Artificial Intelligence

Albert-Ludwigs-Universität Freiburg

Thorsten Schmidt
Abteilung für Mathematische Stochastik

www.stochastik.uni-freiburg.de
thorsten.schmidt@stochastik.uni-freiburg.de
SS 2017

Our goal today

Bayesian Optimization

Literature (incomplete, but growing):

- I. Goodfellow, Y. Bengio und A. Courville (2016). **Deep Learning**. http://www.deeplearningbook.org. MIT Press
- D. Barber (2012). **Bayesian Reasoning and Machine Learning**. Cambridge University Press
- R. S. Sutton und A. G. Barto (1998). **Reinforcement Learning : An Introduction**. MIT Press
- G. James u. a. (2014). **An Introduction to Statistical Learning: With Applications in R**. Springer Publishing Company, Incorporated. ISBN: 1461471370, 9781461471370
- T. Hastie, R. Tibshirani und J. Friedman (2009). **The Elements of Statistical Learning**. Springer Series in Statistics. Springer New York Inc. URL: https://statweb.stanford.edu/~tibs/ElemStatLearn/
- K. P. Murphy (2012). **Machine Learning: A Probabilistic Perspective**. MIT Press
- CRAN Task View: Machine Learning, available at https://cran.r-project.org/web/views/MachineLearning.html
- UCI ML Repository: http://archive.ics.uci.edu/ml/ (371 datasets)
- Warren B Powell (2011). **Approximate Dynamic Programming: Solving the curses of dimensionality**. Bd. 703. John Wiley & Sons
- A nice resourse is https://github.com/aikorea/awesome-rl

Bayesian Optimization (BO)

- Typically we are interested in a problem

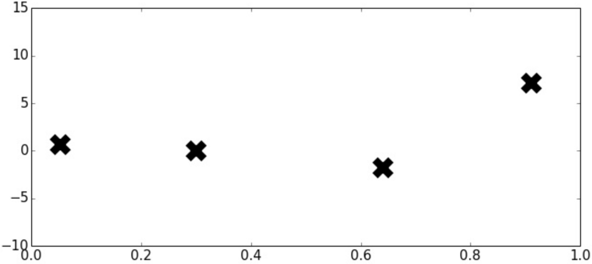
$$x^* = \arg\min_{x \in \mathscr{X}} f(x)$$

with some "well behaved" function $f : \mathscr{X} \to \mathbb{R}^d$.
- However, in many cases f is not explicitly known and it also might be multimodal.
- Also the evaluations of f might contain errors or might be very expensive.
- A nowadays famous application is (hyper-) parameter tuning in Machine Learning. Such parameters are: the number of layers / units per layers, penalties, learning rates, etc.
- A classical example is the optimal design of experiments, or the case when statistics is needed but the likelihood is intractable.

- Currently feasiable are: grid search. This will need many function evaluations, which is not good if evaluations are expensive.
- **Random search** is a well-known alternative. The usage of pseudo-random numbers even improves performance.

The problem

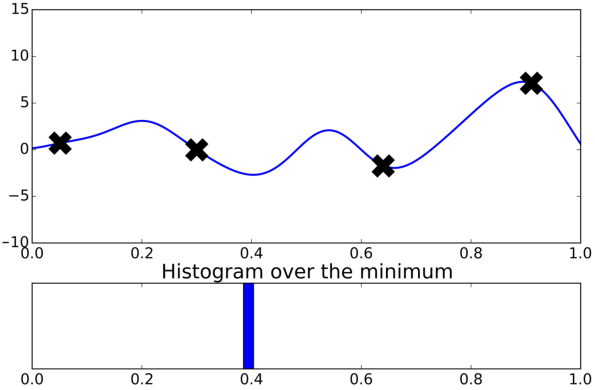
Let us illustrate the problem with a few pictures[1]



Where to choose the next point x where we evaluate $f(x)$??

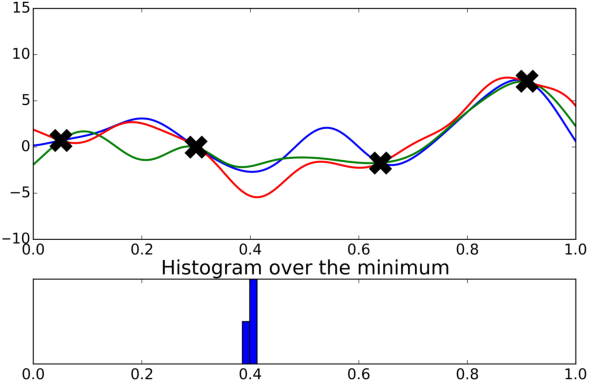
[1] Source: Javier González, Introduction to Bayesian Optimization. Masterclass, 2017 at Lancaster University.

Let us consider some possible curves. Here is one:

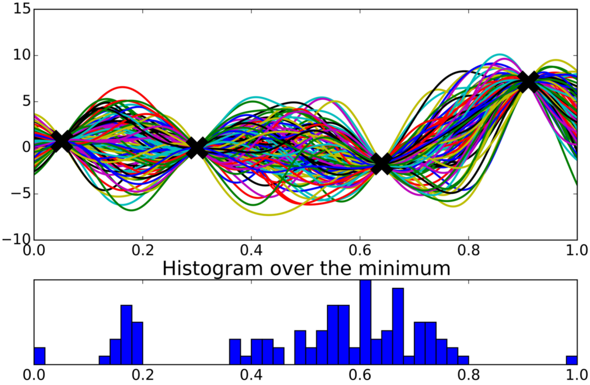


Clearly, we would choose to evaluate at the minimun and are finished. But this is not the only possible curve !

Three curves

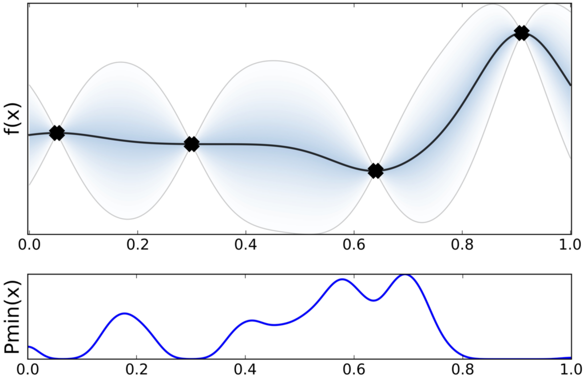


Many curves



If we think of a continuum of course, we arrive at the Bayesian representation of the problem.

We consider a density over the possible curves, which is called **prior**.



Where should we optimally place our next evaluation x^n??

- The approach is clear: we have a prior distribution p.
- Given some data $\mathscr{D}$ we update through Bayes' rule

$$p(x|\mathscr{D}) = \frac{p(\mathscr{D}|x)p(x)}{\mathbb{P}(\mathscr{D})}.$$

- Clearly, this is only possible if $\mathbb{P}(\mathscr{D}) \neq= 0$. If this is the case, we will use a conditionaly density given by

$$f(x|y) = \frac{f(x,y)}{f(y)}$$

  where $f(x,y)$ is the joint density of x and y and $f(y)$ is the marginal density.

Historical overview

- Bayesian optimization dates back at least to works by Kushner[2] in 1964 and Mockus[3] in 1978.
- Since about 10 years there is a considerable interest of these methods in the machine learning community.

[2] Harold J Kushner (1964). „A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise". In: **Journal of Basic Engineering** 86.1, S. 97–106.

[3] J Močkus (1975). „On Bayesian methods for seeking the extremum". In: **Optimization Techniques IFIP Technical Conference**. Springer, S. 400–404.

Mathematical formulation

- In most cases the prior is chosen to be Gaussian - this is the case we will also focus here. There are other variants (Student processes) and interesting research questions in this direction
- A **Gaussian process** is a family $(X(x))_{x\in\mathscr{X}}$ of random variables, where for any (finite) $x_1,\dots,x_n$ the joint distribution of

  $$X(x_1),\dots,X(x_n)$$

  is Gaussian.
- The Gaussian process can be characterized by its **mean function**

  $$m(x) := \mathbb{E}[X(x)]$$

  and its **covariance function**

  $$c(x,y) := \mathrm{Cov}(X(x),X(y)).$$

- We are able to observe (at a certain cost) $X(x)$ for a fixed sample $x_1,\dots,x_n$

- Typically we specify some kind of regression for our setup, like

$$X(x) = \beta x + \varepsilon_x$$

where the $\varepsilon(x_i)$, $i = 1, \ldots, n$ are i.i.d.

- However, if x_1 is close to x_2 we would expect close outcomes rather than independent outcomes.
- This motivatives covariance functions of the form

$$c(x,y) \propto e^{-K(x,y)}$$

with a kernel function K. Often, $K(x,y) = \| x - y \|^{\alpha}$

Example

Gaussian process regression For example suppose that our observation is unbiased, i.e. we observe $Y(x)$ such that

$$\mathbb{E}[Y(x)] = f(x).$$

A model for this is the Gaussian regression

$$Y(x) = f(x) + \varepsilon(x).$$

The **posterior** distribution is given by

$$X(z)|X(x) = f \sim \mathcal{N}(\mu, \sigma^2)$$

where $\mu = \mu(f,x,z)$ and $\sigma = \sigma(f,x,z)$ are given by

$$\mu = m(z) + K(z,x)\frac{f - m(x)}{K(x,x) + \sigma^2 I_n}$$
$$\sigma^2 = K(z,z) - K(z,x)\frac{K(x,z)}{K(x,x) + \sigma^2 I_n}.$$

At the core is the following result. Consider the case where (X,Y) is a two-dimensional normal random variable with mean (a,A) and covariance matrix

$$\begin{pmatrix} b^2 & \rho bB \\ \rho bB & B^2 \end{pmatrix}.$$

Lemma

The conditional distribution of X given Y is Gaussian and

$$\begin{aligned} \mathbb{E}[X|Y] &= a+\rho\frac{b}{B}(Y-A) \\ \mathbb{E}[(X-\mathbb{E}[X|Y])^2|Y] &= b^2(1-\rho^2). \end{aligned}$$

Beweis.

First consider standard normal X and Y with correlation ρ. The conditional density of X given $Y = y$ is

$$f(x|y) = \frac{f(x,y)}{f(y)} = \frac{2\pi\sqrt{1-\rho^2}}{\sqrt{2\pi}} \frac{\exp\left(-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right)}{\exp\left(-\frac{y^2}{2}\right)}$$

$$= \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\right).$$

Note that $\mathbb{E}[X|Y] = \rho Y$ and that $X - \rho Y$ is independent of Y as $\mathrm{Cov}(X - \rho Y, Y) = \rho - \rho = 0$. $\square$

...

For the general case observe that $Z_1 := b^{-1}(X-a)$ and $Z_2 := B^{-1}(Y-A)$ are standard normal and $\mathrm{Cov}(Z_1, Z_2) = \rho$. Hence, X conditional on Y is again normally distributed and

$$\mathbb{E}[X|Y] = \mathbb{E}[a + bZ_1|Y] = a + b\rho Z_2 = a + \frac{\rho b}{B}(Y-A).$$

and we conclude by computing the conditional variance,

$$\mathbb{E}[(X - \mathbb{E}[X|Y])^2] = \mathbb{E}[(bZ_1 - \rho b Z_2)^2|Y] = b^2\mathbb{E}[(Z_1 - \rho Z_2)^2] = b^2(1-\rho^2). \qquad \square$$

Acquisition

- The next step is to **acquire** new data through an acquisition cirterium. Recall we have the observation $X(x)$ where we are now interested in choosing x optimally.
- The predictive variance is
$$\gamma(x) = \frac{f(x^*) - \mu(x)}{\sigma(x)}.$$
- Kushner suggest to study the probability of improvement
$$\alpha_{PI}(x) = \Phi(\gamma(x)).$$
- Mockus suggest the **expected improvement** and a further alternative (Srinivas e.a. 2010) is the lower confidence bound
$$\alpha_{LCB}(x) = \mu(x) - \kappa\sigma(x).$$