Bayesian optimization for solving the inverse problem of polymer translocation

Github repo: BayesianOptimization

1. PRELIMINARIES

1.1. Polymer translocation dynamics

In this problem translocation of a polymer chain into an trans compartment through a narrow window from a *cis* compartment will be considered. The progress of the translocation process is characterized by the degree of translocation s representing the number of chain segments in the *trans* compartment. Such translocation is determined by free energy of the translocating chain

$$\mathcal{F}(N, s),$$

where N - number of monomers ans $0 \leq s \leq N$. $N = const \Rightarrow \mathcal{F}(s)$

Polymer translocation can be considered in terms of random walks along the free energy profile. This process can be described by the Fokker-Planck equation: (Palyulin et al., 2014; Rasmussen et al., 2012).

$$\frac{\partial}{\partial \tau} W(s, \tau) = \frac{\partial}{\partial s}\Big[\frac{\partial \mathcal{F}(N, s)}{\partial s} W(s, \tau) + \frac{\partial}{\partial s} W(s, \tau)\Big],$$

where $W(s, \tau)$ - the probability of translocation of s segments of polymer chain of length N with 1 initial segment compartment at time $\tau = 0$ within time τ.

The dynamics of translocation is determined by numerical solution of FP equation. The time of successful translocation is determined as the probability flux $J_{|s} = N$ at $s = N$, which represents the probability per unit time that the chain will successfully pass through the opening in time τ starting from 1 segment located in trans compartment.

$$P_T = -\Big[k_0 \frac{\partial \mathcal{F}(N, s)}{\partial s} W(s.\tau) + k_0 \frac{\partial}{\partial s} W(s.\tau)\Big]_{s=N}$$

where k_0 is the local friction coefficient (Palyulin et al., 2014).

Respectively, the time of unsuccessful translocation attempt, or the time of return to *cis* compartment, is determined as the probability flux $J_{|s} = 0$ at $s = 0$.

$$P_F = \Big[k_0 \frac{\partial \mathcal{F}(N, s)}{\partial s} W(s.\tau) + k_0 \frac{\partial}{\partial s} W(s.\tau)\Big]_{s=0}$$

Such probability distributions should be normalized by normalizing coefficient:

$$P_{T/F}^{total} = \int_0^\infty P_{T/F}(\tau) d\tau$$

The average times of successful $< \tau_T >$ and unsuccessful $< \tau_F >$ translocation attempts are given by

$$< \tau_{T/F} >= \frac{\int_0^\infty \tau P_{T/F}(\tau) d\tau}{\int_0^\infty P_{T/F}(\tau) d\tau}$$

Also rate of positive translocation may be defined as

$$rate_{T/F} = \frac{\int_0^\infty P_{T/F}(\tau) d\tau}{\int_0^\infty P_T(\tau) d\tau + \int_0^\infty P_F(\tau) d\tau}$$

Fokker-plank equation solver was written using *Fortan* programming language.

1.2. Gaussian processes

Definition 1.

A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution (Williams & Rasmussen, 2006).

$$f(x) \sim GP(\mu(x), k(x, x'))$$

where $\mu(x)$ - its mean function and $k(x, x')$ - covariance function :

$$m(x) = \mathbb{E}[f(x)]$$
$$k(x, x') = \mathbb{E}[(f(x) - \mu(x))(f(x') - \mu(x'))]$$

Usually mean function is taken as 0. In this case the random variables represent the value of the function $f(x)$ at location x. Such GP calles **stationary**.

Gaussian distributions have the following property: they are being closed under *conditioning* and *marginalization*. Being closed under conditioning and marginalization means that the resulting distributions from these operations are also Gaussian.

1.3. GP regression

1.3.1. NOISE-FREE OBSERVATIONS

Let's consider given data points $\{(x_i, y_i)|i = 1, ...n\}$.

Model:

$$y_i = f(x_i)$$

where $f \sim GP(|0, K)$

The prior is $p(f) = \mathcal{N}(f|0, K)$

Let us denote input test point as x^*, and output $y^* = f^*(x^*)$

In this case the joint distribution of training outputs f and test outputs f^*:

$$\begin{bmatrix} f \\ f^* \end{bmatrix} = \mathcal{N}\big(0, \begin{bmatrix} K(X,X) & K(X,X^*) \\ K(X^*,X) & K(X^*,X*) \end{bmatrix} \tag{1}$$

While the joint distribution gives some insight as to how f^* relates to f, at this point no inference has been performed on the predictions f^*. Actually we are interested in the posterior distribution of the predicted GP realizations f^* at the test points X^*.

To get the posterior distribution over functions we need to restrict this joint prior distribution to contain only those functions which agree with the observed data points.

$$f^*|X^*, X, f \sim \mathcal{N}\big(K(X^*,X)K(X,X)^{-1}f,$$
$$K(X^*,X^*) - K(X^*,X)K(X,X)^{-1}K(X,X^*)\big)$$

Function values f^* (corresponding to test inputs x) can be sampled from the joint posterior distribution by evaluating the mean and covariance matrix:

$$\begin{cases} \mu = K(X^*,X)K(X,X)^{-1}f \\ \sigma^2 = K(X^*,X^*) - K(X^*,X)K(X,X)^{-1}K(X,X^*) \end{cases} \tag{2}$$

1.3.2. NOISE OBSERVATIONS

Let's consider giving training data points:

$\{(x_i, y_i)|i = 1, ...n\}$.

Model:

$$y_i = f(x_i) + \epsilon_i$$

where

$$f \sim GP(|0, K)$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- is a white noise

The prior is

$$p(f) = \mathcal{N}(f|0, K)$$

Assuming additive independent identically distributed Gaussian noise ϵ the noise model, or likelihood is

$$p(y|f) = \mathcal{N}(y|f, \sigma^2 I)$$

Integrating over the function variables f we get the *marginal likelihood*

$$p(y) = \int p(y|f)p(f)df = \mathcal{N}(0, K + \sigma^2 I)$$

$$cov(y) = K(X,X) + \sigma^2 I$$

Let us denote input test point as x^*, and output $y^* = f^* + \epsilon^*$

$$\begin{bmatrix} f \\ f^* \end{bmatrix} = \mathcal{N}\big(0, \begin{bmatrix} K(X,X) + \sigma^2 I & K(X,X^*) \\ K(X^*,X) & K(X^{\cdot}X^*) \end{bmatrix}$$

Deriving the conditional distribution we have the key predictive equations for Gaussian process regression.

$$f^*|X, y, X^* \sim \mathcal{N}\big((\overline{f^*}, cov(f^*)\big)$$

$$\begin{cases} \overline{f^*} = \mathbb{E}[f^*|X, y, X^*] = K(X^*,X)[K(X,X) + \sigma^2 I]^{-1}y \\ cov(f^*) = K(X^*,X^*) - K(X^*,X)(K(X,X) + \sigma^2 I)^{-1}K(X,X^*) \end{cases} \tag{3}$$

Predicted mean - is a linear combination of observations y; this is sometimes referred to as a linear predictor. Another way to look at this equation is to see it as a linear combination of n kernel functions, each one centered on a training point, by writing

$$\overline{f^*} = \sum_{i=1}^{n} \alpha_i k(x_i, x^*)$$

where $\alpha = [K(X,X) + \sigma^2 I]^{-1}y$

1.3.3. The Model Selection Problem

As was mentioned in the marginal likelihood:

$$p(y|X) = \int p(y|f, X)p(f|X)df$$

As we have $p(y|f) = \mathcal{N}(y|f, \sigma^2 I)$ and $p(f) = \mathcal{N}(f|0, K)$ it can be rewritten

$$\log p(y|X) = -\frac{1}{2}y^T(K+\sigma^2 I)^{-1}y - \frac{1}{2}\log(K+\sigma^2 I) - \frac{n}{2}\log 2\pi$$

This equation consists of 3 main parts:

1. Data fitting

$$\frac{1}{2}y^T(K + \sigma^2 I)^{-1}y$$

2. Regularization

$$\frac{1}{2}\log(K + \sigma^2 I)$$

3. Normalizing constant

$$\frac{n}{2}\log 2\pi$$

The goal is to maximize likelihood:

$$\frac{\partial}{\partial\Theta}\log p(y|X) = \frac{1}{2}y^T K^{-1}\frac{\partial K}{\partial\Theta}K^{-1} - \frac{1}{2}tr(K^{-1}\frac{\partial K}{\partial\Theta})$$

To obtain results it is needed to reverse the matrix K, that is $O(n^3)$. Usually in practice the Cholesky factorization is used.

1.3.4. Kernels

There are two main type of covariance functions - stationary and non-stationary (Williams & Rasmussen, 2006). Stationary covariance functions are functions of $x - x_0$. Thus it is invariant stationarity to translations in the input space.

If a covariance function depends only on x and x_0 through $x \cdot x_0$ dot product covariance it is called a dot product covariance function.

There are several main types of covariance functions used in this project.

1. Squared exponential (SE)

$$k_{SE}(r) = exp^{-\frac{r^2}{2l^2}}$$

with parameter ‘ defining the characteristic length-scale

2. The Matern class

$$k_{Matern}(r) = \frac{2^{1-\nu}}{U(\nu)}\big(\frac{\sqrt{2\nu}r}{l}\big)^{\nu}K_{\nu}\big(\frac{\sqrt{2\nu}r}{l}\big)$$

with positive parameters ν and l, where K_ν is a modified Bessel function. In this projects 2 main kernels from Matern family are used:

$$k_{3/2}(r) = \big(1 + \frac{\sqrt{3}r}{l}\big)exp^{-\frac{\sqrt{3}r}{l}}$$

$$k_{5/2}(r) = \big(1 + \frac{\sqrt{5}r}{l} + \frac{5r^2}{3l^2}\big)exp^{-\frac{\sqrt{5}r}{l}}$$

3. Rational Quadratic Covariance Function

$$k_{QR} = (1 + \frac{r^2}{2\alpha l^2})^{-\alpha}$$

where $l, \alpha > 0$

4. Exponential quadratic covariance function (EQ)

$$k_{EQ}(r) = \sigma^2\big(1 + \sqrt{5}r + \frac{5}{3}r^2\big)exp^{-\sqrt{5}r}$$

1.4. Bayesian optimization

In this problem a sequential decision approach to global optimization black-box objective smooth functions $f(\cdot) : \mathcal{X} \to \mathbb{R}$ over an index set $\mathcal{X} \in \mathbb{R}^d$ is considered (Wang & de Freitas, 2014; Wilson et al., 2018).

Black-box function is considered to be very costly. So usual gradient methods for optimization are not really good in such problem. That is why Bayesian optimization is applied.

Let’s introduce some definitions:

- **Surrogate Function** (Forrester et al., 2008):

It is the statistical/probabilistic modelling of the “black-box” function. For experimenting with different parameters, this model is used to simulate function output instead of calling the actual costly function. A Gaussian Process Regression (it is a multivariate Gaussian Stochastic process) is used as “surrogate” in Bayesian Optimization.

- **Acquisition Function**:

Technique by which the posterior is used to select the next sample from the search space.

Once additional samples and their evaluation via the objective function $f(x)$ have been collected, they are added to data $\{x\}_i$ and the posterior is then updated.

- **Exploration vs Exploitation**

Acquisition function uses "Exploration vs Exploitation" strategy to decide optimal parameter search in an iterative manner. Inside these iterations, surrogate model helps to get simulated output of the function.

The **main algorithm for Bayesian optimization** (Snoek et al., 2012) may be desribed the following way. At t-th decision round by maximizing acquisition function we obtain an input $x_t \in \mathcal{X}$ and value of a black-box $f(x)$. The returned value may be deterministic $y_t = f(x_t)$ or stochastic $y_t = f(x_t) + \epsilon_t$. The goal is to get max$f(x)$.

This sequential optimisation approach is natural when the function does not have an obvious mathematical representation. But although the function is unknown, we assume that it is smooth.

1.4.1. Acquisition functions more information

Many acquisition functions can be interpreted in Bayesian decision theory as evaluating an expected loss associated with evaluating f at a point x. We then select the point with the lowest expected loss. In the below we drop the f — D subscripts on the mean μ and covariance K functions for f.

Probability of improvement

Suppose

$$f_{min} = \min f$$

- minimum value observed so far. Probability of improvement evaluates f at the point most likely to improve upon this value. This corresponds to the following utility function associated with evaluating f at a given point x:

$$u(x) = \begin{cases} 0 \; f(x) > f_{min} \\ 1 \; f(x) \leq f_{min} \end{cases}$$

So we get a unit reward if $f(x)$ is less than f_{min}. Otherwise - no reward. Probability if improvement:

$$a_{PI}(x) = \mathbb{E}\big[u(x)|x, \mathcal{D}\big] = \int_{-\infty}^{f_{min}} \mathcal{N}\big(f; \mu(x), K(x,x)\big) df = \\ = \Phi\big(f_{min}; \mu(x), K(x,x)\big)$$

The point with the highest probability of improvement (the maximal expected utility) is selected.

Expected improvement

An alternative acquisition function that accounts for the size of the improvement is *expected improvement*. Suppose that f_{min} is the minimal value of observed f. Expected improvement evaluates f at the point that, in expectation, improves upon f_{min} the most. This corresponds to the utility function:

$$u(x) = \max(0, f_{min} - f(x))$$

So here we have the reward equal to the "improvement" $f_{min} f(x)$ if $f(x) \leq f_{min}$, and no reward otherwise.

$$a_{PI}(x) = \mathbb{E}\big[u(x)|x, \mathcal{D}\big] = \int_{-\infty}^{f_{min}} \big(f_{min} - f\big)\mathcal{N}\big(f; \mu(x), K(x,x)\big) df = \\ = \big(f_{min} - \mu(x)\big)\Phi\big(f_{min}; \mu(x), K(x,x)\big) + \\ + K(x,x)\mathcal{N}\big(f_{min}; \mu(x), K(x,x)\big)$$

The point with the highest probability of improvement (the maximal expected utility) is selected.

The expected improvement has two components. The First can be increased by reducing the mean function $\mu(x)$. The second can be increased by increasing the variance $K(x,x)$. These two terms can be interpreted as explicitly encoding a tradeoff between *exploitation* and exploration (evaluating at points with high uncertainty.

Upper confidence bound

Alternative acquisition function is typically known as *ucb*. In the context of minimization, the acquisition function would take the form

$$a_{ucb}(x, \beta) = \mu(x) - \beta\sigma(x)$$

where $\beta > 0$ - is a tradeoff parameter and $\sigma(x) = \sqrt{K(x,x)}$ is the marginal standard deviation of $f(x)$. This acquisition function contains explicit exploitation ($\mu(x)$) and exploration ($\sigma(x)$) terms.

Abbr.	Acquisition Function
EI	$\mathbb{E}_y\big[max(ReLU(y-\alpha))\big]$
PI	$\mathbb{E}_y\big[max(1^-(y-\alpha))\big]$
UCB	$\mathbb{E}_y\big[max(\mu + \sqrt{\frac{\beta\pi}{2}}\|\gamma\|)\big]$

(Wilson et al., 2018)

Entropy search Last considered alternative is *entropy search*. Here, we want to minimize the uncertainty we have in the location of the optimal value.

$$x^* = \text{argmin}_{x \in X} f(x)$$

We believe that f induce a distribution over x^*, $p(x^*|D)$. Unfortunately, there is no closed-form expression for this distribution.

Entropy search aims to evaluate points so as to minimize the entropy of the induced distribution $p(x^*|D)$.

The utility function:

$$u(x) = H\big[x^*|D\big] - H\big[x^*|D, x, f(x)\big]$$

As in probability of improvement and expected improvement, we may build an acquisition function by evaluating the expected utility provided by evaluating f at a point x. Due to the nature of the distribution $p(x^*|D)$, this is somewhat complicated, and a series of approximations must be made.

2. Problem statement

The problem, that should be solved, can be formulated the following way. Suppose there is a discretized translocation time distribution for a successful case P_T, and similar discretized time distribution for an unsuccessful translocation P_F, as well as a percentage of successful translocation r (1.1). It is necessary to reproduce the energy profile $\mathcal{F}(s)$ corresponding to these conditions as accurately as possible, using Bayesian optimization (1.4).

2.1. Parametrization

First of all we should parametrized $\mathcal{F}(N, s)$ using Gaussian functions on the interval $(0; N)$ where N - number of monomers:

$$g_i(x) = \frac{A_i}{\sqrt{2\pi\sigma_i^2}} \exp^{\frac{-(x-\mu_i)^2}{2\sigma_i^2}}$$

where μ_i - fixed mean value for i gaussian distribution from $(a; b)$, where a and b - just parameters, characterized interval for mean value of gaussians. and A_i and σ_i^2 - optimized parameters: amplitude and variance of i Gaussian distribution. So if there are used n gaussians for parametrization of $\mathcal{F}(N, s)$ we totally have vector of parameters $p \in \mathbb{R}^{2n}$.

2.2. Target function

In our problem we consider initial time distribution for successful translocation $P_T(p_{init})$,similar time distribution for an unsuccessful translocation $P_F(p_{init})$, and a percentage of successful translocation $r(p_{init})$.

For each vector $p \in \mathbb{R}^{2n}$ using FP formalism distribution for successful translocation $P_T(p)$, for an unsuccessful translocation $P_F(p)$, and a percentage of successful translocation $r(p)$ can be obtained.

The objective function may be considered in a several ways:

1. $$f(p, p_{init}) = MSE(P_T(p_{init}), P_T(p)) + \\ + MSE(P_F(p_{init}), P_F(p)) + \alpha \cdot |r(p_{init}) - r(p)|$$
   where α - is a parameter, that can be chosen in a different ways.
2. $$f(p, p_{init}) = \frac{|r^{init} - r^p|}{r^{init}}$$

Or some other ways. (**will be completed later**)

2.3. Experiment setup

First of all we have initial dataset $\{p_{train}\}_i^m, p_i \in \mathbb{R}^{2n}$, where n - number of gaussians for parametrization and m - number of samples in train dataset. Also there is a true energy distribution, parametrized with vector $p_{true} \in \mathbb{R}^{2n}$. For each element of train dataset the objective function is obtained:

$$\{y_{train}\}_i^m = \{f(p^i_{train}, p_{true})\}_i^m, f_i(p^i_{train}, p_{true}) \in \mathbb{R}$$

Then the target function is approximated by Gaussian Processes (1.3.1, 1.3.2) with chosen kernel (1.3.4). After it we obtain new sample $p_{new} \in \mathbb{R}^{2n}$ while maximizing acquisition function (1.4.1). And finally the value of objective function $f(p_{new}, p_{true})$ is obtained.

References

Forrester, A., Sobester, A., and Keane, A. *Engineering design via surrogate modelling: a practical guide*. John Wiley & Sons, 2008.

Palyulin, V. V., Ala-Nissila, T., and Metzler, R. Polymer translocation: the first two decades and the recent diversification. *Soft matter*, 10(45):9016–9037, 2014.

Rasmussen, C. J., Vishnyakov, A., and Neimark, A. V. Translocation dynamics of freely jointed lennard-jones chains into adsorbing pores. *The Journal of chemical physics*, 137(14):144903, 2012.

Snoek, J., Larochelle, H., and Adams, R. P. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25, 2012.

Wang, Z. and de Freitas, N. Theoretical analysis of bayesian optimisation with unknown gaussian process hyper-parameters. *arXiv preprint arXiv:1406.7758*, 2014.

Williams, C. K. and Rasmussen, C. E. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

Wilson, J. T., Hutter, F., and Deisenroth, M. P. Maximizing acquisition functions for bayesian optimization. *arXiv preprint arXiv:1805.10196*, 2018.