Probabilistic numerics for deep learning

Presenter: Shijia Wang

Michael Osborne

Department of Engineering Science, University of Oxford

Deep Learning (DLSS) and Reinforcement Learning (RLSS) Summer School, Montreal 2017

Outline

1. Introduction
   - Probabilistic Numerics
2. Components
   - Probabilistic modeling of functions
   - Bayesian optimization
   - Bayesian stochastic optimization
   - Integration beats Optimization
3. Conclusion
   - Experiments
   - Papers

Outline

Definition

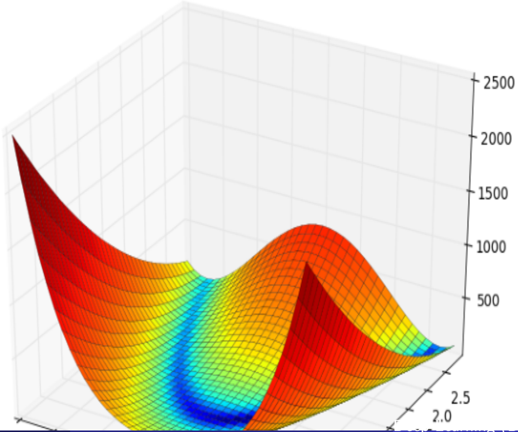
- Take the things were most interested in achieving and apply to computation
- Apply probability theory to numerics (computation cores)

Learning Algorithms

- Use numeric functions as learning algorithms
- Idea is to use Bayesian probability theories

Rosenbrock

$$f(x,y) = (1-x)^2 + 100(y-x^2)^2$$



Rosenbrock

- Easy to graph on a computer
- No easy way of finding its global minimum since it lies in a flat parabolic region
- Minimum $f(x, y) = 0$ when $(x, y) = (1, 1)$
- Reason: computational limits from the optimization problem

Uncertainty

- Epistemically uncertain about the function due to being unable to afford computation
- Probabilistically model function and use tools from decision theory to make optimal use of computation

Outline

Probability

- Probability is an expression of confidence in a proposition
- Probability theory can quantify inverse of logic expression
- Depends on the agent's prior knowledge

Gaussian Distribution

- Allows for distributions for variables conditioned on any other observed variables.
- Multivariate Gaussian Distribution:

$$\frac{1}{\sqrt{det(2\pi\Sigma)}} e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)}$$

- μ is mean
- Σ is covariance matrix

Gaussian Process

- **Gaussian Process** is a collection of random variables that any finite subset of the variables has a multivariate Gaussian distribution.
- Defined by mean and covariance function.
- Generalizes to potentially infinite number of variables.

Kernels/Covariance Function

- Squared exponential kernel:

$$K_{SE}(x_1, x_2) = Aexp(-\frac{1}{2}\sum_{d \in D}\frac{(x_{1d} - x_{2d})^2}{h_d})$$

- A the signal variance matrix, describes variation from the mean
- h_d the lengthscale, describes smoothness

Kernels/Covariance Function

- Matern kernel:

$$K_{Matern(3/2)}(x_1, x_2) = A(1 + \sqrt{3}r)exp(-\sqrt{3}r)$$

$$K_{Matern(5/2)}(x_1, x_2) = A(1 + \sqrt{5}r + \frac{5}{3}r^2)exp(-\sqrt{5}r)$$

- A the signal variance matrix, describes variation from the mean
- $r = \sqrt{\sum_{d \in D} \frac{(x_{2d} - x_{1d})^2}{h_d}}$
- h_d the lengthscale, describes smoothness

Mean

- Posterior estimates:

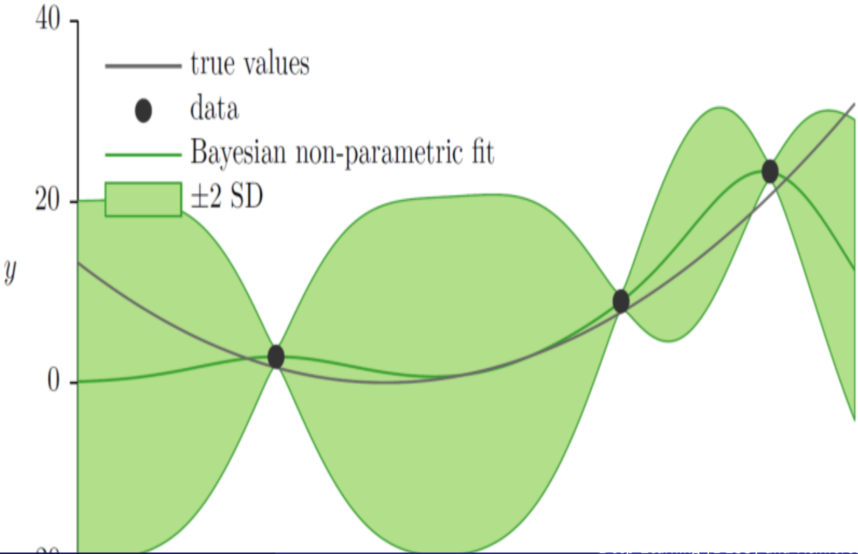
$$m(x|D) = \frac{1}{K}\sum_{k=0}^{K} m(x|D, \lambda_k)$$

- K number of draws of the hyperparameter values that have been made by slice sampling
- λ prior data observed

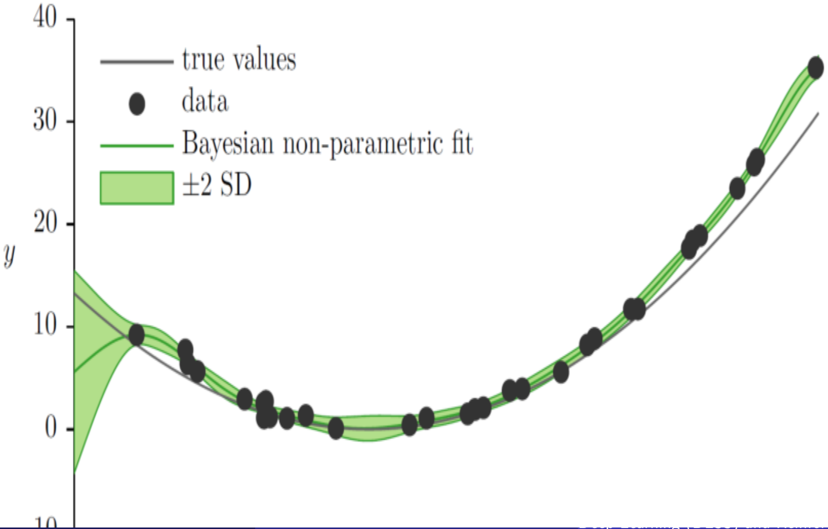
Gaussian Process

- Complexity that grows with data
- Robust to overfitting

Gaussian Process



Gaussian Process



Outline

Bayesian Optimization

- Bayesian optimization is the approach of probabilistically modelling $f(x, y)$ and using decision theory to make optimal use of computation
- By defining the costs of observation and uncertainty, we can select evaluations optimally by minimizing the expected loss with respect to a probability distribution
- Representing the core components: cost evaluation and degree of uncertainty

Acquisition Function

- **Acquisition Function** $\alpha(x)$ quantifies how valuable evaluating at x is expected to be
- Evaluated on the GP rather than the objective.
- Since working on GP is less costly, can find its global maximum and use the point as the next evaluation of the objective function.

Entropy Search AF

- Optimization is viewed as gaining knowledge about the location of the global minimum.
- Prior belief about the location of the global minimum of the objective is represented as a probability distribution $p(x_*)$. The probability that $x_* = argmin_x f(x)$
- Selects points to maximize the relative entropy of this distribution from the uniform distribution:

$$x_{n+1} = argmax_x(H[p(x_*|D_n)] - E_{x_*}[H[p(x_*|D_n, x, y)]])$$

- $H[p] = -\sum_i p_i log p_i$ entropy

Predictive Entropy Search

- The acquisition function $\alpha(x)$ is the expected information gain about the value at x_{n+1} given a true observation of the global minimum:

$$\alpha(x_{n+1}) = H[y_{n+1}|D_n, x_{n+1}] - H[y_{n+1}|D_n, x_{n+1}, x_*]$$

Loss

- loss function - lowest function value found after algorithm ends
- Take a myopic approximation and consider only the next evaluation
- The expected loss is the expected lowest value of the function we've evaluated after the next iteration

Myopic Loss

Consider only with one evaluation remaining, the loss of returning value y with current lowest value μ

$$\lambda(y) \triangleq \begin{cases} y; & y < \eta \\ \eta; & y \geq \eta \end{cases}.$$

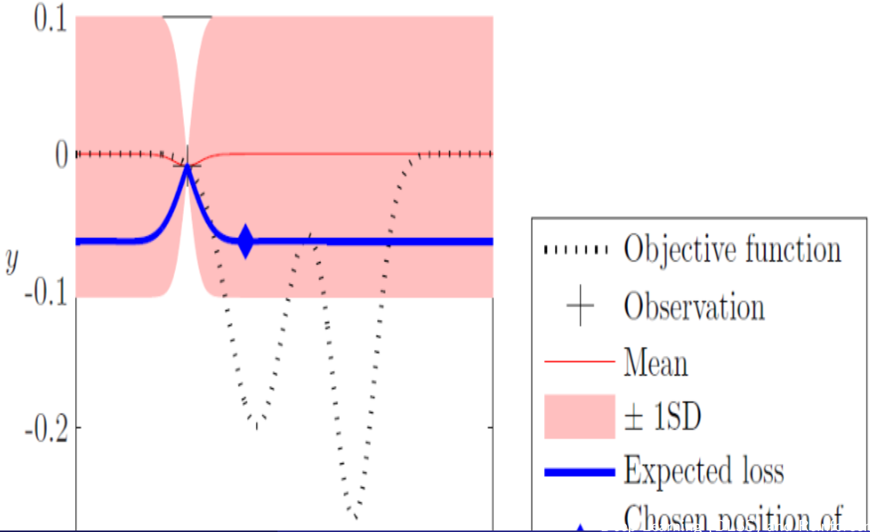
Expected Loss

Expected loss is the expected lowest value

$$\int \lambda(y)\, p(y \mid x, I_0)\, \mathrm{d}y$$

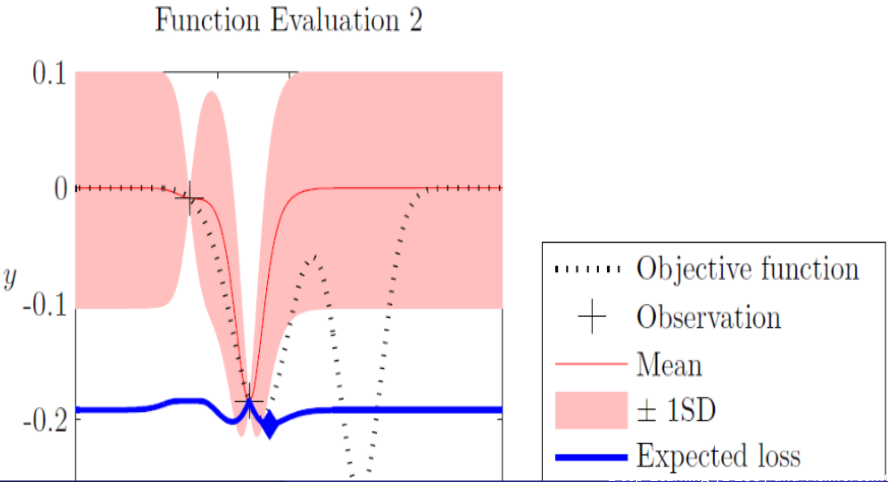
Expected Loss

Use a Gaussian process as the probability distribution for the objective function



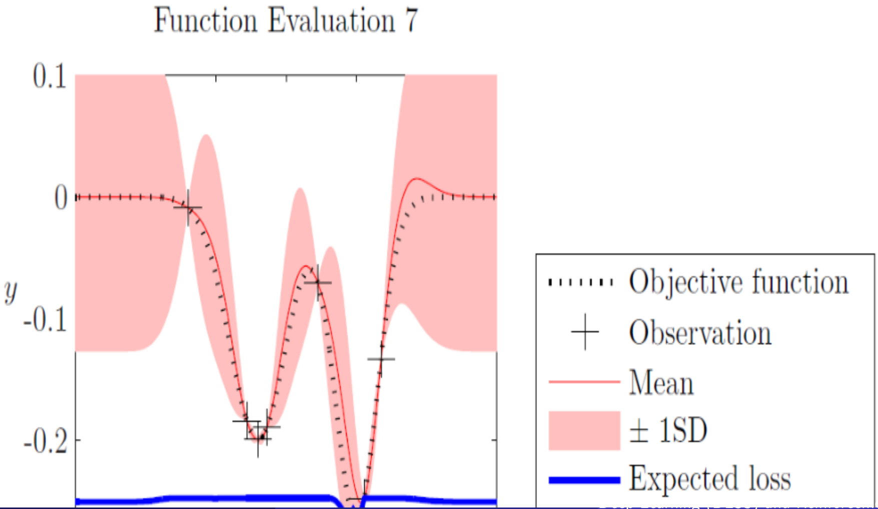
Expected Loss

Exploitative step



Expected Loss

Exploratory step



Outline

Noise

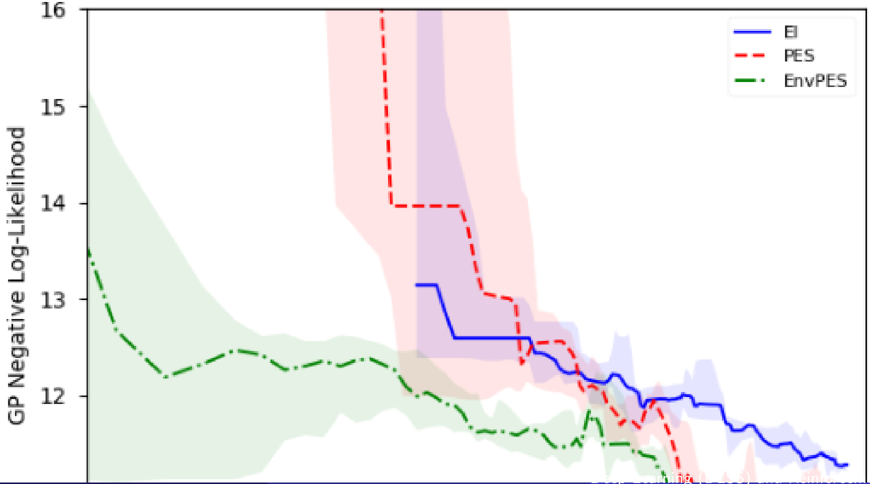
- Using only a subset of the data gives a noisy likelihood evaluation
- Use Bayesian optimization for stochastic learning

Noise

- Within Bayesian Optimization noise is not a problem
- If additional noise in the random variable we can just add a noise likelihood to complement model
- Encode that cost as a function of the number of data
- Intelligently choose the size of data that it needs at runtime to best optimization

Bayesian Optimization

Batch size
Klein, Falkner, Bartels, Hennig, Hutter (2017);
McLeod, Osborne, Roberts (2017)



Outline

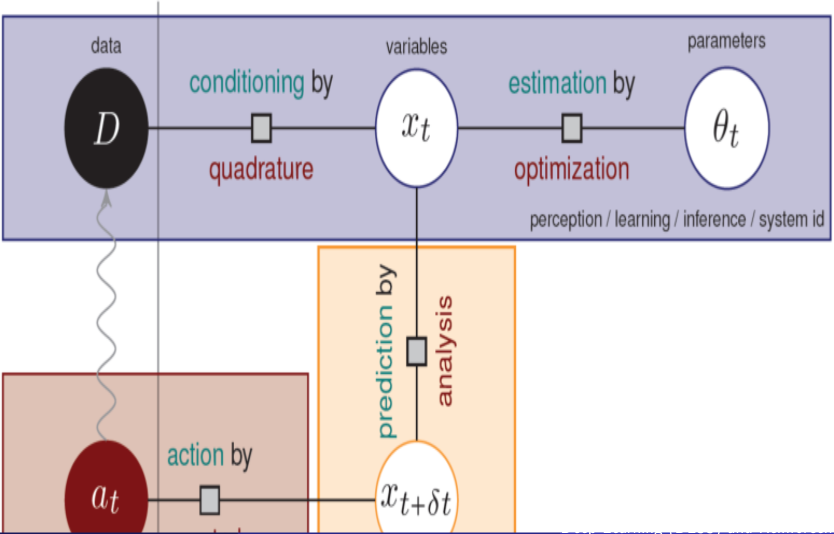
Integration

- Normally we want integration rather than optimization
- Average over the calculated parameters and functions by their likelihoods
- Reduces uncertainty of calculated functions
- Uses Bayesian quadrature for numerical integration

Outline

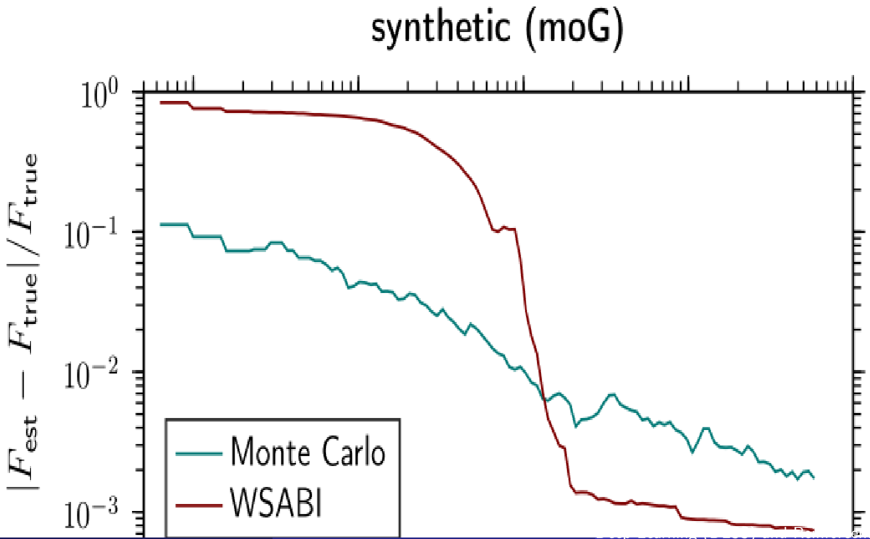
Model

Propagates uncertainty



Model

Converges



Outline

Papers

- McLeod, Osborne & Roberts (2017). Practical Bayesian Optimization for Variable Cost Objectives. arxiv.org/abs/1703.04335
- Gunter, T., Osborne, M. A., Garnett, R., Hennig, P., & Roberts, S. J. (2014). Sampling for Inference in Probabilistic Models with Fast Bayesian Quadrature. In Advances in Neural Information Processing Systems (NIPS).