SageMaker Automatic Model Tuning

Part 1: Theory

© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Many hyperparameters for training

Model:

- Number of layers: 1, 2, 3, …
- Activation functions: Sigmoid, tanh, RELU, …

Optimization:

- Method: SGD, Adam, AdaGrad, …
- Learning Rate: 0.01 to 2

Data:

- Batch Size: 8, 16, 32 …
- Augmentation: Resize, Normalize, Color Jitter, …

© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Choose function to optimize

Optimize model-params: λ

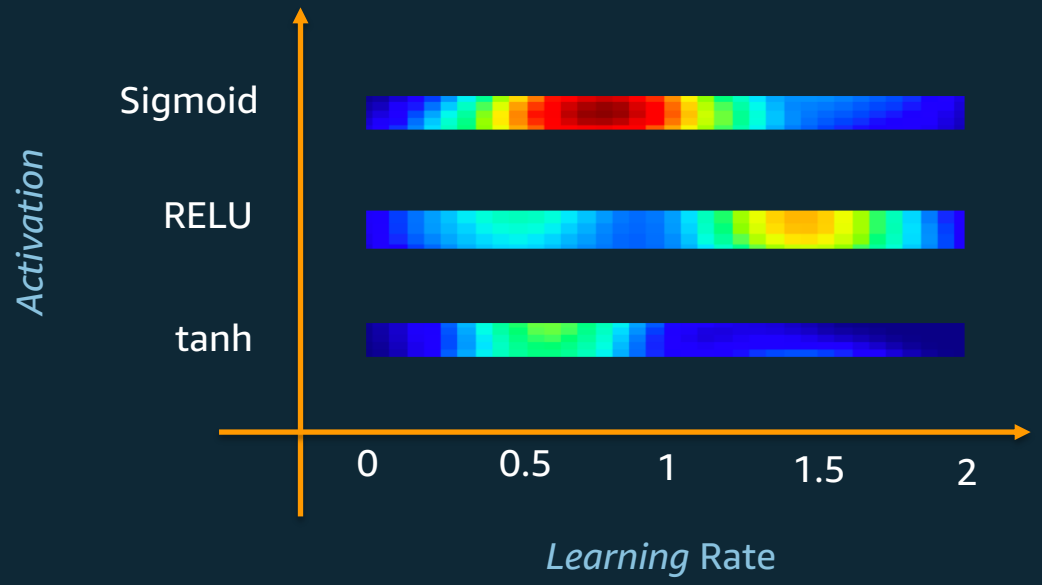
$$\min_{\lambda} f(\lambda|\theta)$$

Optimize hyper-params: θ

$$\min_{\theta} h(\theta)$$

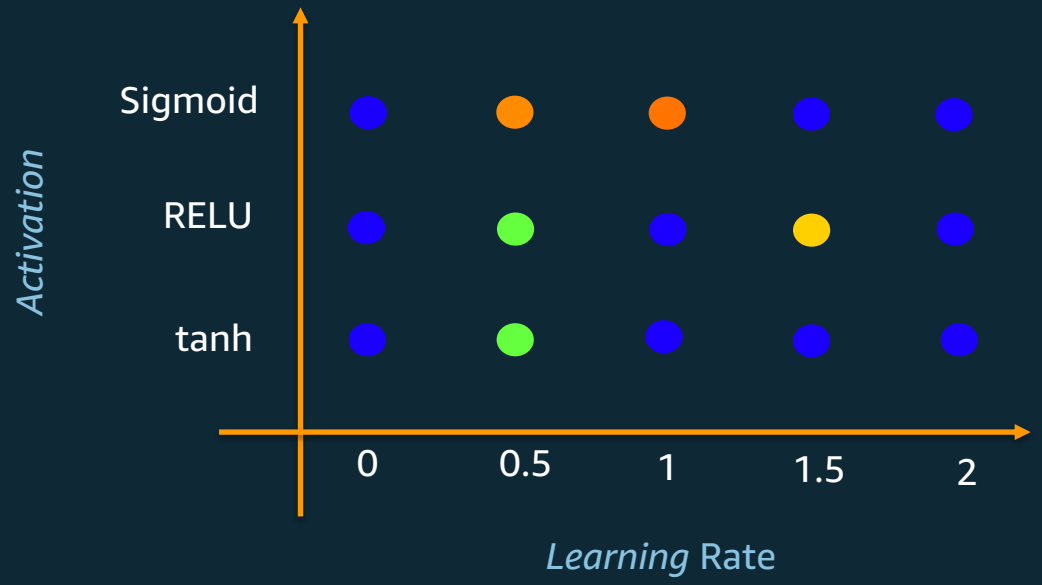
© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Grid Search



© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Grid Search



© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Grid Search

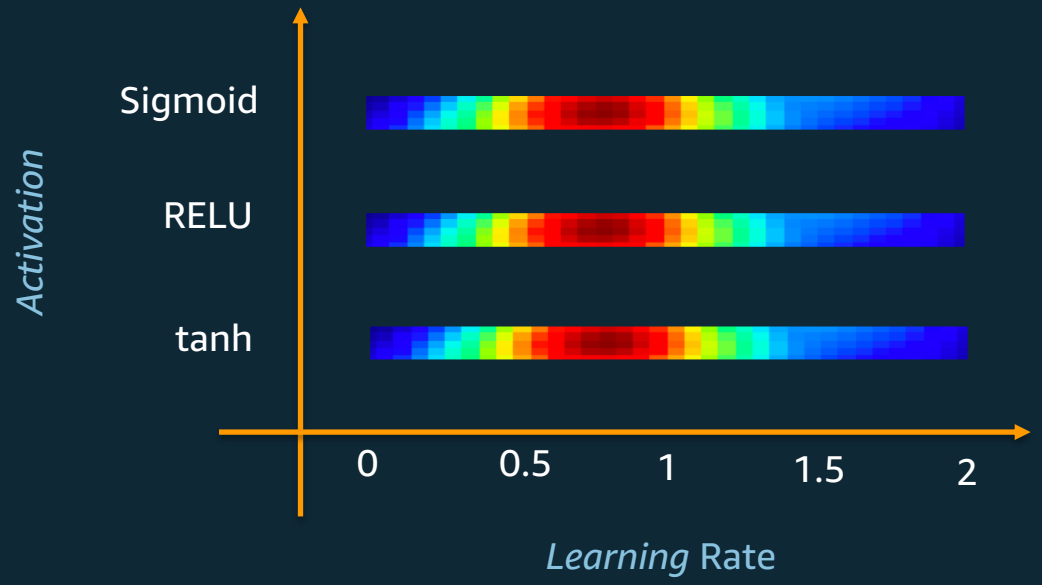
Assume each hyper-param has 5 options
e.g. Learning Rate: 0, 0.5, 1, 1.5, 2

1 HP = 5 combinations
2 HPs = 5*5 = 25 combinations
3 HPs = 5*5*5 = 125 combinations
…
10 HPs = 5^10 = 9,765,625 combinations

N HPs = 5^N combinations

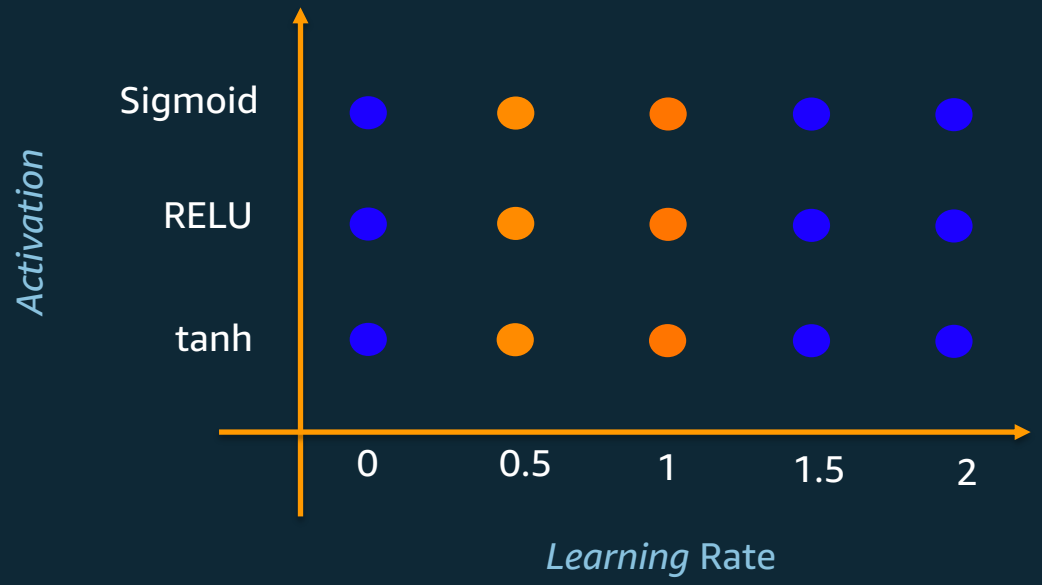
© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Grid Search



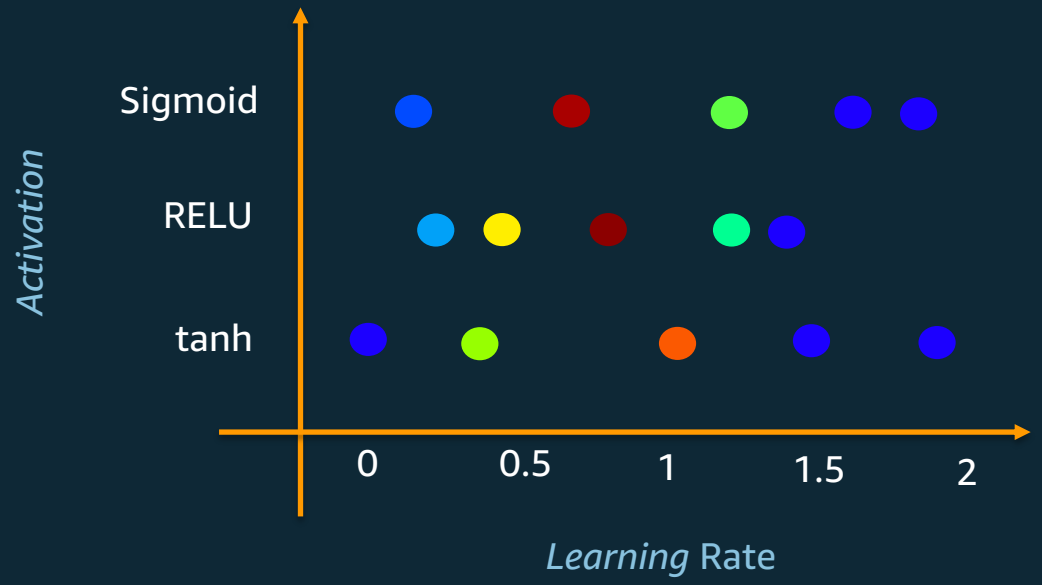
© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Grid Search



© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Random Grid Search



© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Create model of model loss

Approximate expensive $h(\theta)$

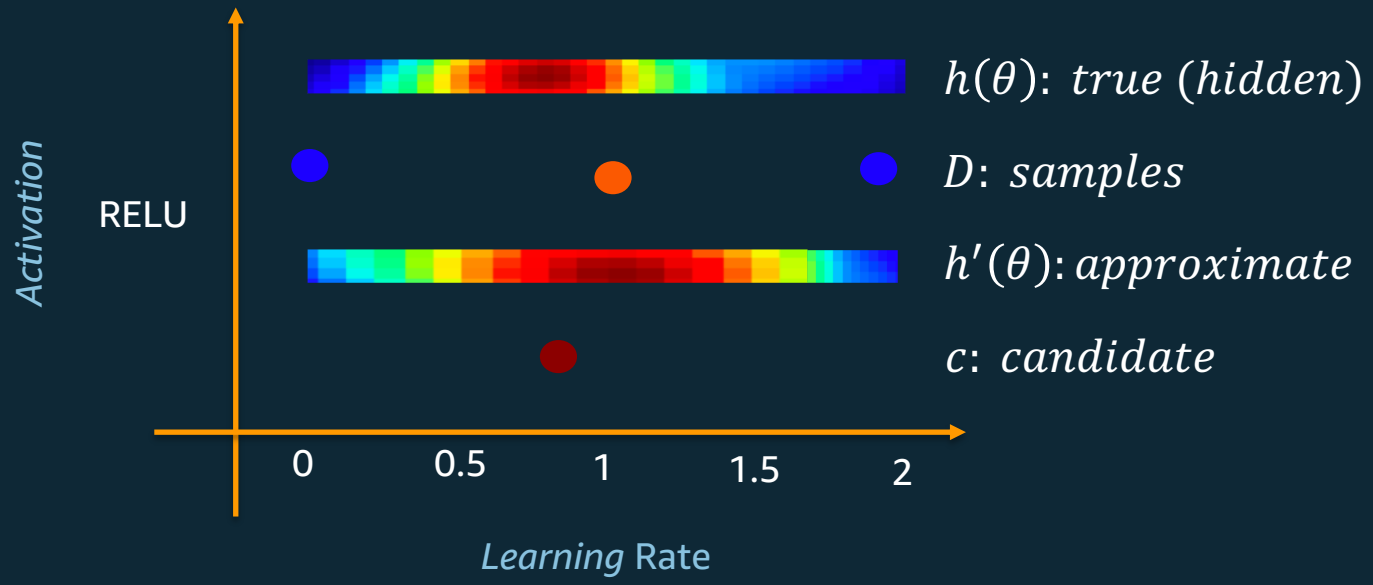
$$h(\theta) \approx h'(\theta)$$

Use previous evaluations: D

$$h'(\theta|D)$$

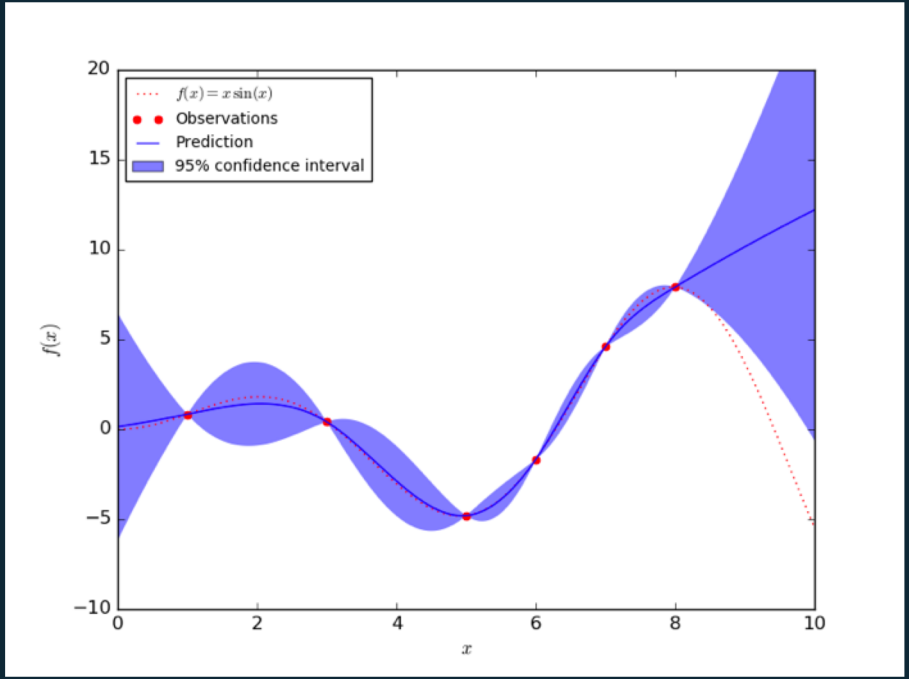
© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Use model of model loss



© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Gaussian Process for model of model loss



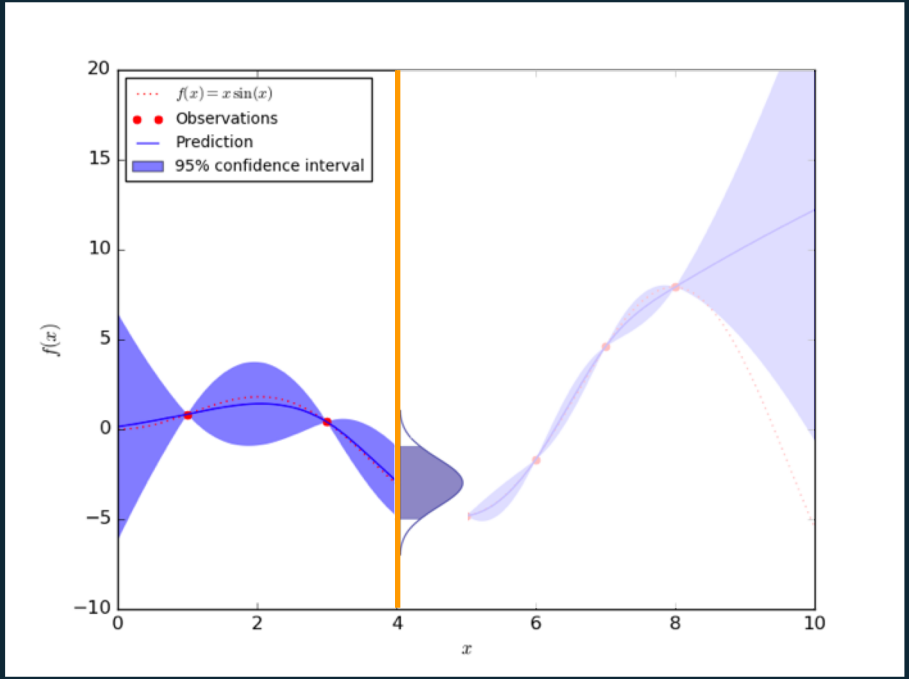
© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Gaussian Process

$$f(X_{t_1}), f(X_{t_2}), \ldots, f(X_{t_n}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

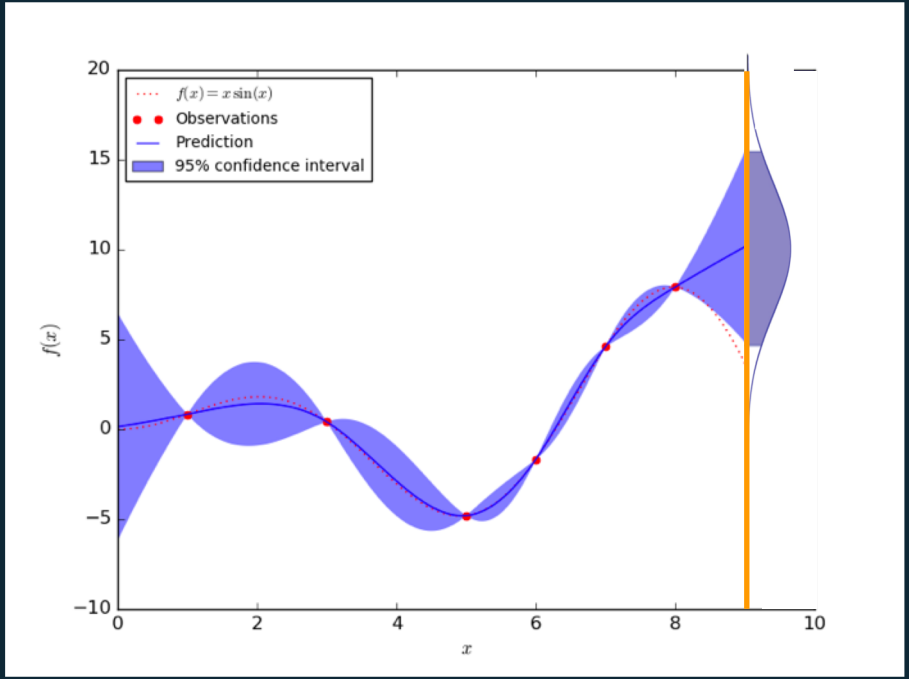
© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Gaussian Process for model of model loss



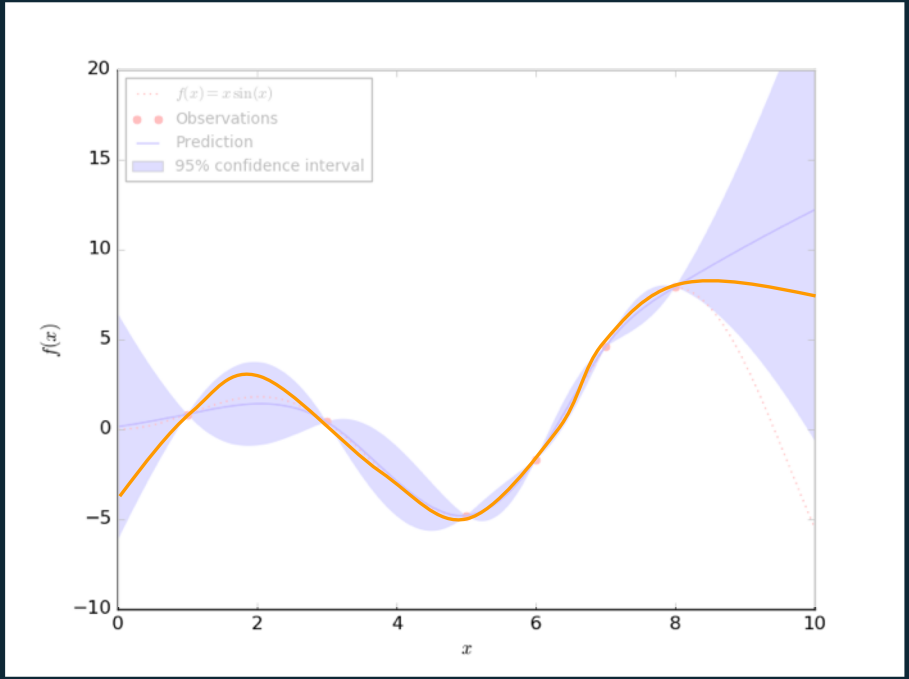
© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Gaussian Process for model of model loss



© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Gaussian Process for model of model loss

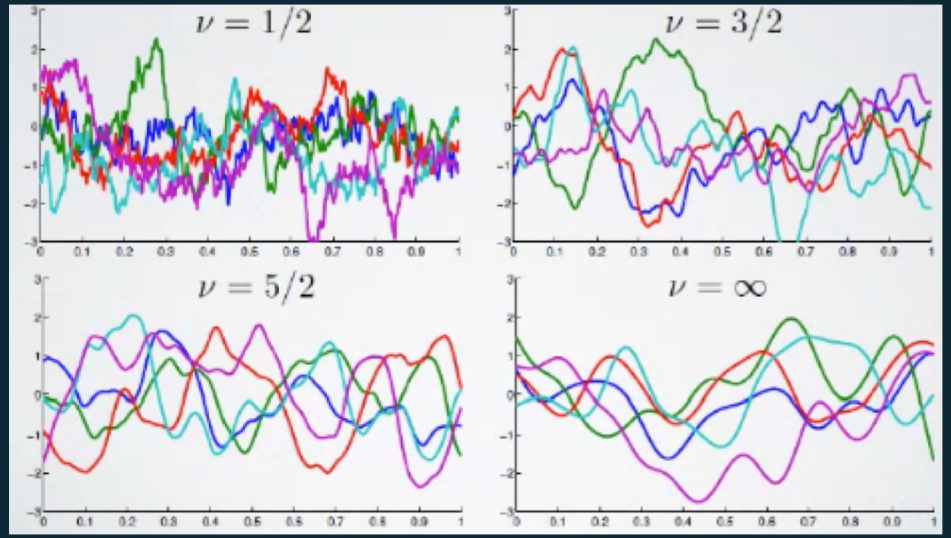


© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Covariance Matrix

Similarity between 2 points: controls ‘smoothness’.

SageMaker uses Matérn kernel with $\upsilon = 5/2$



© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Closed form expressions

- For any two vectors $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$:

$$K_\theta(\mathbf{x}, \mathbf{x}') \triangleq \theta_0 \cdot \left[1 + \sqrt{5} r(\mathbf{x}, \mathbf{x}') + \frac{5}{3} r^2(\mathbf{x}, \mathbf{x}')\right] \cdot e^{-\sqrt{5} r(\mathbf{x}, \mathbf{x}')}$$

with

$$r^2(\mathbf{x}, \mathbf{x}') \triangleq \sum_{j=1}^{d} \frac{(\mathbf{x}_j - \mathbf{x}'_j)^2}{\theta_j^2}.$$

$$\begin{aligned}
\mu_\theta(f(\mathbf{x}) | \{\mathbf{x}^c, y^c\}_{c \in \mathcal{C}}) &\triangleq \mathbf{k}_\mathcal{C}(\mathbf{x})^\top [\mathbf{K}_\mathcal{C} + \theta_\sigma^2 \mathbf{I}]^{-1} \mathbf{y}_\mathcal{C} \\
\Sigma_\theta(f(\mathbf{x}) | \{\mathbf{x}^c, y^c\}_{c \in \mathcal{C}}) &\triangleq K(\mathbf{x}, \mathbf{x}) - \mathbf{k}_\mathcal{C}(\mathbf{x})^\top [\mathbf{K}_\mathcal{C} + \theta_\sigma^2 \mathbf{I}]^{-1} \mathbf{k}_\mathcal{C}(\mathbf{x})
\end{aligned}$$

© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Acquisition Function

Given posterior distribution of functions…

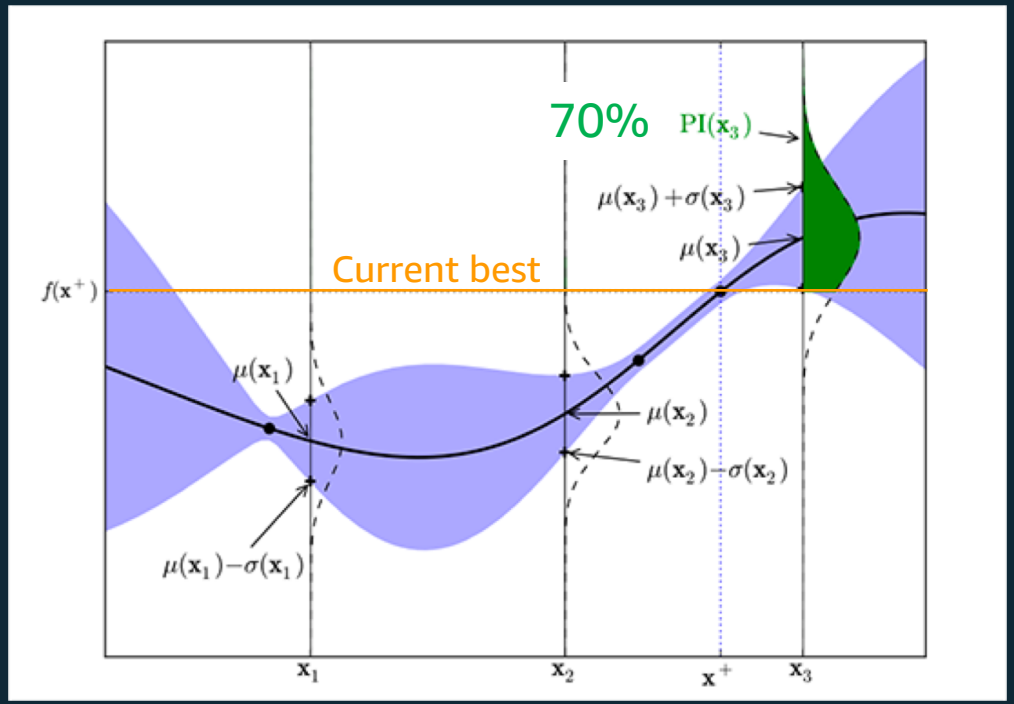
Used as criteria for selecting next candidate hyperparams for evaluation.

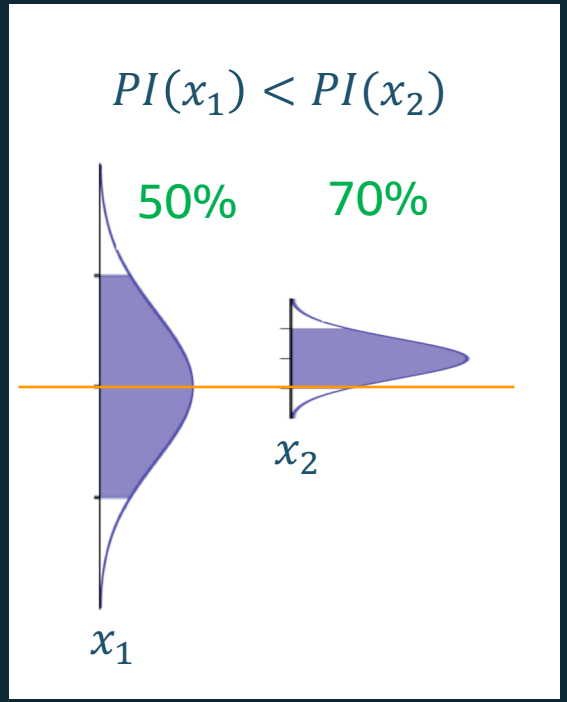
Often depends on the best hyperparams seen so far in search.

Controls exploration vs exploitation in search.

© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

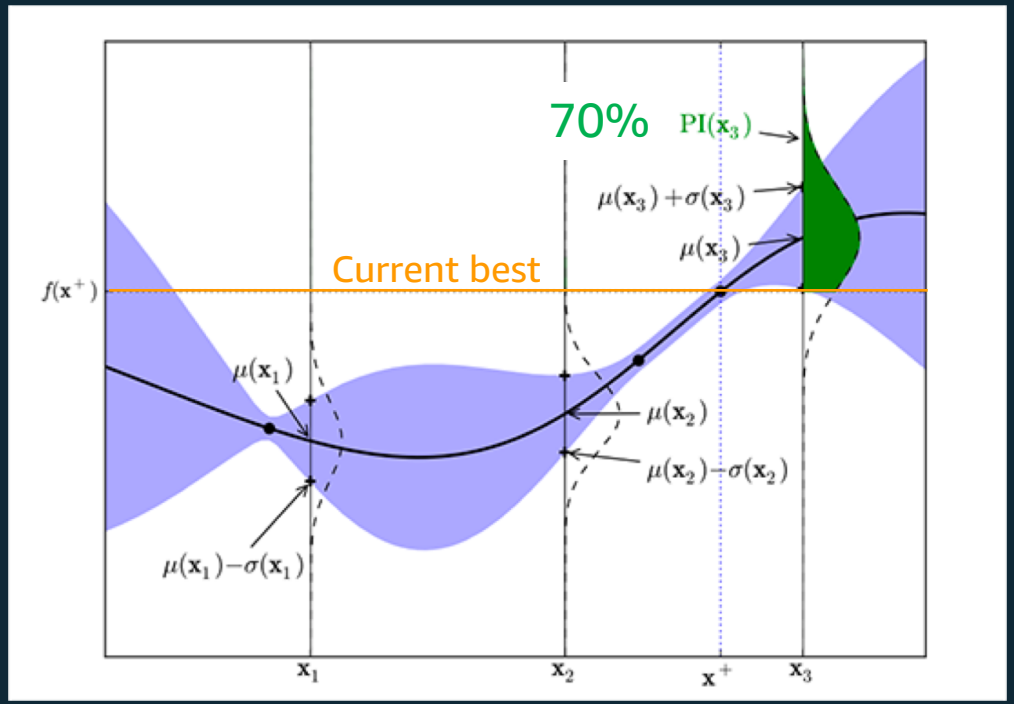
Acquisition Function: Probability of Improvement

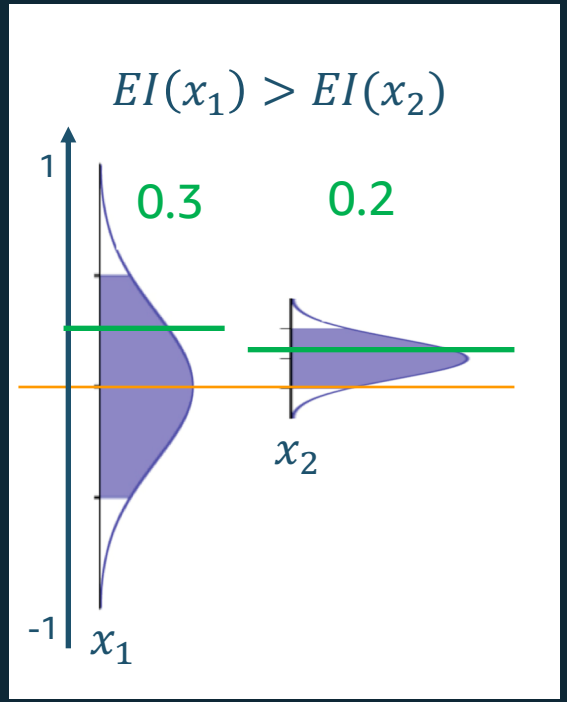




© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

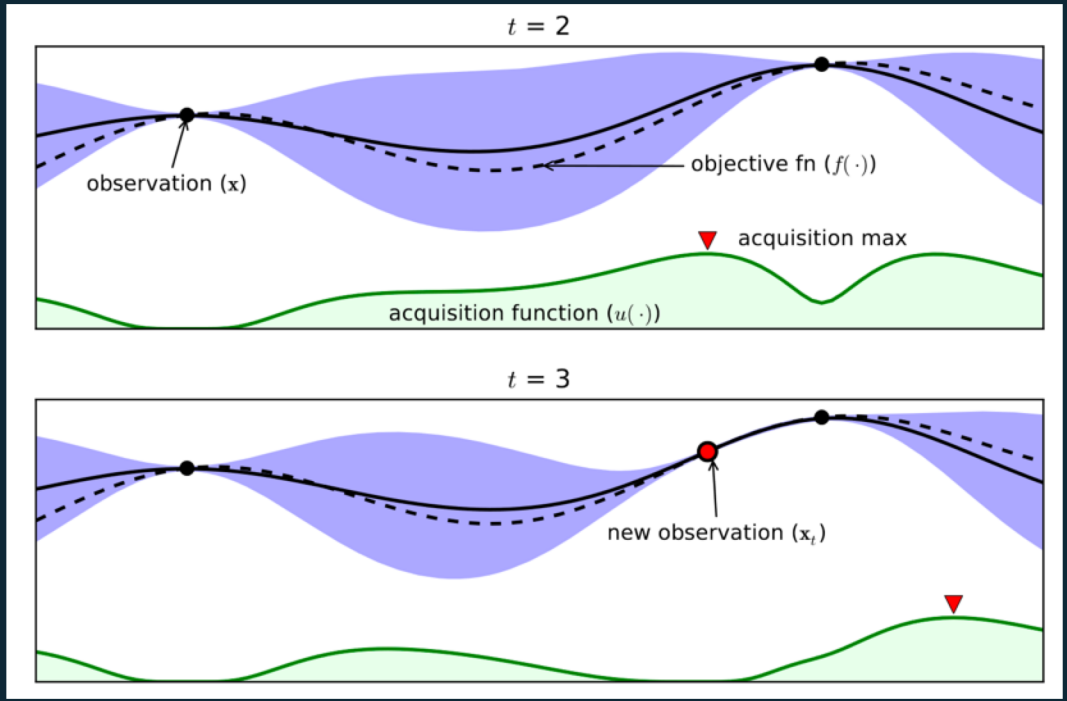
Acquisition Function: Expected Improvement





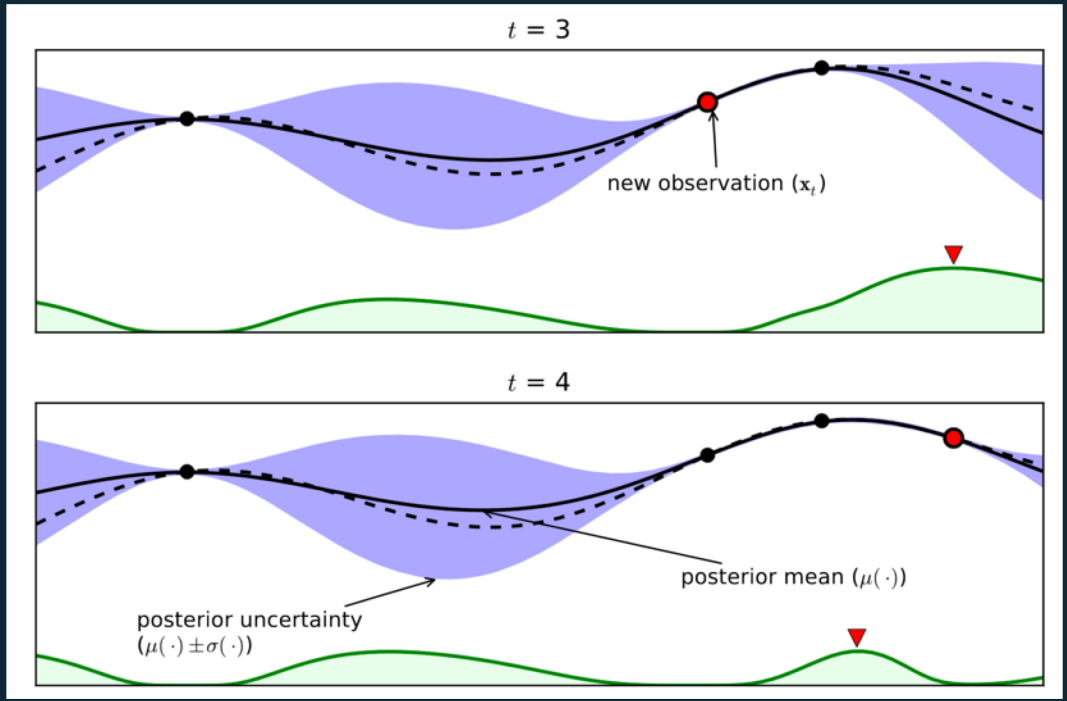
© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Using Acquisition Functions



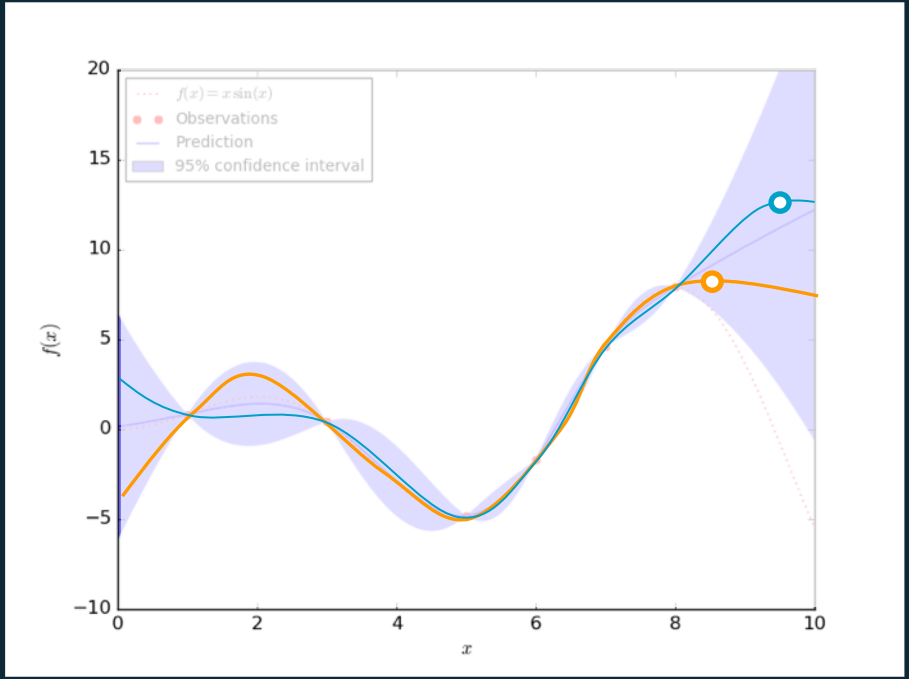
© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Using Acquisition Functions



© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Add parallelism with Thompson Sampling



© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Input Mapping

Create a ***unit hypercube*** from all input hyperparameters

Continuous – e.g. learning rate

$$[0,2] \rightarrow [0,1]$$

Integral – e.g. number of epochs

$$\{1,2,\ldots,50\} \rightarrow [0,1]$$

Categorical – e.g. type of activation

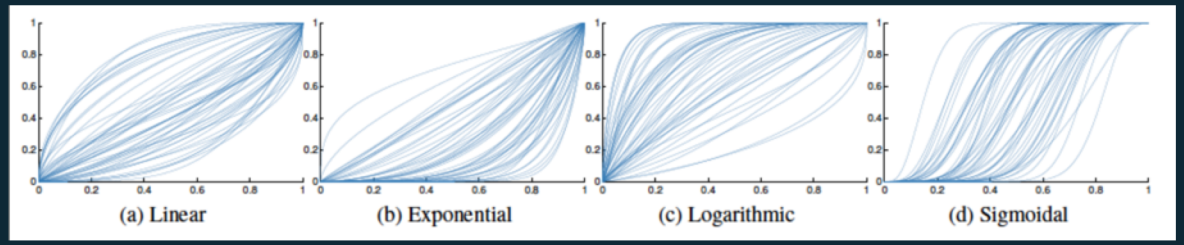
$$\{Sigmoid, RELU, tanh\} \rightarrow [0,1]^3$$

© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Input Warping

Apply transformation of variables within the hypercube.

Automatic selection of transform per variable from Beta CDF:



e.g. for learning rate

{0.3, 0.6, 0.9} less effective search space than

{0.01, 0.1, 1} (i.e. log transform)

© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Part 2: Practical

© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.