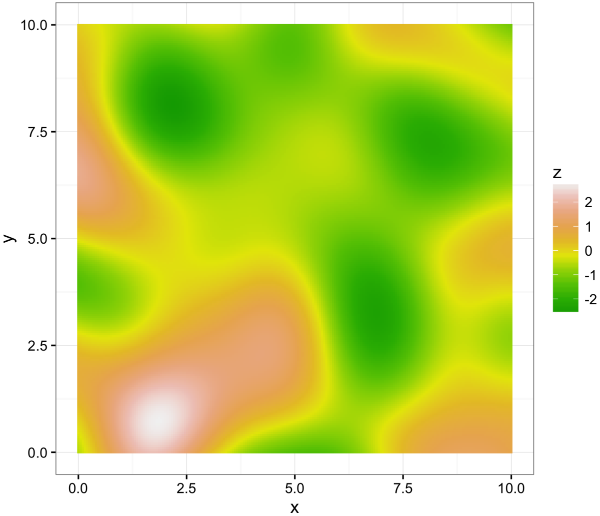
Comparison of different exponential power correlation functions with $\theta = 2$

Comparison of Correlation Functions



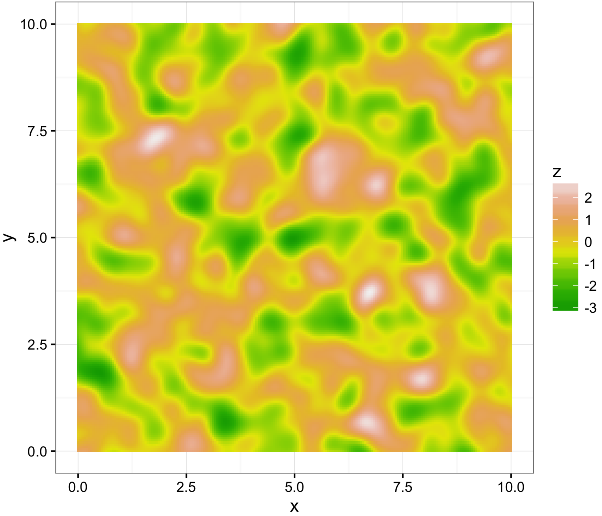
Comparison of different exponential power correlation functions with $q = 2$

Comparison of Correlation Functions



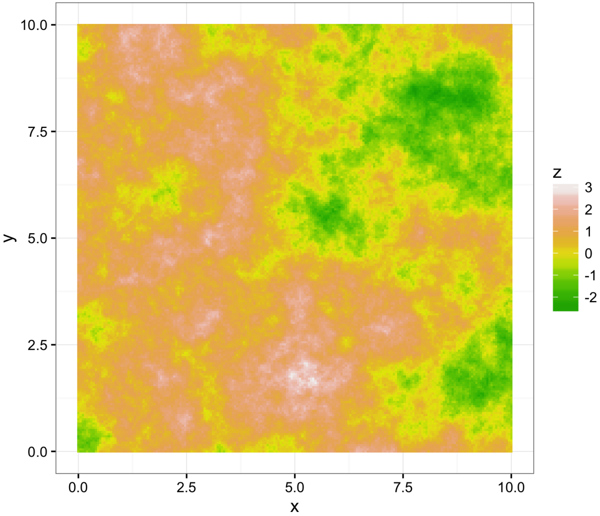
A Gaussian process with $q = 2$ and $\theta_k = .5$

Comparison of Correlation Functions



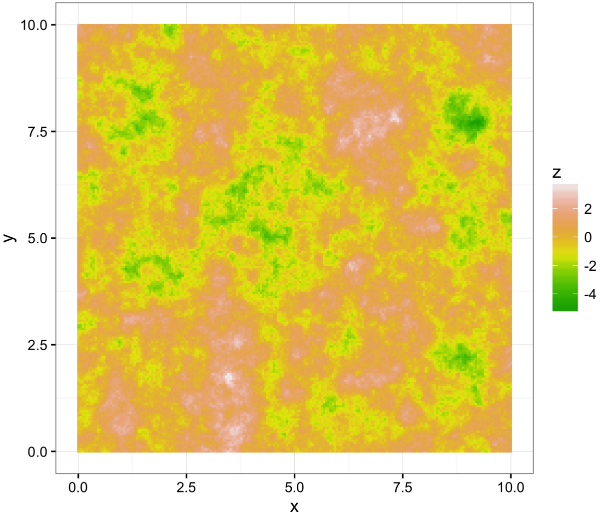
A Gaussian process with $q = 2$ and $\theta_k = 2$

Comparison of Correlation Functions



A Gaussian process with $q = 1$ and θ_k= 0.5

Comparison of Correlation Functions



A Gaussian process with $q = 1$ and $\theta_k = 2$

Estimation of GP Parameters

- The unknown parameters involved in (1) are σ^2, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ and β.
- Given $\boldsymbol{\theta}$, the estimated σ^2 and β are

$$\hat{\beta} = \left(\mathbf{F}^\top \mathbf{R}^{-1} \mathbf{F}\right)^{-1} \mathbf{F}^\top \mathbf{R}^{-1} \mathbf{Y}, \tag{4}$$

$$\hat{\sigma}^2 = \frac{(\mathbf{Y} - \mathbf{F}\hat{\beta})^\top \mathbf{R}^{-1} (\mathbf{Y} - \mathbf{F}\hat{\beta})}{n}, \tag{5}$$

  where $\mathbf{Y} = (Y_1, \ldots, Y_n)$, $\mathbf{R}$ is the $n \times n$ matrix with entries $R(\mathbf{x}_i, \mathbf{x}_j)$ defined in (3) for $i, j = 1, \ldots, n$ and $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_n]$.
- Given $\hat{\beta}$ and $\hat{\sigma}^2$, the correlation parameters $\boldsymbol{\theta}$ can be estimated by maximizing the log likelihood function

$$-\frac{n}{2} \log \hat{\sigma}^2 - \frac{1}{2} \log |\mathbf{R}|. \tag{6}$$

Gaussian process (GP)

- We treat the deterministic response $y(\mathbf{x})$ as a realization of a Gaussian stochastic process

$$Y(\mathbf{x}) = \mu + Z(\mathbf{x}).$$

Gaussian process (GP)

- We treat the deterministic response $y(\mathbf{x})$ as a realization of a Gaussian stochastic process

$$Y(\mathbf{x}) = \mu + Z(\mathbf{x}).$$

- μ is the constant mean.

Gaussian process (GP)

- We treat the deterministic response $y(\mathbf{x})$ as a realization of a Gaussian stochastic process

$$Y(\mathbf{x}) = \mu + Z(\mathbf{x}).$$

- μ is the constant mean.
- $Z(\mathbf{x})$ is a zero-mean, stationary, Gaussian stochastic process with variance σ^2 and correlation function $r(\mathbf{x}, \mathbf{x}')$.

Gaussian process (GP)

- We treat the deterministic response $y(\mathbf{x})$ as a realization of a Gaussian stochastic process

$$Y(\mathbf{x}) = \mu + Z(\mathbf{x}).$$

- μ is the constant mean.
- $Z(\mathbf{x})$ is a zero-mean, stationary, Gaussian stochastic process with variance σ^2 and correlation function $r(\mathbf{x}, \mathbf{x}')$.
- A popular choice:

$$r(\mathbf{x}, \mathbf{x}') = \exp\left\{ -\sum_{k=1}^{p} \theta_k |x_k - x'_k|^{p_k} \right\}$$

Gaussian process (GP)

- We treat the deterministic response $y(\mathbf{x})$ as a realization of a Gaussian stochastic process

$$Y(\mathbf{x}) = \mu + Z(\mathbf{x}).$$

- μ is the constant mean.
- $Z(\mathbf{x})$ is a zero-mean, stationary, Gaussian stochastic process with variance σ^2 and correlation function $r(\mathbf{x}, \mathbf{x}')$.
- A popular choice:

$$r(\mathbf{x}, \mathbf{x}') = \exp\left\{-\sum_{k=1}^{p} \theta_k |x_k - x_k'|^{p_k}\right\}$$

- This model is also called Kriging, or more specific ordinary Kriging.

Gaussian process (GP)

The BLUP predictor can be expressed by

$$\hat{y}(\mathbf{x}) = \hat{\mu} + \mathbf{r}' R^{-1}(\mathbf{y} - \mathbf{1}\hat{\mu}),$$

where

- $\hat{\mu} = (\mathbf{1}^\top R^{-1}\mathbf{1})^{-1}(\mathbf{1}^\top R^{-1}\mathbf{y})$
- $\mathbf{r} = (r(\mathbf{x}, \mathbf{x}_1), \ldots, r(\mathbf{x}, \mathbf{x}_n))^\top$
- R is an $n \times n$ matrix with entries $r(\mathbf{x}_i, \mathbf{x}_j)$.

The BLUP predictor can be expressed by

$$\hat{y}(\mathbf{x}) = \hat{\mu} + \mathbf{r}' R^{-1}(\mathbf{y} - \mathbf{1}\hat{\mu}),$$

where

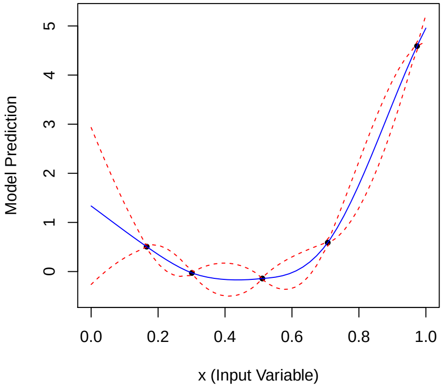
- $\hat{\mu} = (\mathbf{1}^\top R^{-1}\mathbf{1})^{-1}(\mathbf{1}^\top R^{-1}\mathbf{y})$
- $\mathbf{r} = (r(\mathbf{x}, \mathbf{x}_1), \ldots, r(\mathbf{x}, \mathbf{x}_n))^\top$
- R is an $n \times n$ matrix with entries $r(\mathbf{x}_i, \mathbf{x}_j)$.

By substituting BLUP into $\mathrm{MSE}(\hat{y}(\mathbf{x}))$, we have that

$$\mathrm{MSE}(\hat{y}(\mathbf{x})) = \sigma^2 \left(1 - \mathbf{r}' R^{-1}\mathbf{r} + \frac{(1 - \mathbf{1}^\top R^{-1}\mathbf{r})^2}{\mathbf{1}^\top R^{-1}\mathbf{1}}\right)$$

which is the variance of $\hat{y}(\mathbf{x})$.

An Illustration of Interpolator



Expected Improvement

Goal: $\min_{x \in \mathcal{X}} f(\mathbf{x})$, where $f(\mathbf{x})$ is a deterministic blackbox function with inputs $\mathbf{x}$.
Assume that the prior of $f(\mathbf{x})$ is a GP, denoted by $Y(\mathbf{x})$.
The expected improvement can be expressed by

$$\mathrm{EI}(\mathbf{x}) = \mathrm{E}[\max(\mathrm{f_{min}} - \mathrm{Y}(\mathbf{x}), 0)]$$

$$= (f_{min} - \hat{y}(\mathbf{x}))\Phi\left(\frac{f_{min} - \hat{y}(\mathbf{x}}{s(\mathbf{x})}\right) + s(\mathbf{x})\phi\left(\frac{f_{min} - \hat{y}(\mathbf{x})}{s(\mathbf{x})}\right),$$

where $Y(\mathbf{x}) \sim N(\hat{y}(\mathbf{x}), s(\mathbf{x}))$.

Illustration

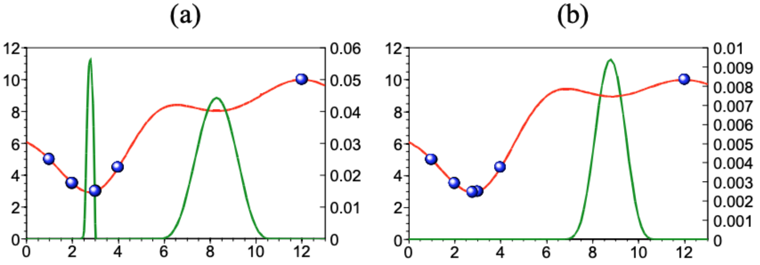


Figure 11. (a) The expected improvement function when only five points have been sampled; (b) the expected improvement function after adding a point at $x = 2.8$. In both (a) and (b) the left scale is for the objective function and the right scale is for the expected improvement.

Maximization of EI

- We have no concave or convex property of $EI(\mathbf{x})$.
- Develop a branch-and-bound algorithm to maximize $EI(\mathbf{x})$ to guaranteed optimality.

$$\frac{\partial EI(\mathbf{x})}{\partial \hat{y}(\mathbf{x})} = -\Phi\left(\frac{f_{min} - \hat{y}(\mathbf{x})}{s(\mathbf{x})}\right)$$

and

$$\frac{\partial EI(\mathbf{x})}{\partial s(\mathbf{x})} = \phi\left(\frac{f_{min} - \hat{y}(\mathbf{x})}{s(\mathbf{x})}\right)$$

- Because of this monotonicity, to find an upper bound on $EI(\mathbf{x})$ over a box for $\mathbf{x}$ is suffices to find a lower bound on $\hat{y}$ and an upper bound on s over the box.