Bayesian Optimization

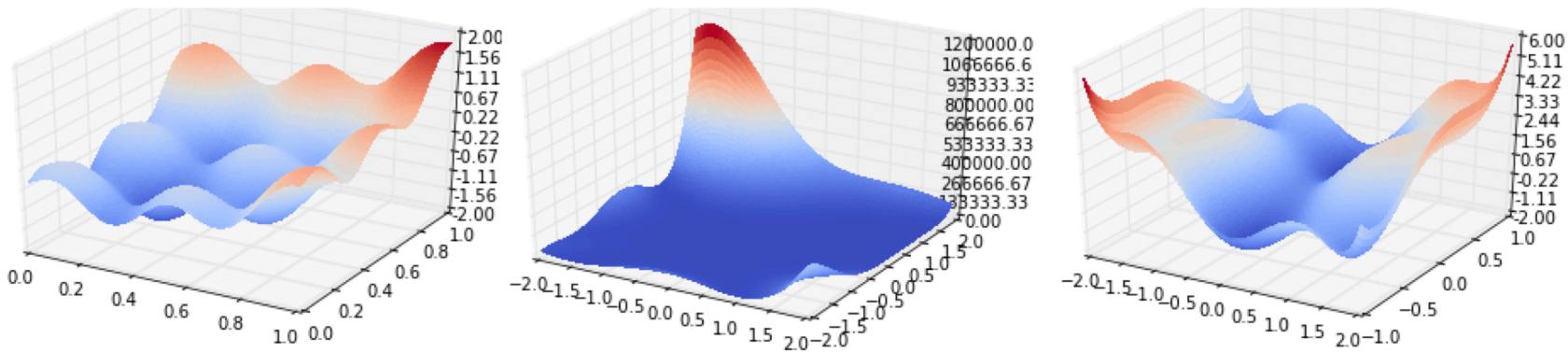
Evgeny Burnaev, Alexey Artemov

Skoltech, Moscow, Russia

Outline

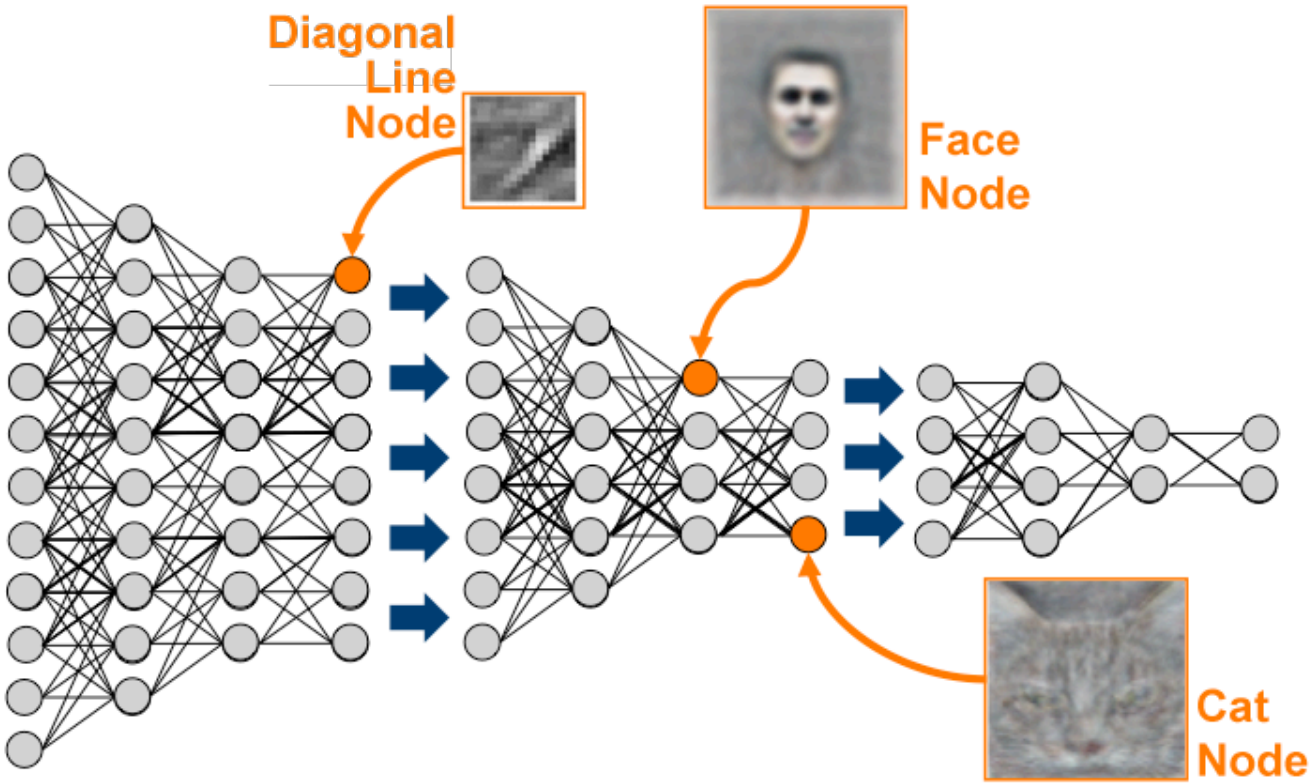
Consider a "well behaved" function $f : \mathcal{X} \to \mathbb{R}$, with $\mathcal{X} \subseteq \mathbb{R}^d$ being a compact set

$$x_{min} = \arg\min_{x \in \mathcal{X}} f(x)$$



- f is explicitly unknown and multimodal
- Evaluations of f may be perturbed
- Evaluations of f are expensive $\Rightarrow$
  - Gradient and Hessian are not computable
  - Grid search is not possible

Parameter tuning in ML algorithms



- Number of layers/units per layer
- Types of each layer
- Regularization coefficients
- Learning rates, etc.

Parameter tuning in ML algorithms: Example of DNN

Input x:

- Number of layers/units per layer
- Types of each layer
- Regularization coefficients
- Learning rates, etc.

Output: f(x)

- Deep Neural Network accuracy
- Estimated using cross-validation and/or test set
- Very time-consuming

Other expensive functions

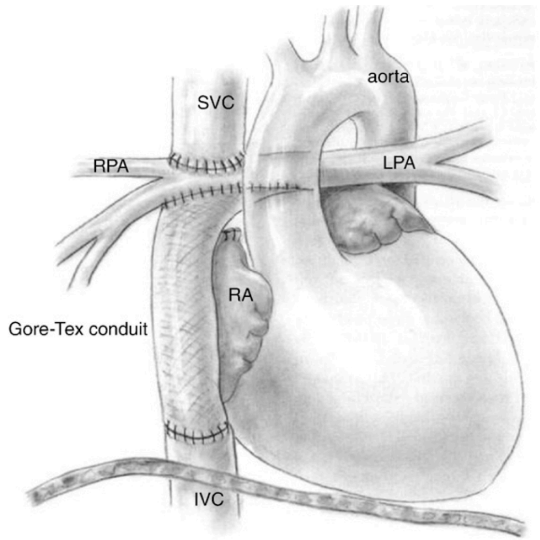


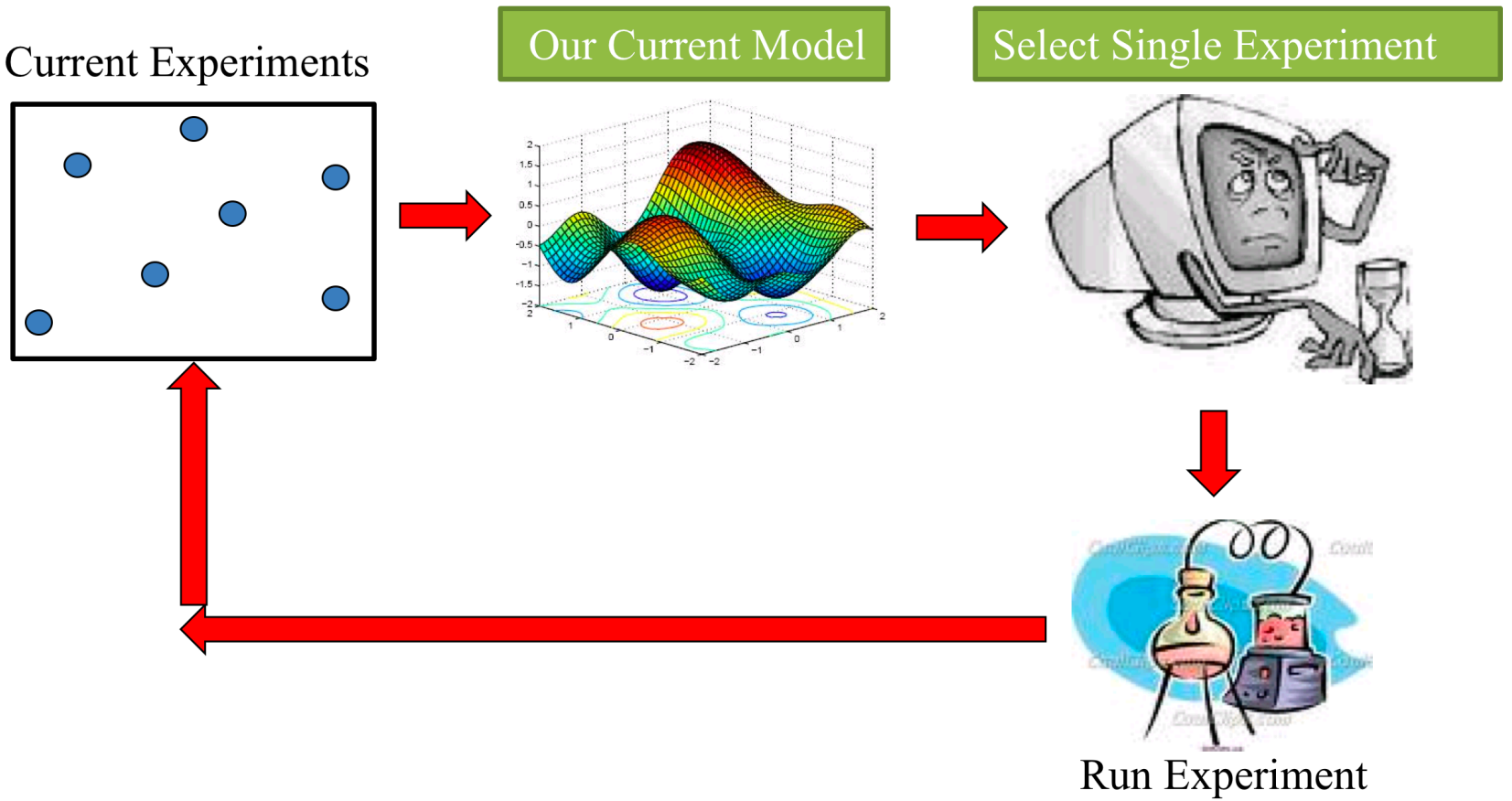
Fig. 1. Extracardiac total cavopulmonary connection. The IVC is disconnected from the right atrium (RA) and connected to the PAs via a Gore-Tex conduit. Figure taken from Reddy et al. [13].

- Design of grafts to be used in heart surgery
- Design of aeorodynamic structures, e.g., cars, airplanes
- Calibrating parameters of complex physical models to experimental data
- A/B testing data to optimize the web design to maximize sign-ups, downloads, purchases, etc.

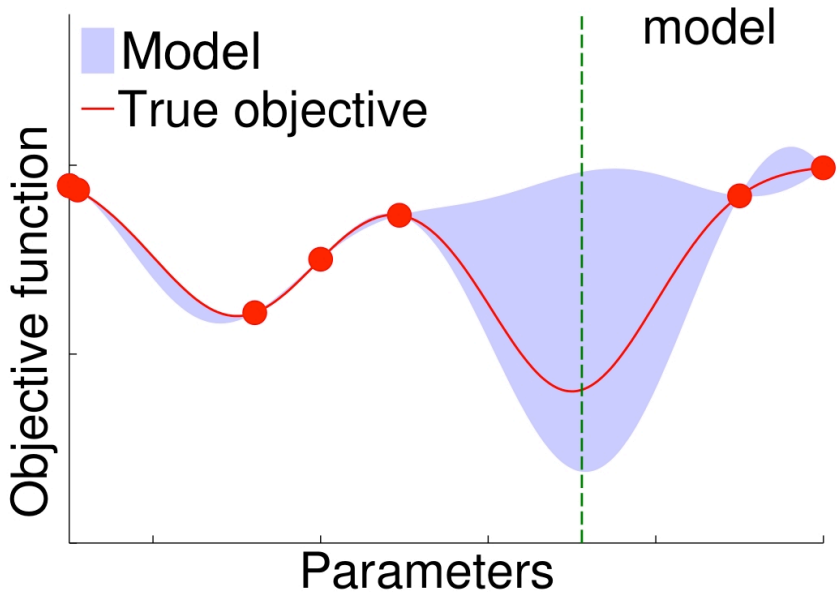
NB! There exists commercial services for optimizing black-box functions: SIGOPT, Google Vizier, etc.

Big Picture

- Since Running experiment is very expensive we use BO
- Select one experiment to run at a time based on results of previous experiments

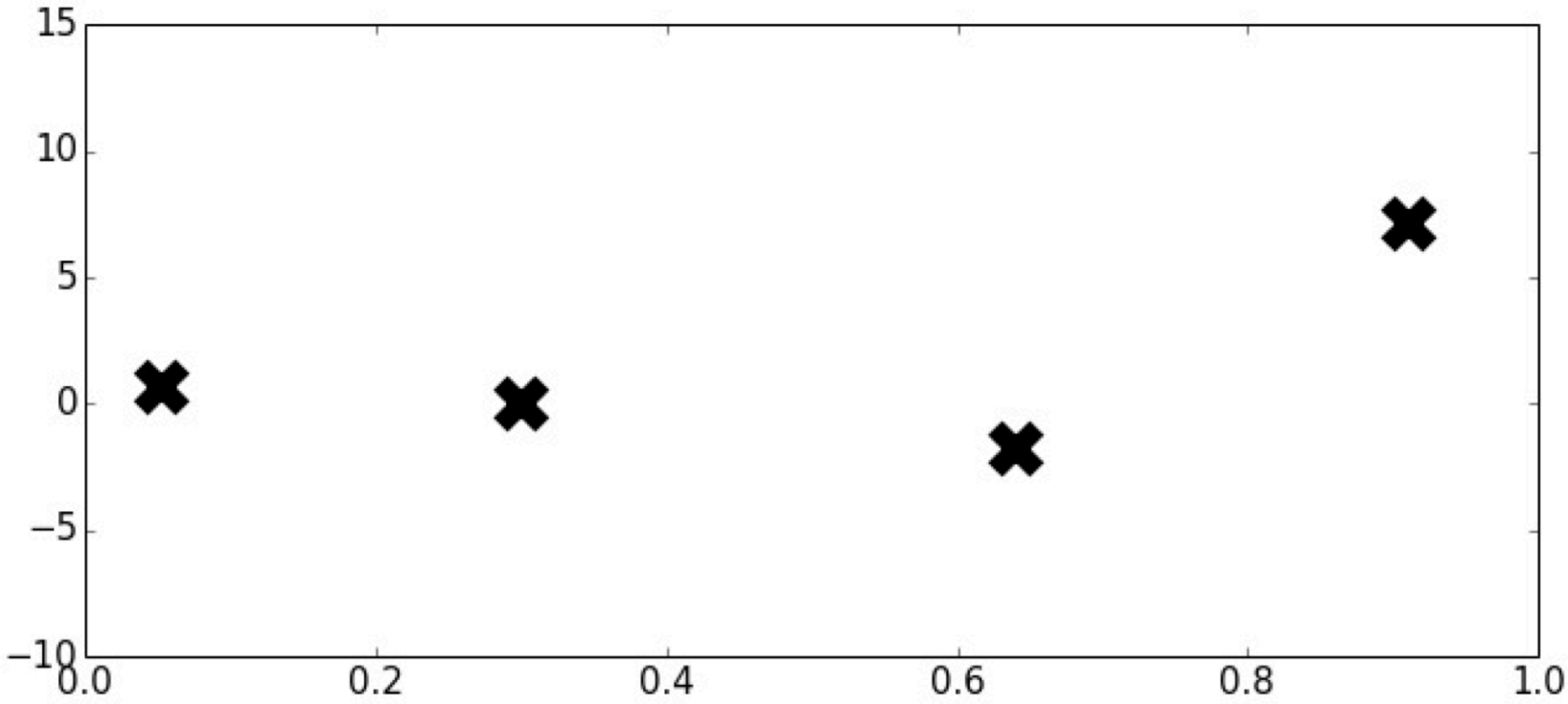


- In high-dimensional case we need many functions evaluations
- Often each evaluation is costly, e.g. in case of experiments



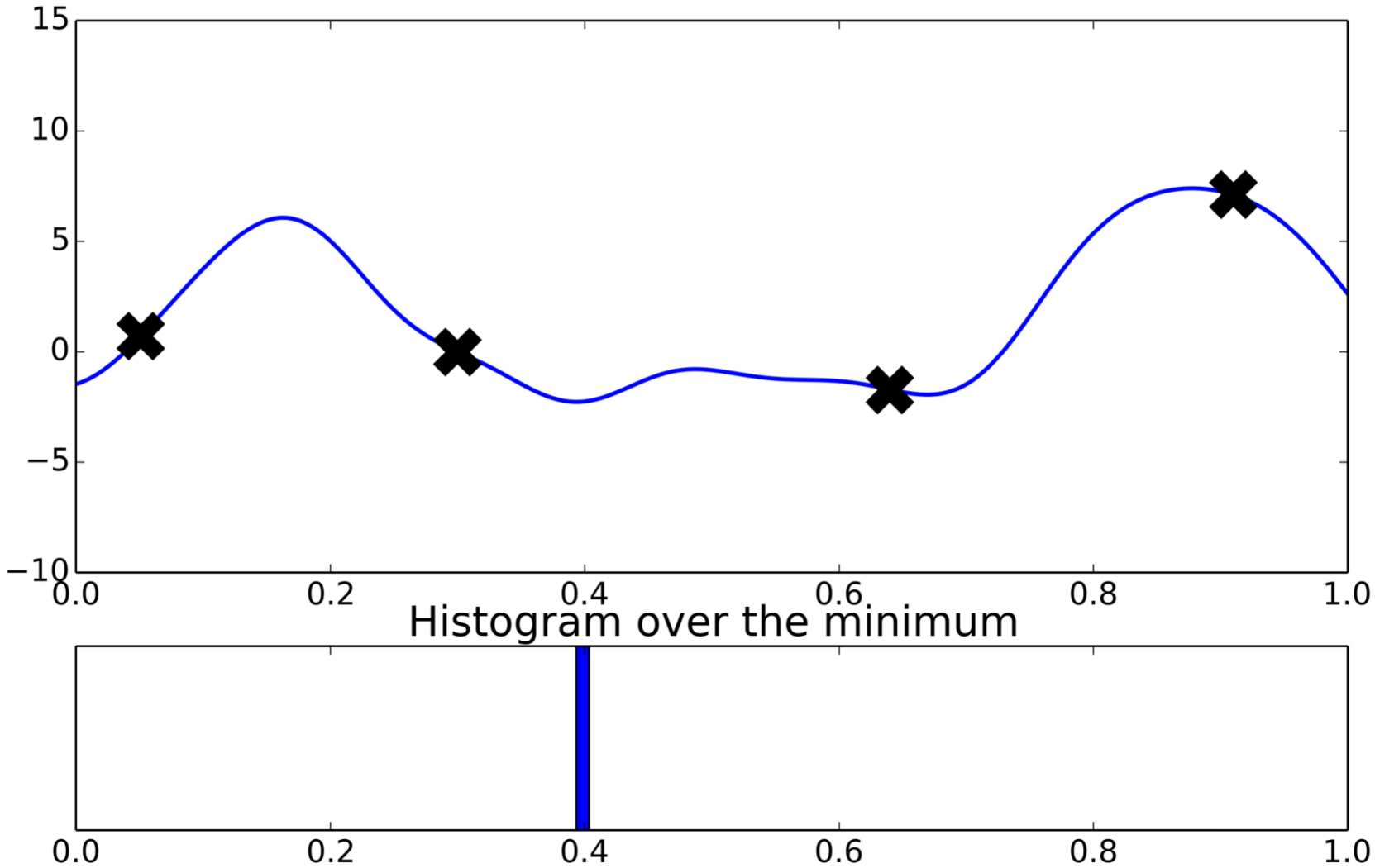
- Error bars are needed to see if a region is still promising

Typical situation

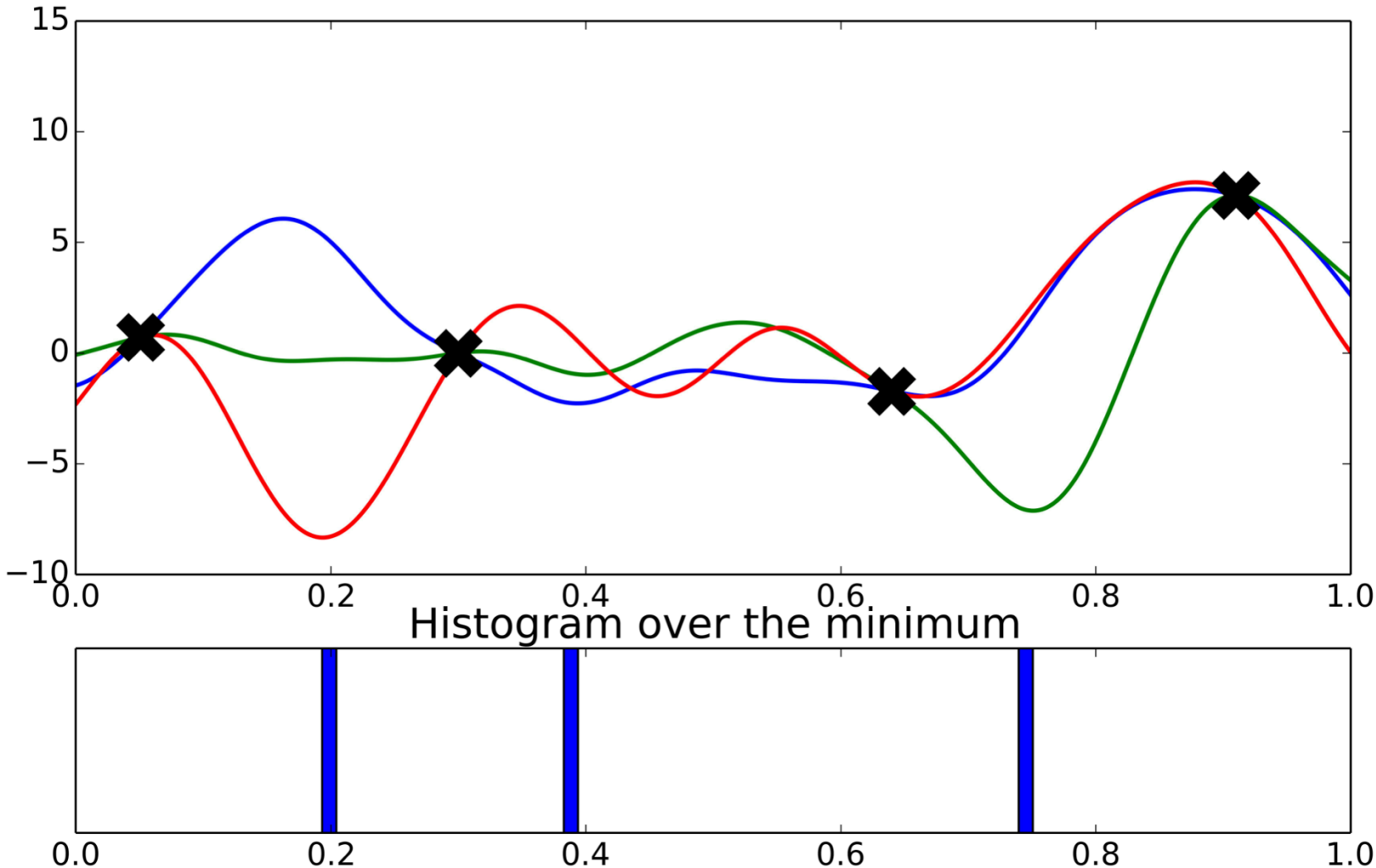


Where is the minimum of f?
Where should we evaluate the function next?

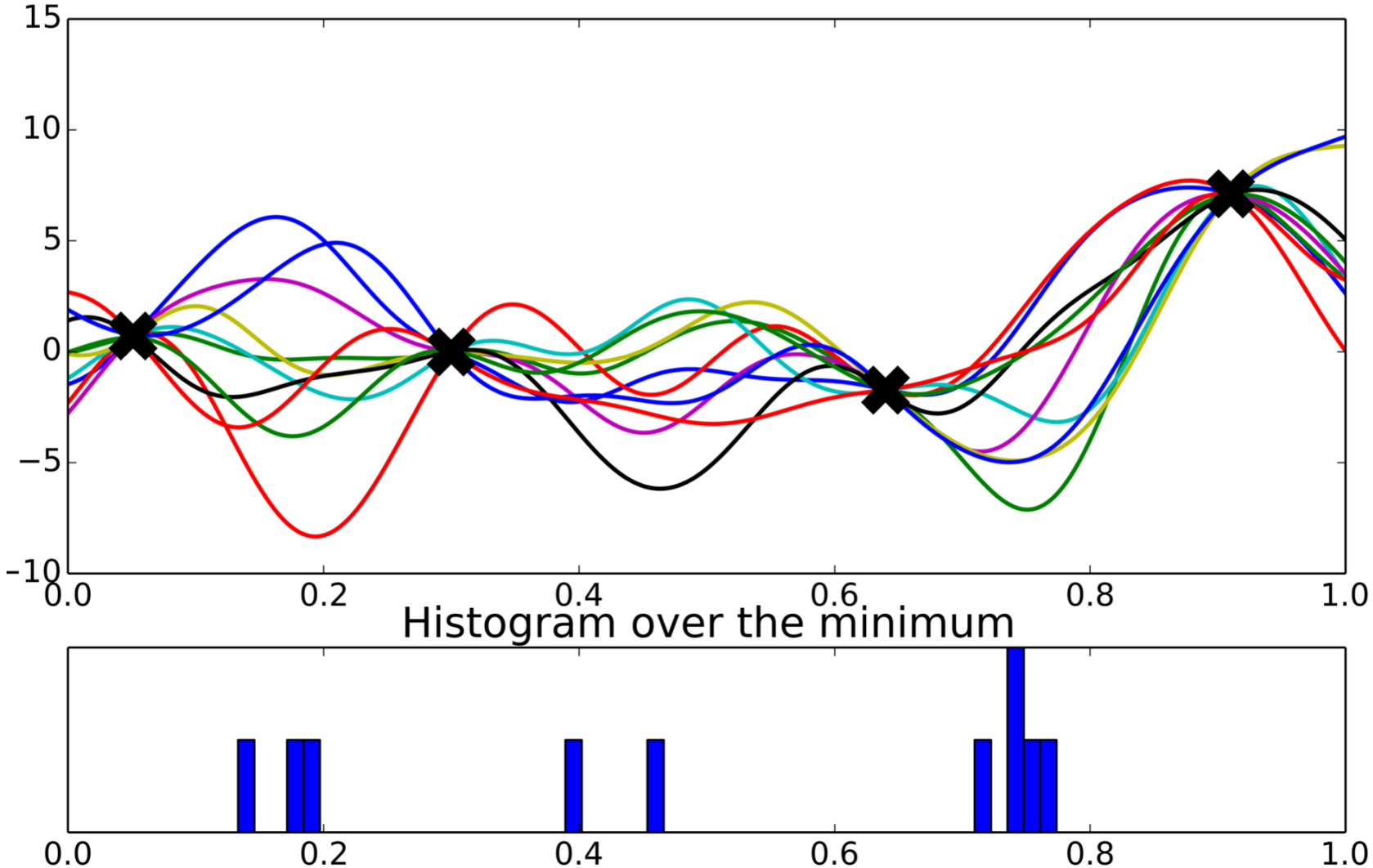
Intuition: one curve



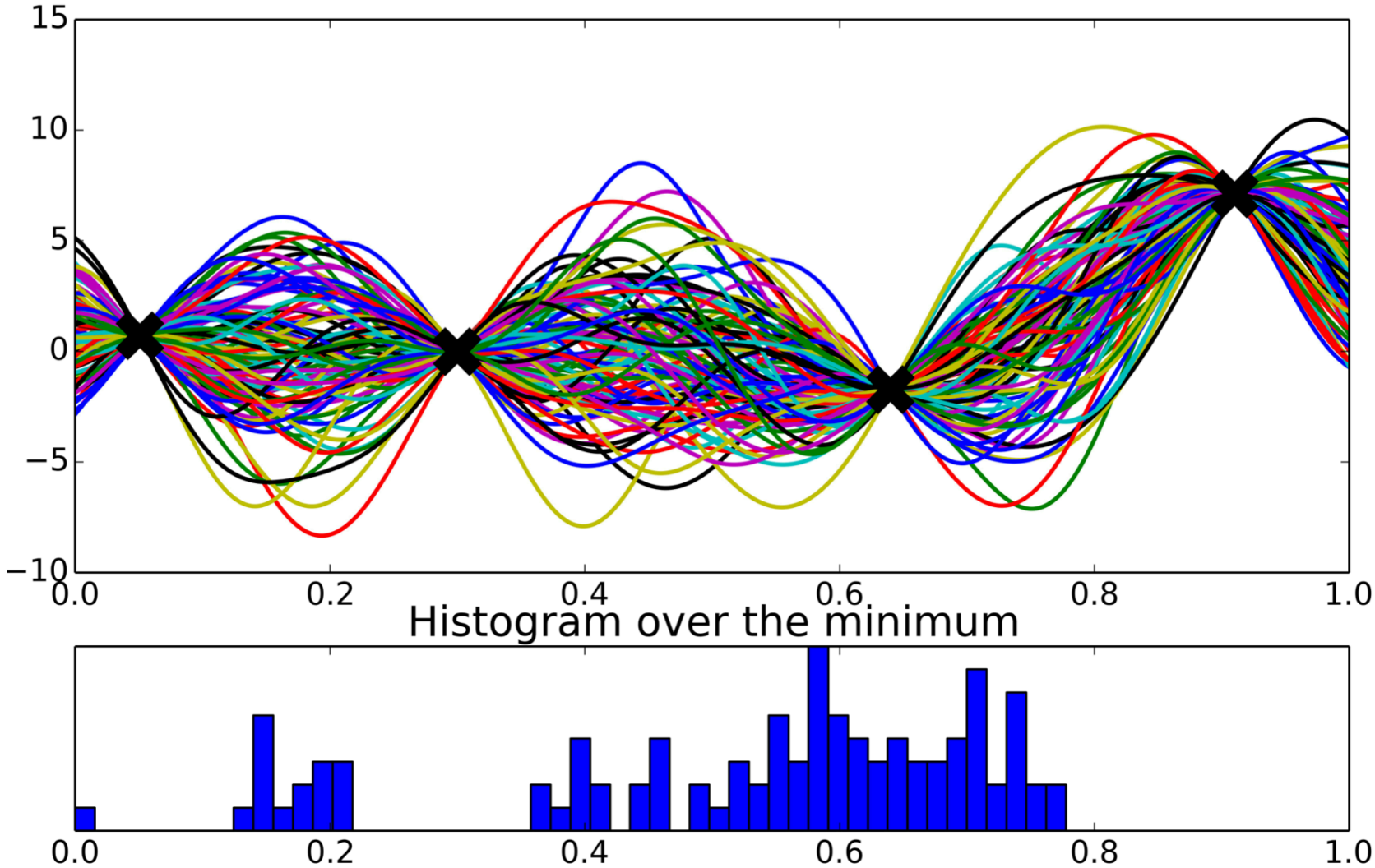
Intuition: three curves



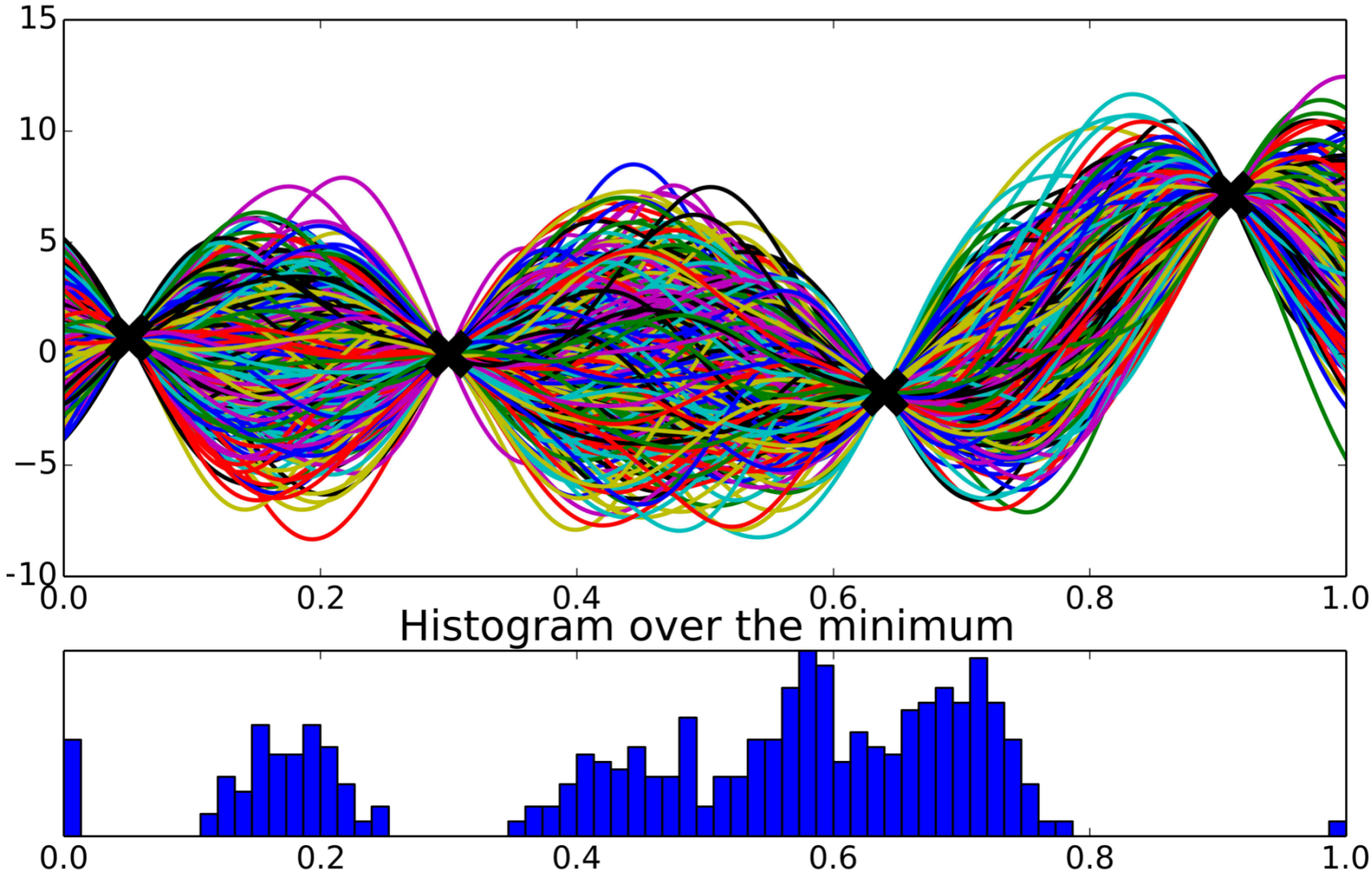
Intuition: ten curves



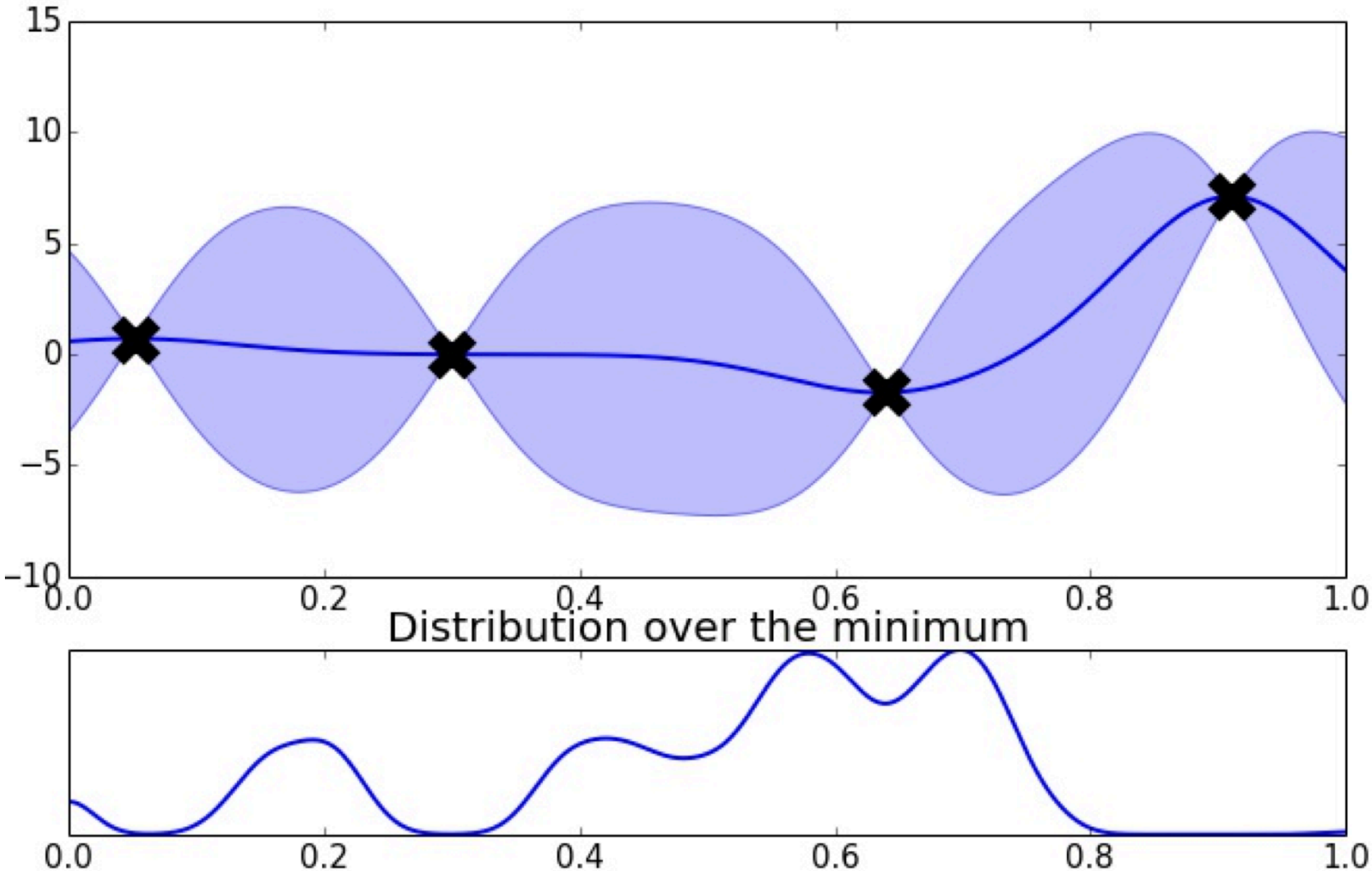
Intuition: hundred curves



Intuition: many curves



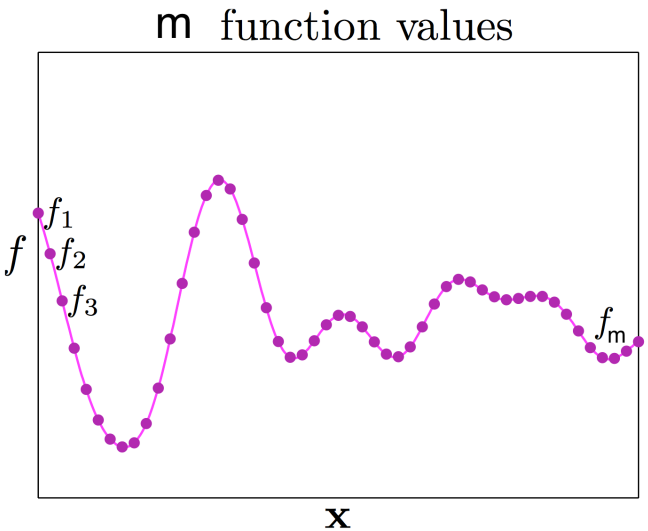
Intuition: infinite number of curves

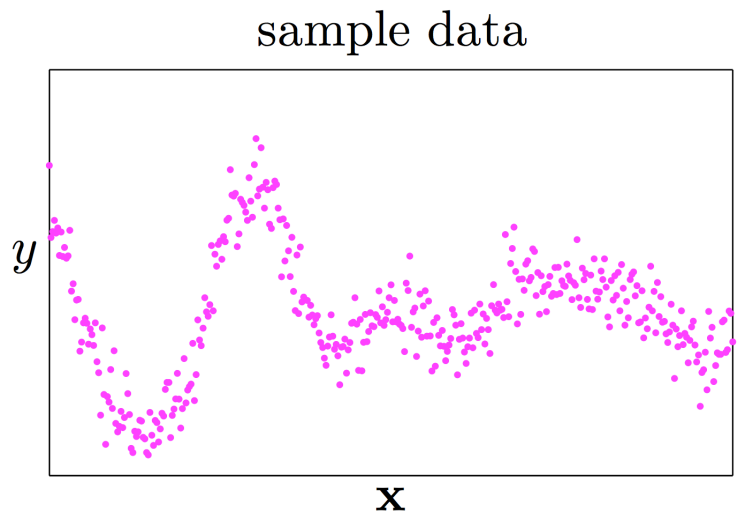


- We made prior assumption about f
- Information about the minimum is now encoded in a new function (the probability distribution p_{min} over the minimum in this case)
- We can use p_{min} (or a functional of it) to decide where to sample next
- Other functions to encode relevant information about the minimum are possible, e. g. the "marginal expected gain" at each location.

- Training data set $\mathrm{S_m} = \{\mathrm{X}, \mathrm{y}\} = \{(\mathrm{x_i}, \mathrm{y_i})\}_{\mathrm{i}=1}^{\mathrm{m}}$
- Model:

$$\mathrm{y_i} = \mathrm{f}(\mathrm{x_i}) + \varepsilon_\mathrm{i},$$
$$\mathrm{f} \sim \mathcal{GP}(\cdot|0, \mathrm{K}), \text{ with } \mathrm{K}(\mathrm{x}, \mathrm{x}') = \mathrm{cov}(\mathrm{f}(\mathrm{x}), \mathrm{f}(\mathrm{x}')),$$
$$\varepsilon_\mathrm{i} \sim \mathcal{N}(0, \sigma^2) \text{ is a white noise}$$





- Training data set $\mathrm{S_m} = \{\mathrm{X}, \mathrm{y}\} = \{(\mathrm{x_i}, \mathrm{y_i})\}_{\mathrm{i}=1}^{\mathrm{m}}$
- Model:

$$\begin{aligned} \mathrm{y_i} &= \mathrm{f}(\mathrm{x_i}) + \varepsilon_\mathrm{i} \\ \mathrm{f} &\sim \mathcal{GP}(\cdot|0, \mathrm{K}) \\ \varepsilon_\mathrm{i} &\sim \mathcal{N}(0, \sigma^2) \text{ is a white noise} \end{aligned}$$

- The prior is

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

- The noise model, or likelihood is

$$\mathrm{p}(\mathrm{y}|\mathrm{f}) = \mathcal{N}\left(\mathrm{y}|\mathrm{f}, \sigma^2 \mathrm{I_m}\right)$$

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

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

- Let us denote input test point as x_*, and output

$$\mathrm{y}_* = \mathrm{f}_* + \varepsilon_*, \ \mathrm{f}_* = \mathrm{f}(\mathrm{x}_*)$$

- Consider joint training and test marginal likelihood

$$\mathrm{p}(\mathrm{y}, \mathrm{f}_*) = \mathcal{N}\left(\begin{array}{c} \mathrm{y} \\ \mathrm{f}(\mathrm{x}_*) \end{array} \middle| 0, \left[\begin{array}{cc} \mathrm{K} + \sigma^2 \mathrm{I}_\mathrm{m} & \mathrm{k}_* \\ \mathrm{k}_*^\top & \mathrm{K}_{**} \end{array} \right] \right),$$

  where $\mathrm{k}_* = \{\mathrm{K}(\mathrm{x}_*, \mathrm{x}_\mathrm{i})\}_{\mathrm{i}=1}^{\mathrm{m}}$ and $\mathrm{K}_{**} = \mathrm{K}(\mathrm{x}_*, \mathrm{x}_*)$

- What we know about noiseless value $\mathrm{f}(\mathrm{x}_*)$?

- Joint training and test marginal likelihood

$$\mathrm{p}(\mathrm{y}, \mathrm{f}_*) = \mathcal{N}\left(\begin{array}{c} \mathrm{y} \\ \mathrm{f}(\mathrm{x}_*) \end{array} \middle| 0, \left[\begin{array}{cc} \mathrm{K} + \sigma^2 \mathrm{I}_\mathrm{m} & \mathrm{k}_* \\ \mathrm{k}_*^\top & \mathrm{K}_{**} \end{array} \right] \right),$$

  where $\mathrm{k}_* = \{\mathrm{K}(\mathrm{x}_*, \mathrm{x}_\mathrm{i})\}_{\mathrm{i}=1}^{\mathrm{m}}$ and $\mathrm{K}_{**} = \mathrm{K}(\mathrm{x}_*, \mathrm{x}_*)$

- Condition on training outputs y we get

$$\mathrm{p}(\mathrm{f}_*|\mathrm{y}) = \mathcal{N}\left(\mathrm{f}_* | \mu_*, \sigma_*^2\right),$$

  where

$$\begin{aligned} \mu_*(\mathrm{x}_*) &= \mathrm{k}_*^\top [\mathrm{K} + \sigma^2 \mathrm{I}_\mathrm{m}]^{-1} \mathrm{y} = \\ &= \sum_{\mathrm{i}=1}^{\mathrm{m}} \alpha_\mathrm{i} \mathrm{K}(\mathrm{x}_*, \mathrm{x}_\mathrm{i}) \text{ with } \boldsymbol{\alpha} = [\mathrm{K} + \sigma^2 \mathrm{I}_\mathrm{m}]^{-1} \mathrm{y} \quad \text{(aka KRR)} \\ \sigma_*^2(\mathrm{x}_*) &= \mathrm{K}_{**} - \mathrm{k}_*^\top [\mathrm{K} + \sigma^2 \mathrm{I}_\mathrm{m}]^{-1} \mathrm{k}_* \end{aligned}$$

- $\mathrm{p}(\mathrm{y}_*|\mathrm{y}) = \mathcal{N}\left(\mathrm{y}_* | \mu_*, \sigma_*^2 + \sigma^2\right)$ predicts what we'll see next

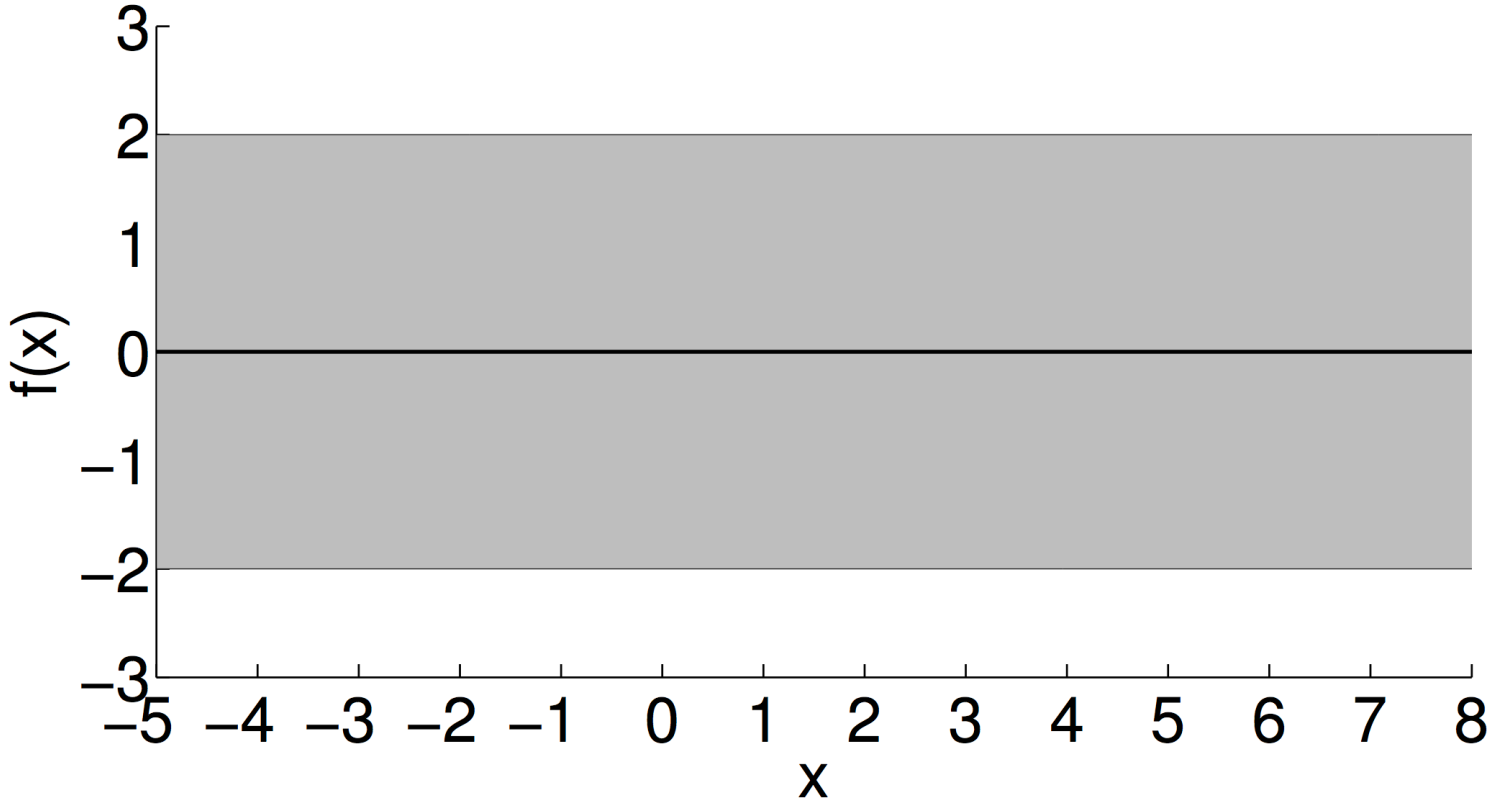


Figure – Prior belief about the function

Predictive (marginal) mean and variance

$$\mathbb{E}[\mathrm{f}(\mathrm{x}_*)|\mathrm{x}_*, \varnothing] = \mu_*(\mathrm{x}_*) = 0$$
$$\mathrm{Var}[\mathrm{f}(\mathrm{x}_*)|\mathrm{x}_*, \varnothing] = \sigma_*^2(\mathrm{x}_*) = \mathrm{K}(\mathrm{x}_*, \mathrm{x}_*)$$

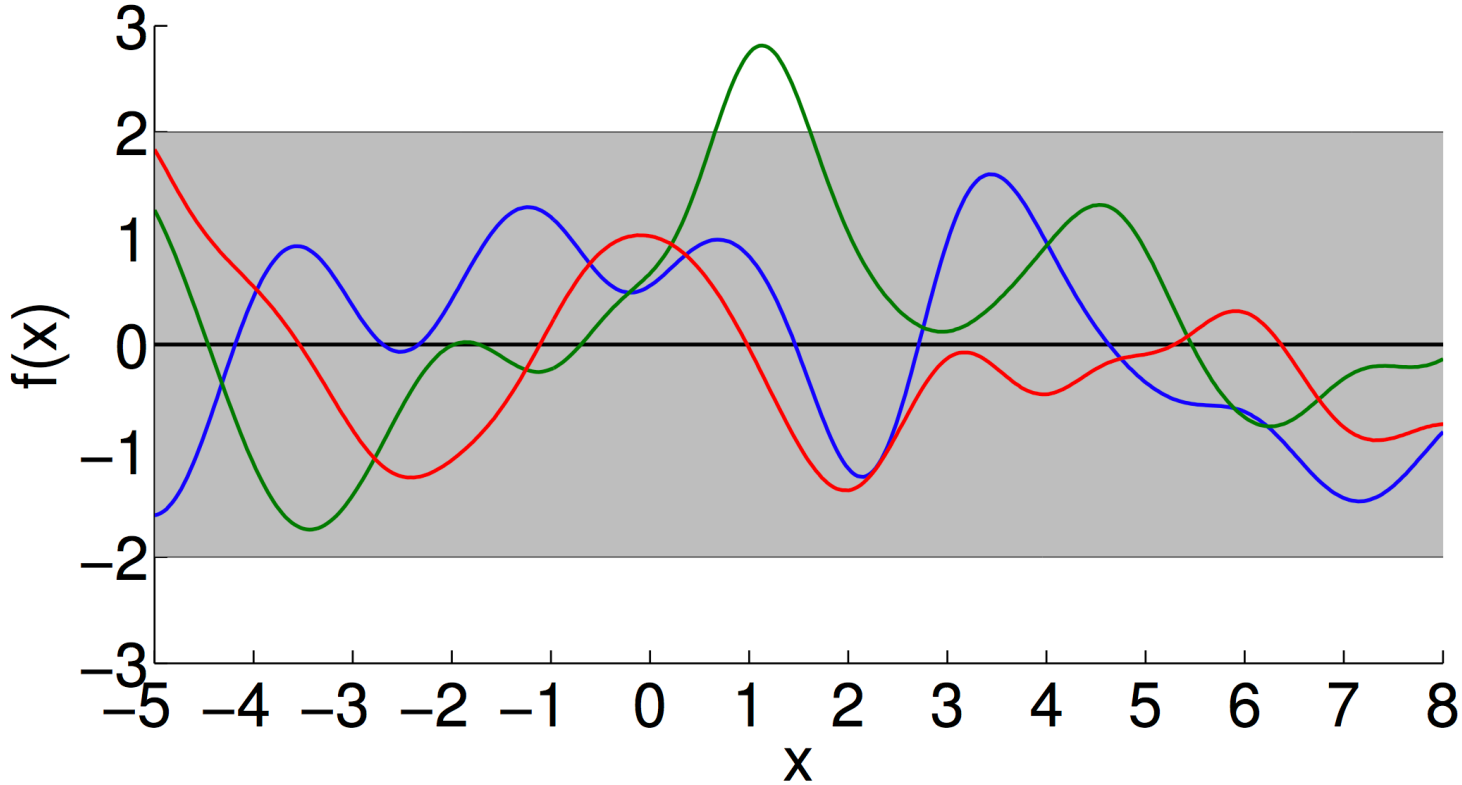


Figure – Prior belief about the function

Predictive (marginal) mean and variance

$$\mathbb{E}[\mathrm{f}(\mathrm{x}_*)|\mathrm{x}_*, \varnothing] = \mu_*(\mathrm{x}_*) = 0$$
$$\mathrm{Var}[\mathrm{f}(\mathrm{x}_*)|\mathrm{x}_*, \varnothing] = \sigma_*^2(\mathrm{x}_*) = \mathrm{K}(\mathrm{x}_*, \mathrm{x}_*)$$

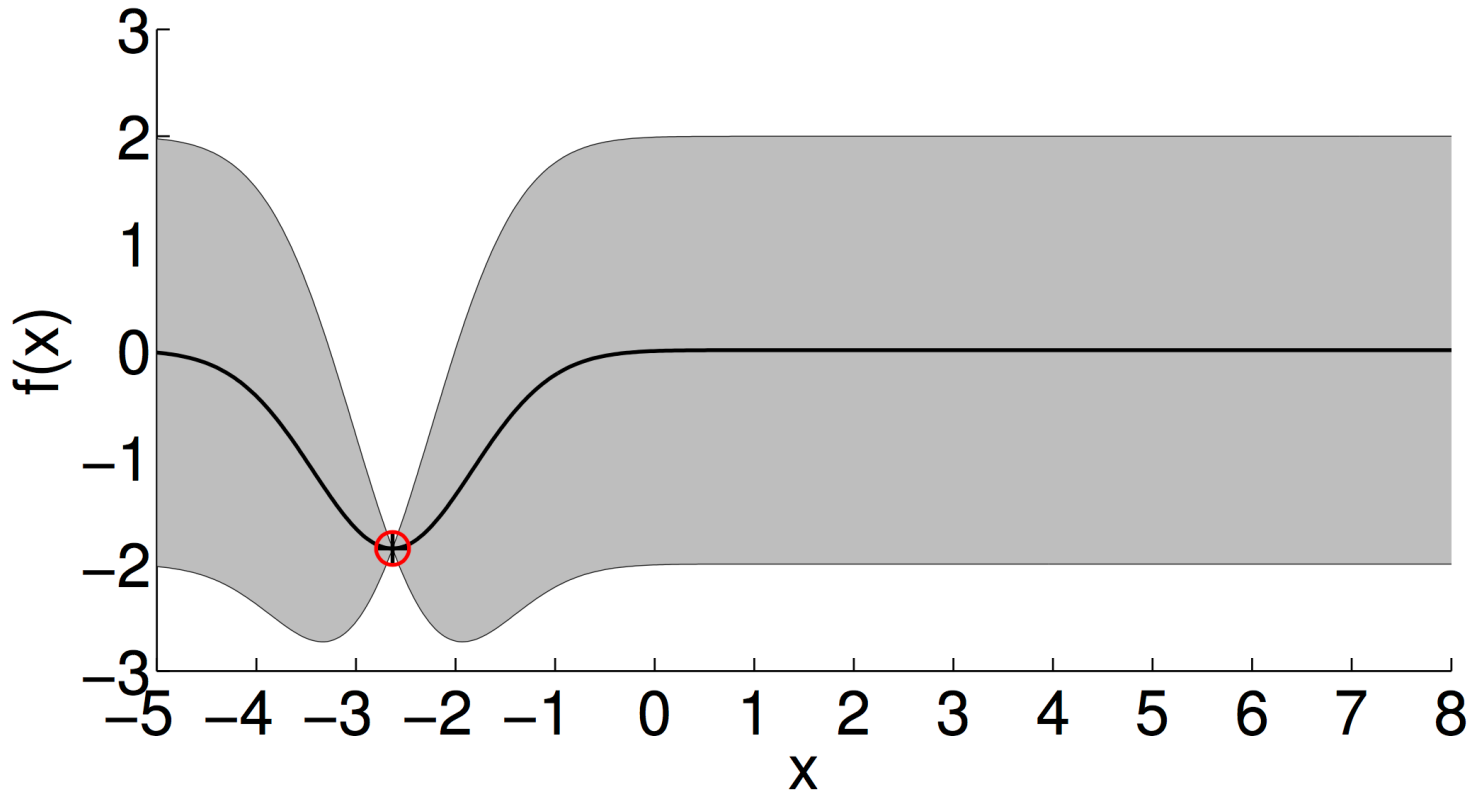


Figure – Posterior belief about the function

Predictive (marginal) mean and variance

$$\mathbb{E}[\mathrm{f}(\mathrm{x}_*)|\mathrm{x}_*, \mathrm{X}, \mathrm{y}] = \mu_*(\mathrm{x}_*) = \mathrm{k}_*^\top [\mathrm{K} + \sigma^2 \mathrm{I_m}]^{-1} \mathrm{y}$$

$$\mathrm{Var}[\mathrm{f}(\mathrm{x}_*)|\mathrm{x}_*, \mathrm{X}, \mathrm{y}] = \sigma_*^2(\mathrm{x}_*) = \mathrm{K}_{**} - \mathrm{k}_*^\top [\mathrm{K} + \sigma^2 \mathrm{I_m}]^{-1} \mathrm{k}_*$$

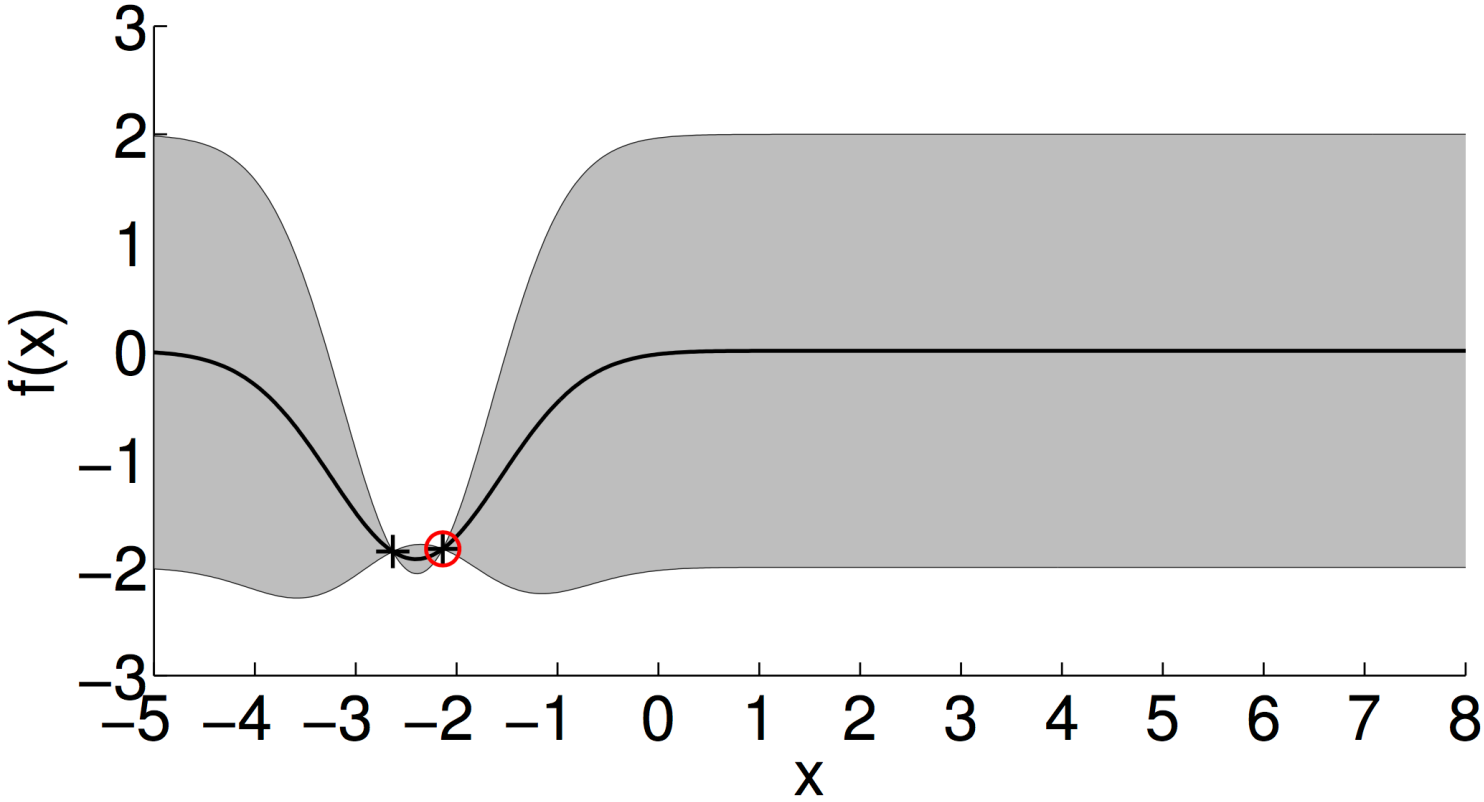


Figure – Posterior belief about the function

Predictive (marginal) mean and variance

$$\mathbb{E}[\mathrm{f}(\mathrm{x}_*)|\mathrm{x}_*, \mathrm{X}, \mathrm{y}] = \mu_*(\mathrm{x}_*) = \mathrm{k}_*^\top [\mathrm{K} + \sigma^2 \mathrm{I}_\mathrm{m}]^{-1} \mathrm{y}$$

$$\mathrm{Var}[\mathrm{f}(\mathrm{x}_*)|\mathrm{x}_*, \mathrm{X}, \mathrm{y}] = \sigma_*^2(\mathrm{x}_*) = \mathrm{K}_{**} - \mathrm{k}_*^\top [\mathrm{K} + \sigma^2 \mathrm{I}_\mathrm{m}]^{-1} \mathrm{k}_*$$

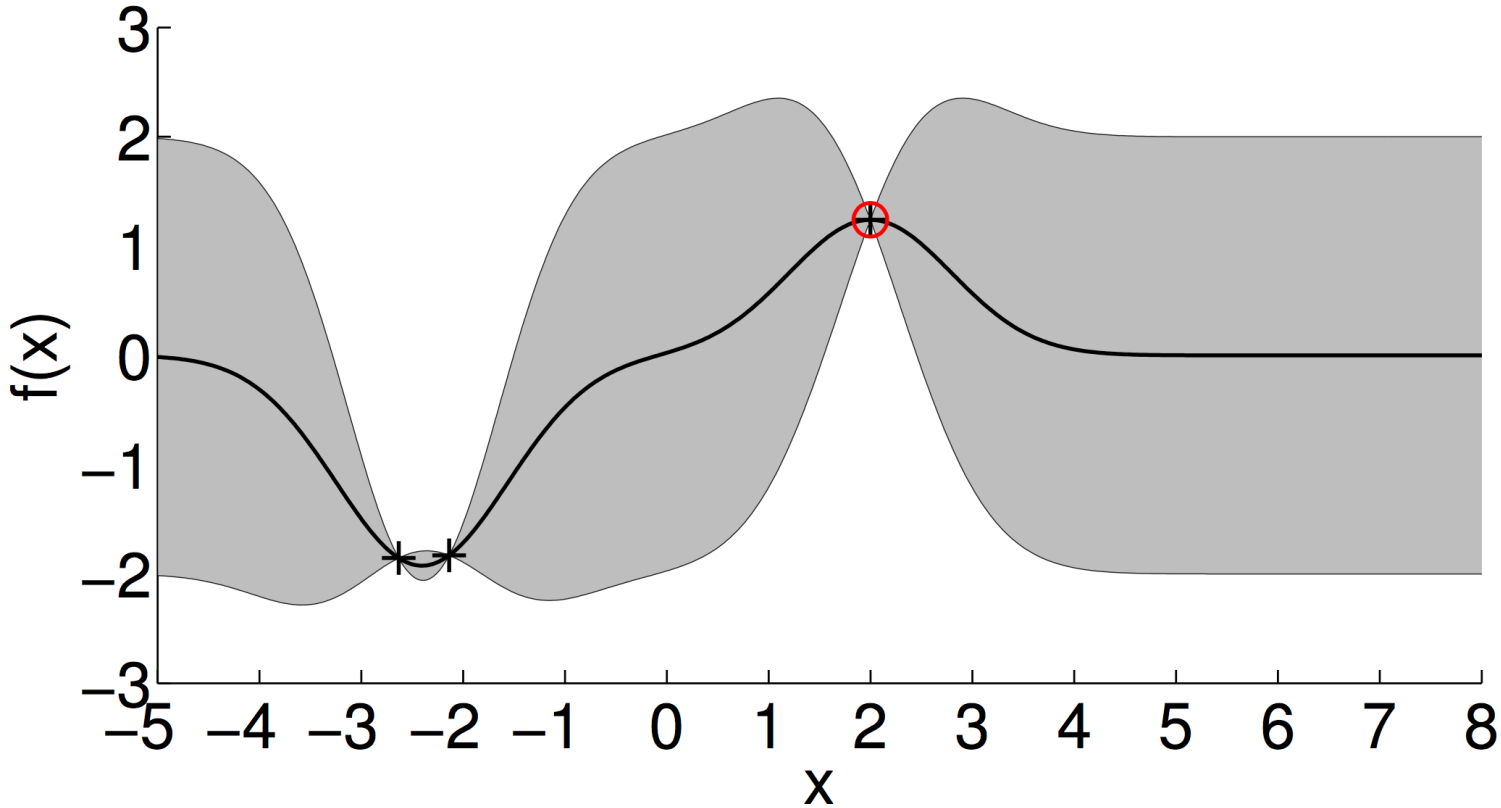


Figure – Posterior belief about the function

Predictive (marginal) mean and variance

$$\mathbb{E}[\mathrm{f}(\mathrm{x}_*)|\mathrm{x}_*, \mathrm{X}, \mathrm{y}] = \mu_*(\mathrm{x}_*) = \mathrm{k}_*^\top [\mathrm{K} + \sigma^2 \mathrm{I_m}]^{-1} \mathrm{y}$$

$$\mathrm{Var}[\mathrm{f}(\mathrm{x}_*)|\mathrm{x}_*, \mathrm{X}, \mathrm{y}] = \sigma_*^2(\mathrm{x}_*) = \mathrm{K}_{**} - \mathrm{k}_*^\top [\mathrm{K} + \sigma^2 \mathrm{I_m}]^{-1} \mathrm{k}_*$$

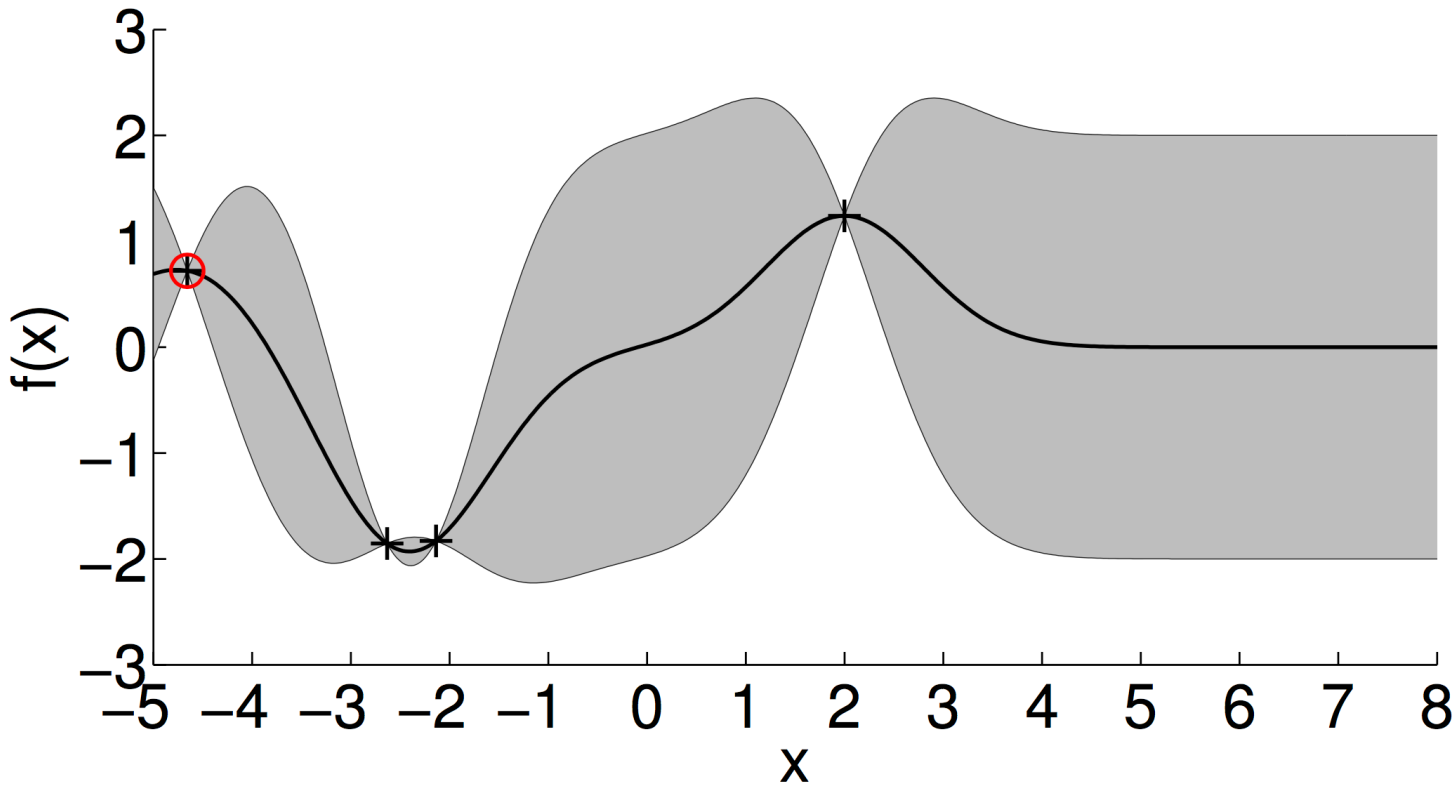


Figure – Posterior belief about the function

Predictive (marginal) mean and variance

$$\mathbb{E}[\mathrm{f}(\mathrm{x}_*)|\mathrm{x}_*, \mathrm{X}, \mathrm{y}] = \mu_*(\mathrm{x}_*) = \mathrm{k}_*^\top [\mathrm{K} + \sigma^2 \mathrm{I_m}]^{-1} \mathrm{y}$$

$$\mathrm{Var}[\mathrm{f}(\mathrm{x}_*)|\mathrm{x}_*, \mathrm{X}, \mathrm{y}] = \sigma_*^2(\mathrm{x}_*) = \mathrm{K}_{**} - \mathrm{k}_*^\top [\mathrm{K} + \sigma^2 \mathrm{I_m}]^{-1} \mathrm{k}_*$$

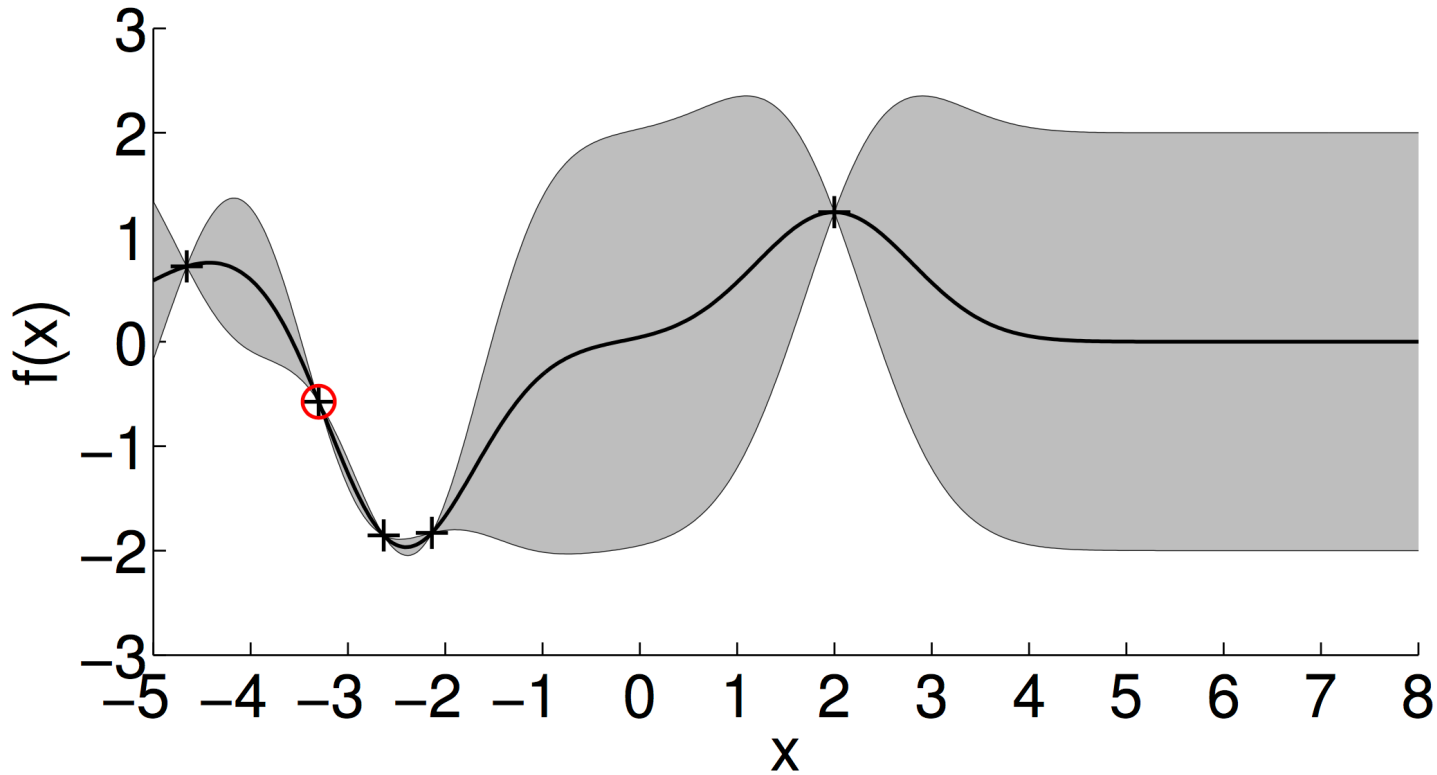


Figure – Posterior belief about the function

Predictive (marginal) mean and variance

$$\mathbb{E}[\mathrm{f}(\mathrm{x}_*)|\mathrm{x}_*, \mathrm{X}, \mathrm{y}] = \mu_*(\mathrm{x}_*) = \mathrm{k}_*^\top [\mathrm{K} + \sigma^2 \mathrm{I_m}]^{-1} \mathrm{y}$$

$$\mathrm{Var}[\mathrm{f}(\mathrm{x}_*)|\mathrm{x}_*, \mathrm{X}, \mathrm{y}] = \sigma_*^2(\mathrm{x}_*) = \mathrm{K}_{**} - \mathrm{k}_*^\top [\mathrm{K} + \sigma^2 \mathrm{I_m}]^{-1} \mathrm{k}_*$$

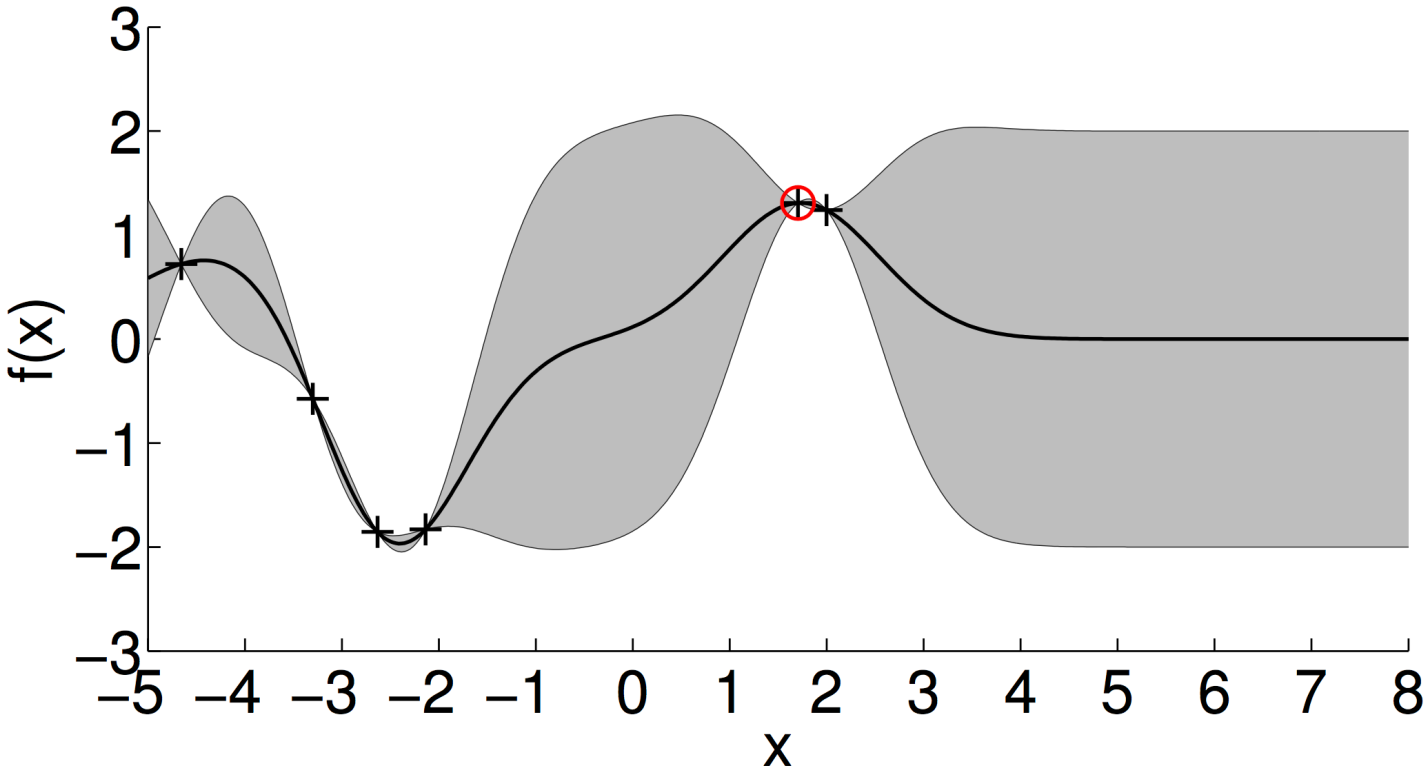


Figure – Posterior belief about the function

Predictive (marginal) mean and variance

$$\mathbb{E}[\mathrm{f}(\mathrm{x}_*)|\mathrm{x}_*, \mathrm{X}, \mathrm{y}] = \mu_*(\mathrm{x}_*) = \mathrm{k}_*^\top [\mathrm{K} + \sigma^2 \mathrm{I_m}]^{-1} \mathrm{y}$$

$$\mathrm{Var}[\mathrm{f}(\mathrm{x}_*)|\mathrm{x}_*, \mathrm{X}, \mathrm{y}] = \sigma_*^2(\mathrm{x}_*) = \mathrm{K}_{**} - \mathrm{k}_*^\top [\mathrm{K} + \sigma^2 \mathrm{I_m}]^{-1} \mathrm{k}_*$$

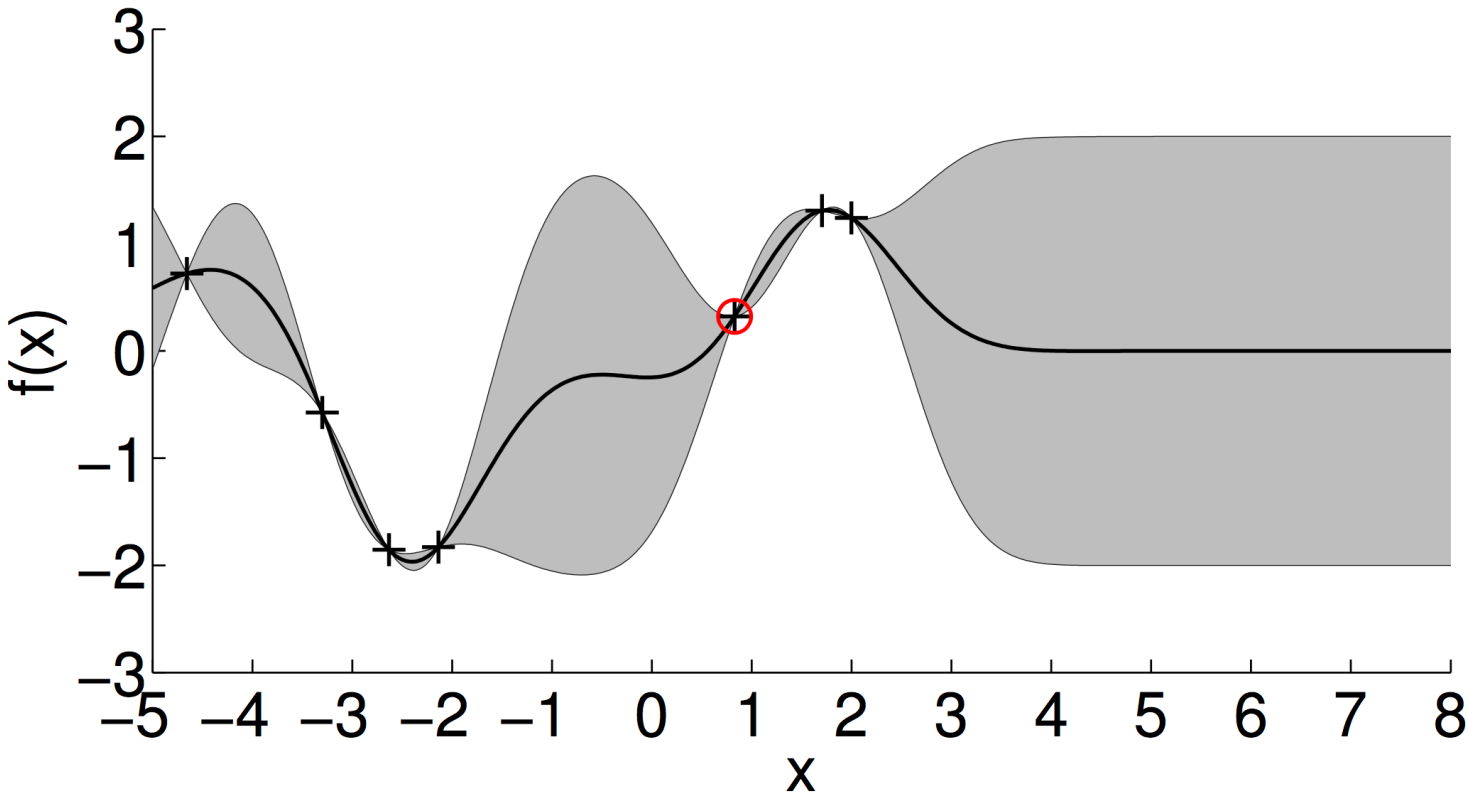


Figure – Posterior belief about the function

Predictive (marginal) mean and variance

$$\mathbb{E}[\mathrm{f}(\mathrm{x}_*)|\mathrm{x}_*, \mathrm{X}, \mathrm{y}] = \mu_*(\mathrm{x}_*) = \mathrm{k}_*^\top [\mathrm{K} + \sigma^2 \mathrm{I_m}]^{-1} \mathrm{y}$$

$$\mathrm{Var}[\mathrm{f}(\mathrm{x}_*)|\mathrm{x}_*, \mathrm{X}, \mathrm{y}] = \sigma_*^2(\mathrm{x}_*) = \mathrm{K}_{**} - \mathrm{k}_*^\top [\mathrm{K} + \sigma^2 \mathrm{I_m}]^{-1} \mathrm{k}_*$$

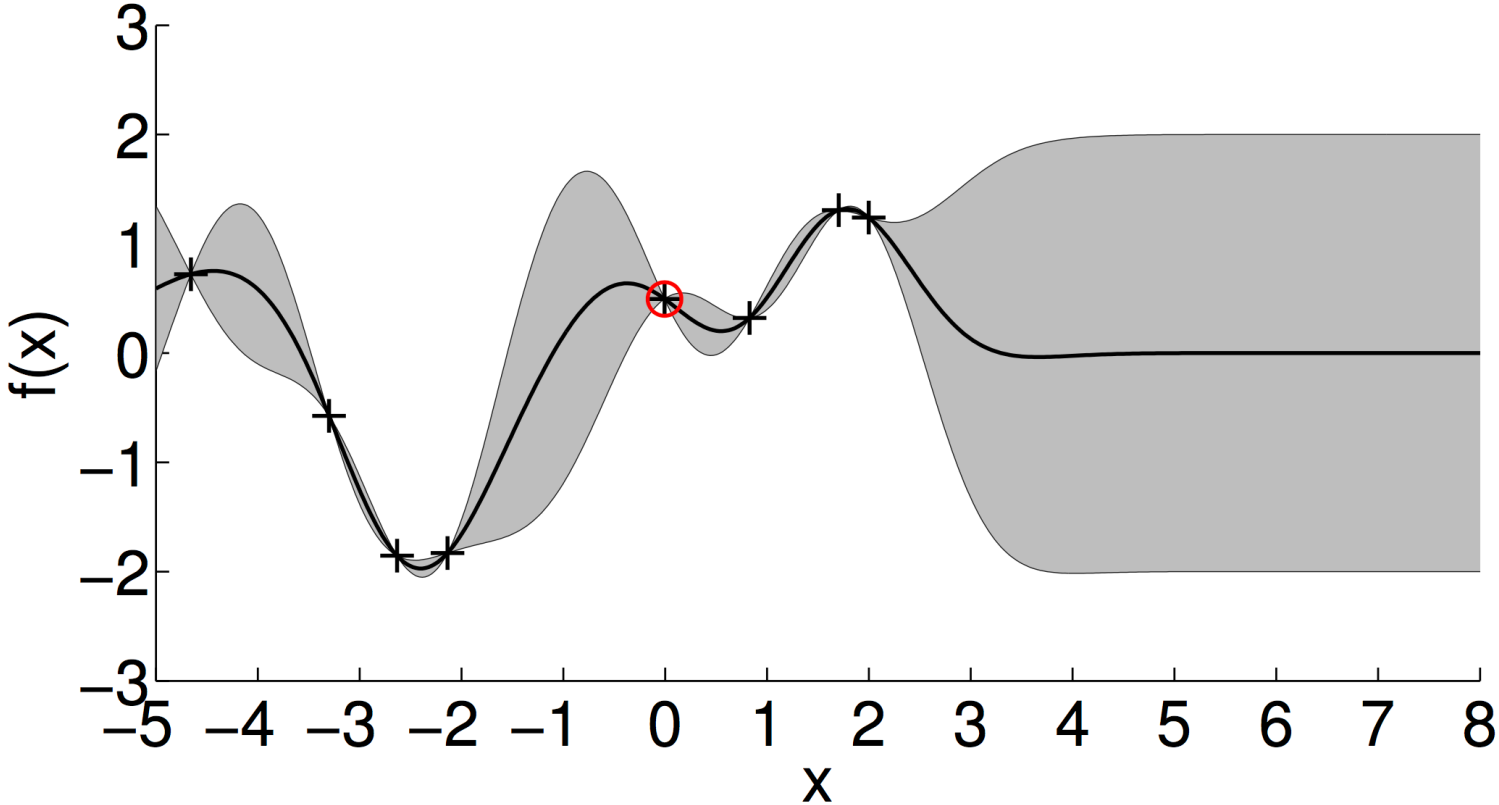


Figure – Posterior belief about the function

Predictive (marginal) mean and variance

$$\mathbb{E}[\mathrm{f}(\mathrm{x}_*)|\mathrm{x}_*, \mathrm{X}, \mathrm{y}] = \mu_*(\mathrm{x}_*) = \mathrm{k}_*^\top [\mathrm{K} + \sigma^2 \mathrm{I_m}]^{-1} \mathrm{y}$$

$$\mathrm{Var}[\mathrm{f}(\mathrm{x}_*)|\mathrm{x}_*, \mathrm{X}, \mathrm{y}] = \sigma_*^2(\mathrm{x}_*) = \mathrm{K}_{**} - \mathrm{k}_*^\top [\mathrm{K} + \sigma^2 \mathrm{I_m}]^{-1} \mathrm{k}_*$$

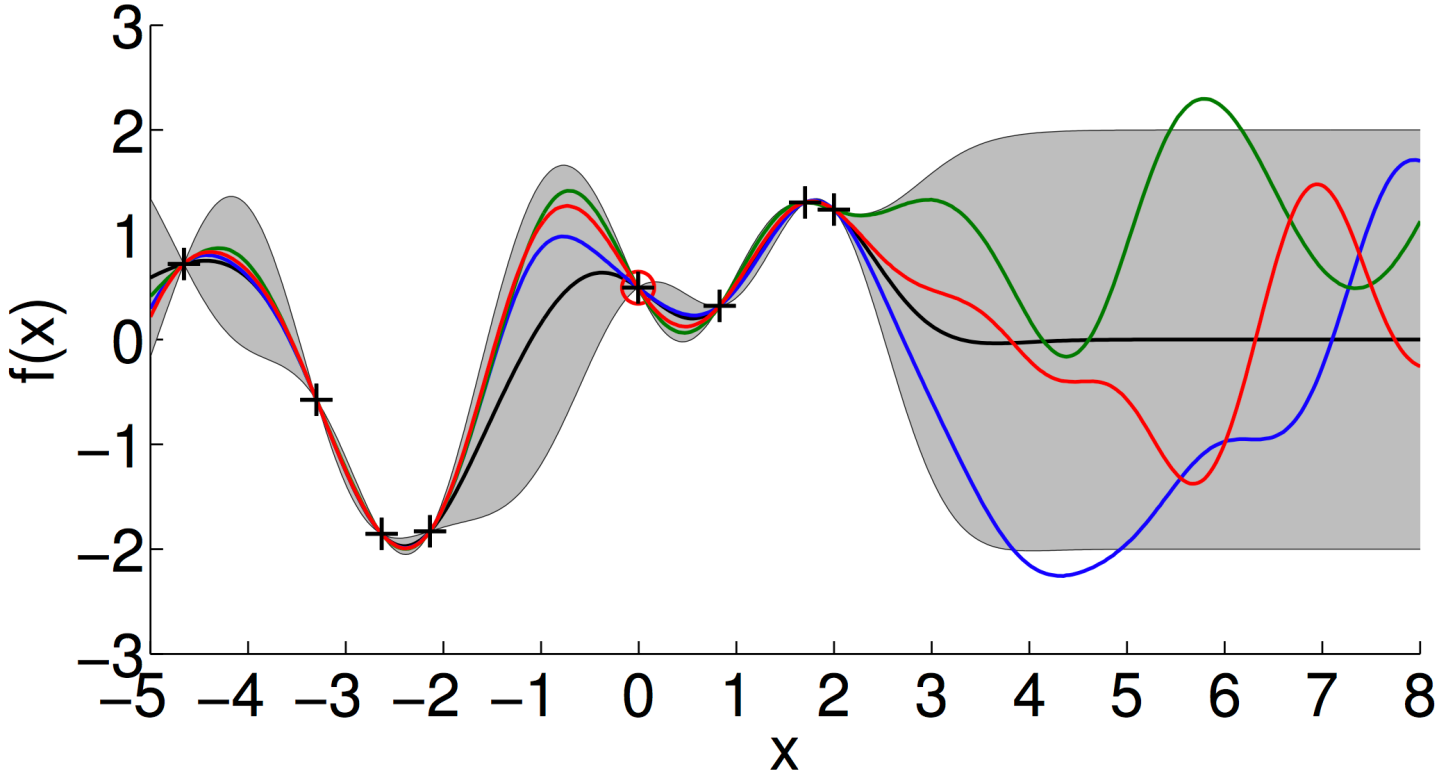


Figure – Posterior belief about the function

Predictive (marginal) mean and variance

$$\mathbb{E}[\mathrm{f}(\mathrm{x}_*)|\mathrm{x}_*, \mathrm{X}, \mathrm{y}] = \mu_*(\mathrm{x}_*) = \mathrm{k}_*^\top [\mathrm{K} + \sigma^2 \mathrm{I_m}]^{-1} \mathrm{y}$$

$$\mathrm{Var}[\mathrm{f}(\mathrm{x}_*)|\mathrm{x}_*, \mathrm{X}, \mathrm{y}] = \sigma_*^2(\mathrm{x}_*) = \mathrm{K}_{**} - \mathrm{k}_*^\top [\mathrm{K} + \sigma^2 \mathrm{I_m}]^{-1} \mathrm{k}_*$$

Methodology to perform global optimization of multimodal black-box functions

1. Choose some prior measure over the space of possible objectives f
2. Combine prior and the likelihood to get a posterior over the objective given some observations
3. Use the posterior to decide where to take the next evaluation according to some acquisition function
4. Augment the data set
5. Iterate between 2 and 4 until the evaluation budget is over

Comment: BO can be theoretically formalized in the framework of dynamic programming principle

- Use GP $\mathcal{GP}(\cdot|\mu(\mathrm{x}), \mathrm{K}(\mathrm{x}, \mathrm{x}'))$ as a prior for $\mathrm{f}(\cdot)$
- GP has marginal closed-form for the posterior mean $\mu_*(\mathrm{x})$ and variance $\sigma_*^2(\mathrm{x}) \Rightarrow$ efficient calculation of acquisition function
  - Exploration: Evaluate in places where the variance is large
  - Exploitation: Evaluate in places where the mean is low

Acquisition functions balance these two factors to determine where to evaluate next

- BO is an strategy to transform the problem

$$x_{min} = \arg\min_{x \in \mathcal{X}} f(x)$$

unsolvable!

into a series of problems

$$x_{t+1} = \arg\max_{x \in \mathcal{X}} \alpha(x|S_t),$$

solvable!

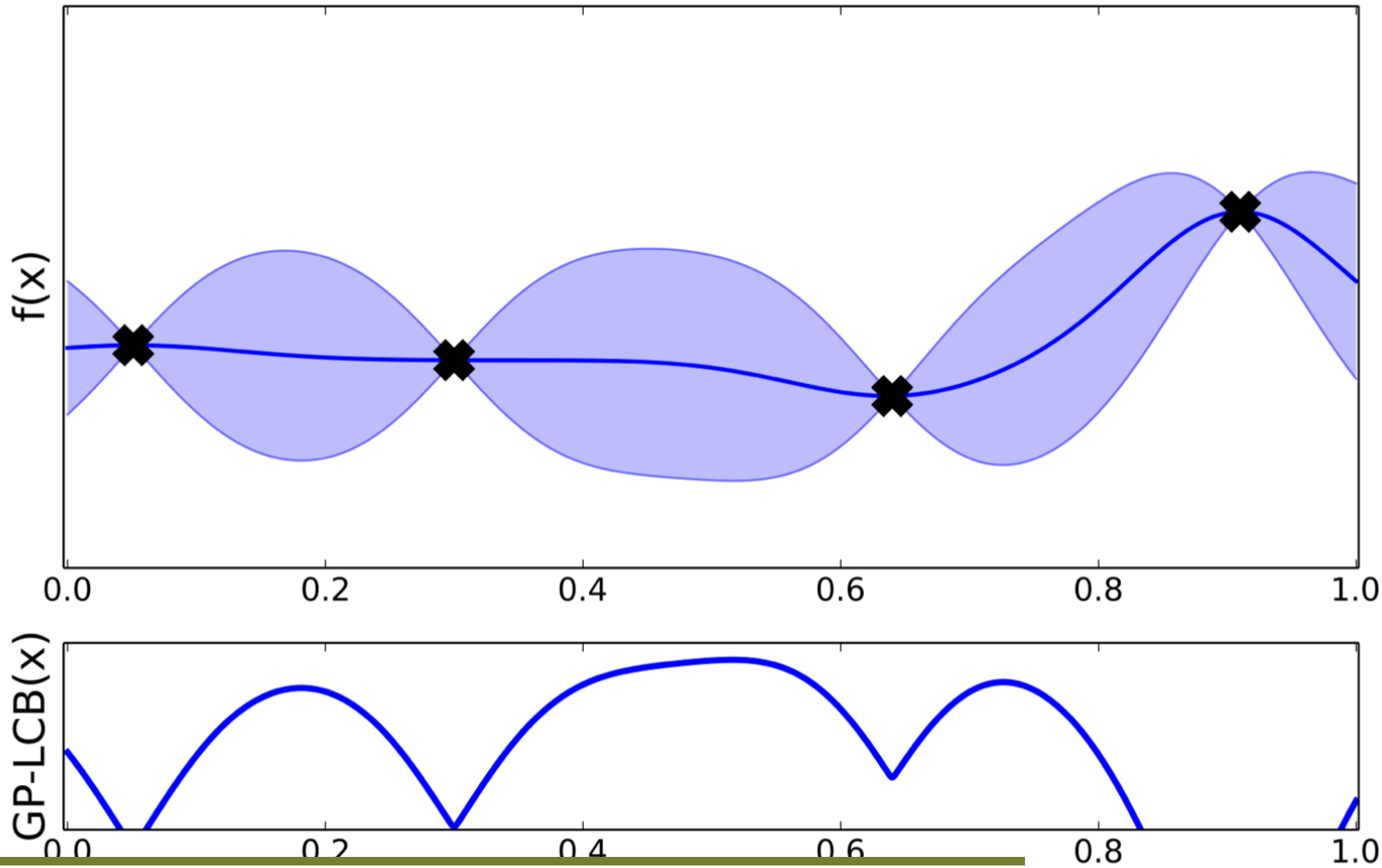
where

  - $\alpha(x)$ is not so expensive to evaluate
  - Gradients of $\alpha(x)$ are typically available
  - Still need to find x_{t+1}: DIRECT, gradient methods, SA

GP Upper (lower) Confidence Band

Direct balance between exploration and exploitation (ζ is a user-defined parameter):

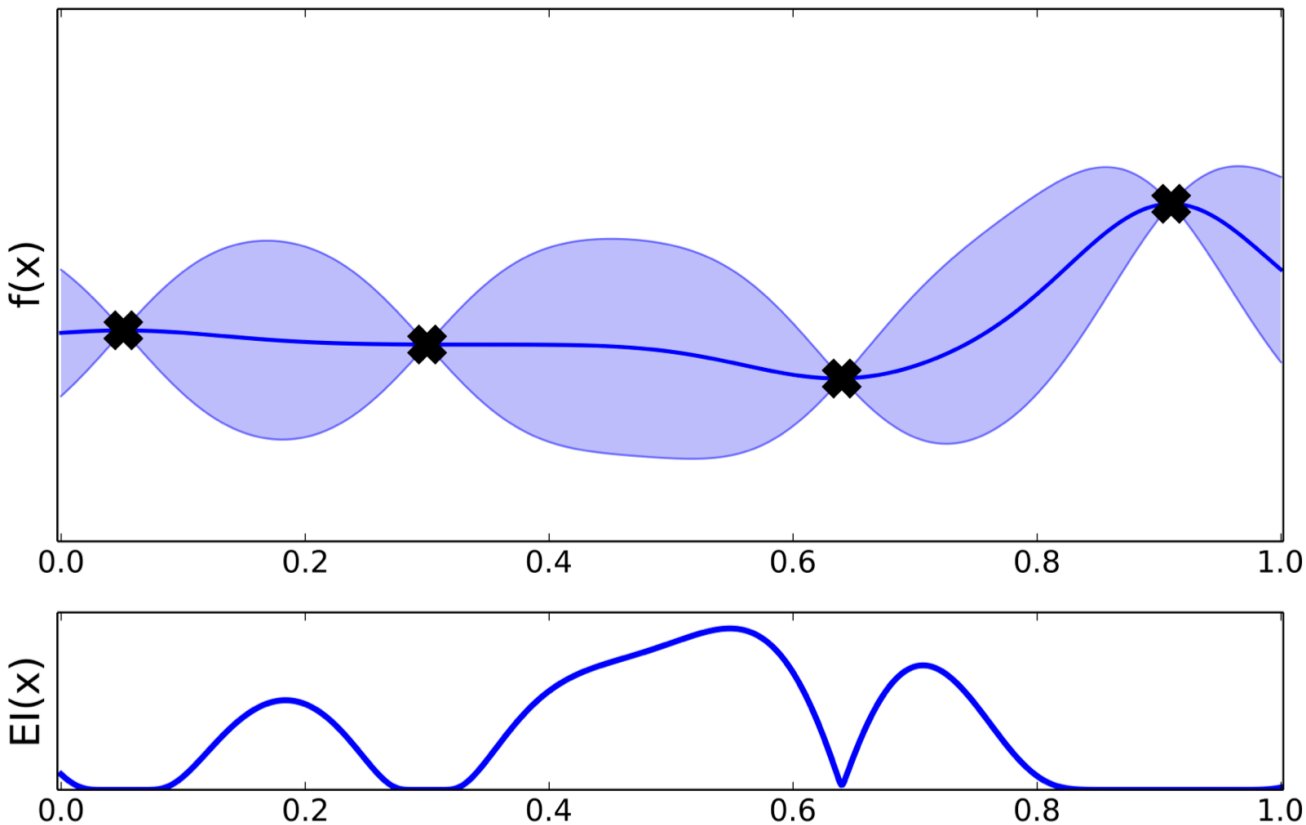
$$\alpha_{\mathrm{LCB}}(\mathrm{x}) = -\mu_*(\mathrm{x}) + \zeta \cdot \sigma_*(\mathrm{x})$$



Let us denote by $\Delta(\mathrm{x}) = \mathrm{y}_{\mathrm{best}} - \mu_*(\mathrm{x})$, then

$$\alpha_{\mathrm{EI}}(\mathrm{x}) = \int_{\mathrm{y}} \max(0, \mathrm{y}_{\mathrm{best}} - \mathrm{y}_*)\mathrm{p}(\mathrm{y}_*|\mathrm{x})\mathrm{dy}_* =$$

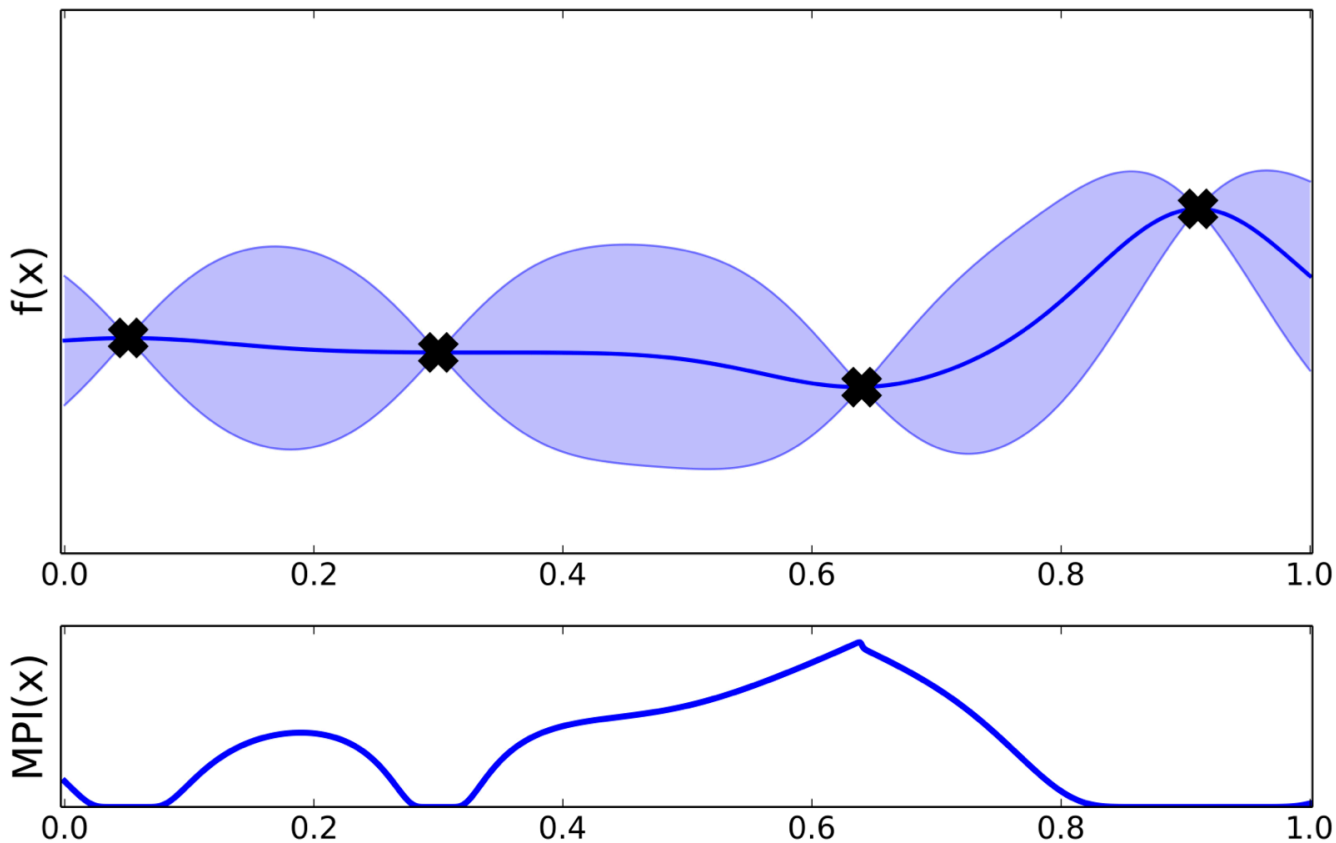
$$= \max(0, \Delta(\mathrm{x})) - \sigma_*(\mathrm{x})\varphi\left(\frac{\Delta(\mathrm{x})}{\sigma_*(\mathrm{x})}\right) + |\Delta(\mathrm{x})|\Phi\left(-\frac{|\Delta(\mathrm{x})|}{\sigma_*(\mathrm{x})}\right)$$



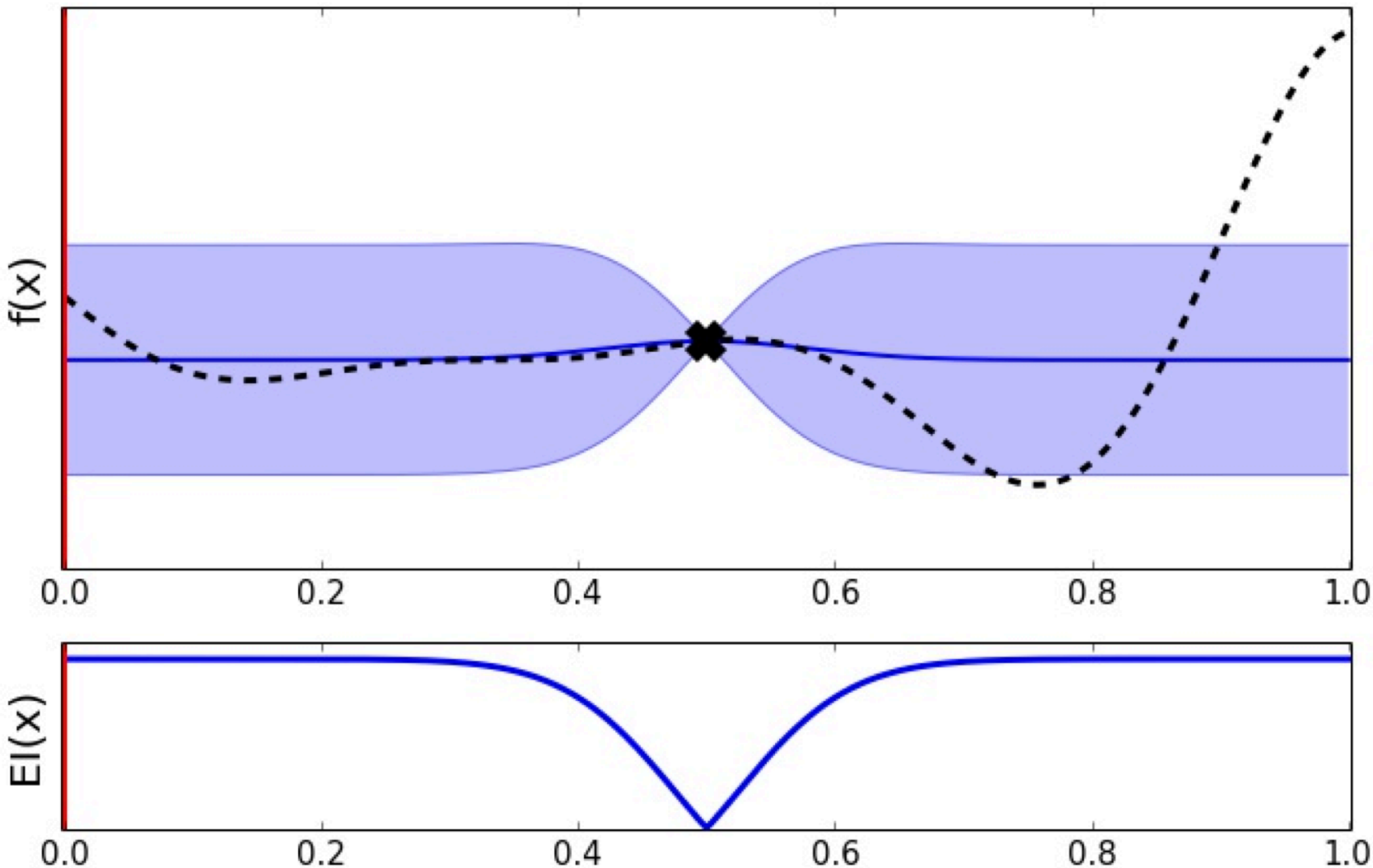
Maximum Probability of Improvement

$$\gamma(\mathrm{x}) = \frac{\mu(\mathrm{x}) - \mathrm{y}_{\mathrm{best}}}{\sigma(\mathrm{x})}$$

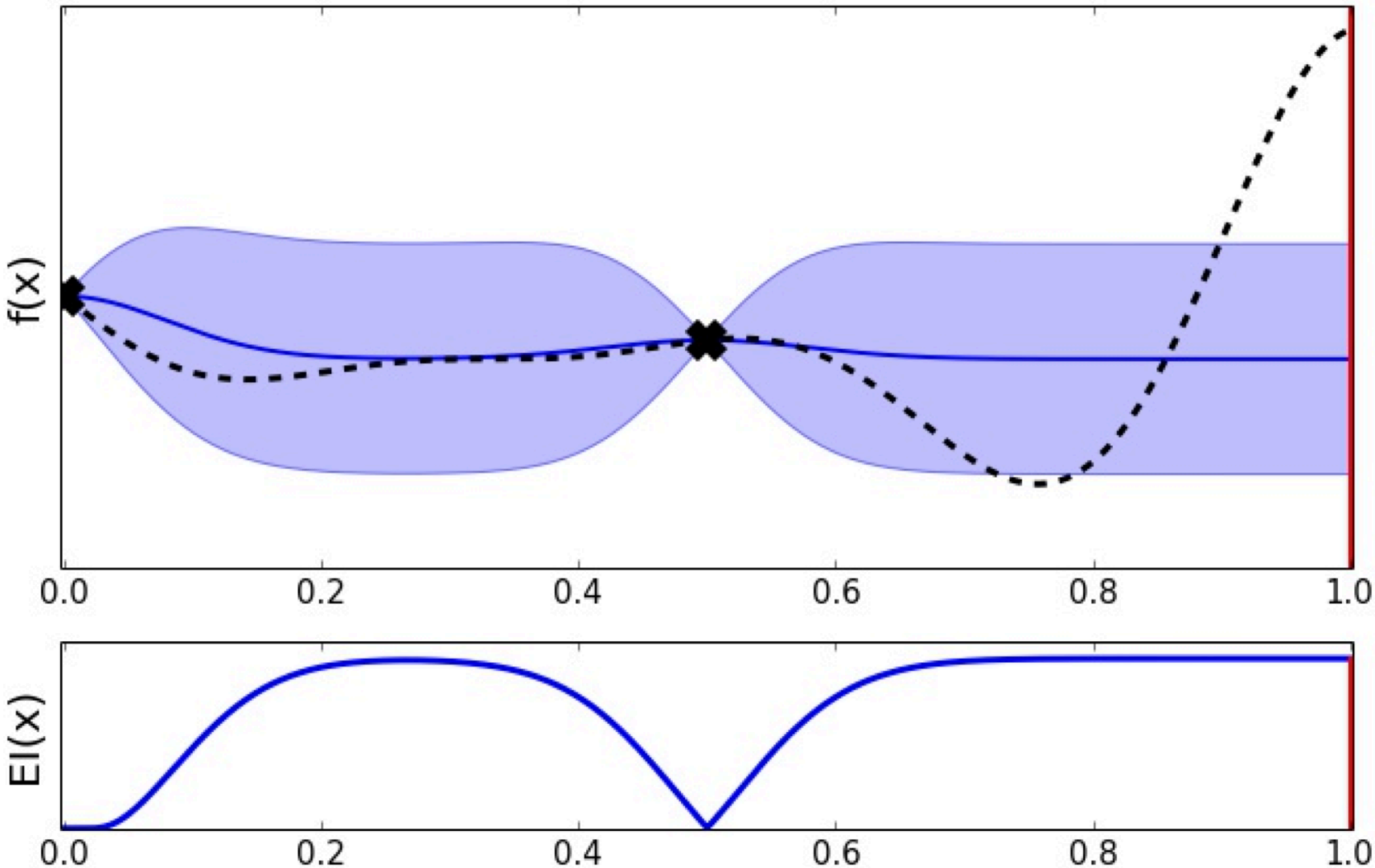
$$\alpha_{\mathrm{MPI}}(\mathrm{x}) = \mathbb{P}(\mathrm{f}(\mathrm{x}) < \mathrm{y}_{\mathrm{best}}) = \Phi(\gamma(\mathrm{x}))$$



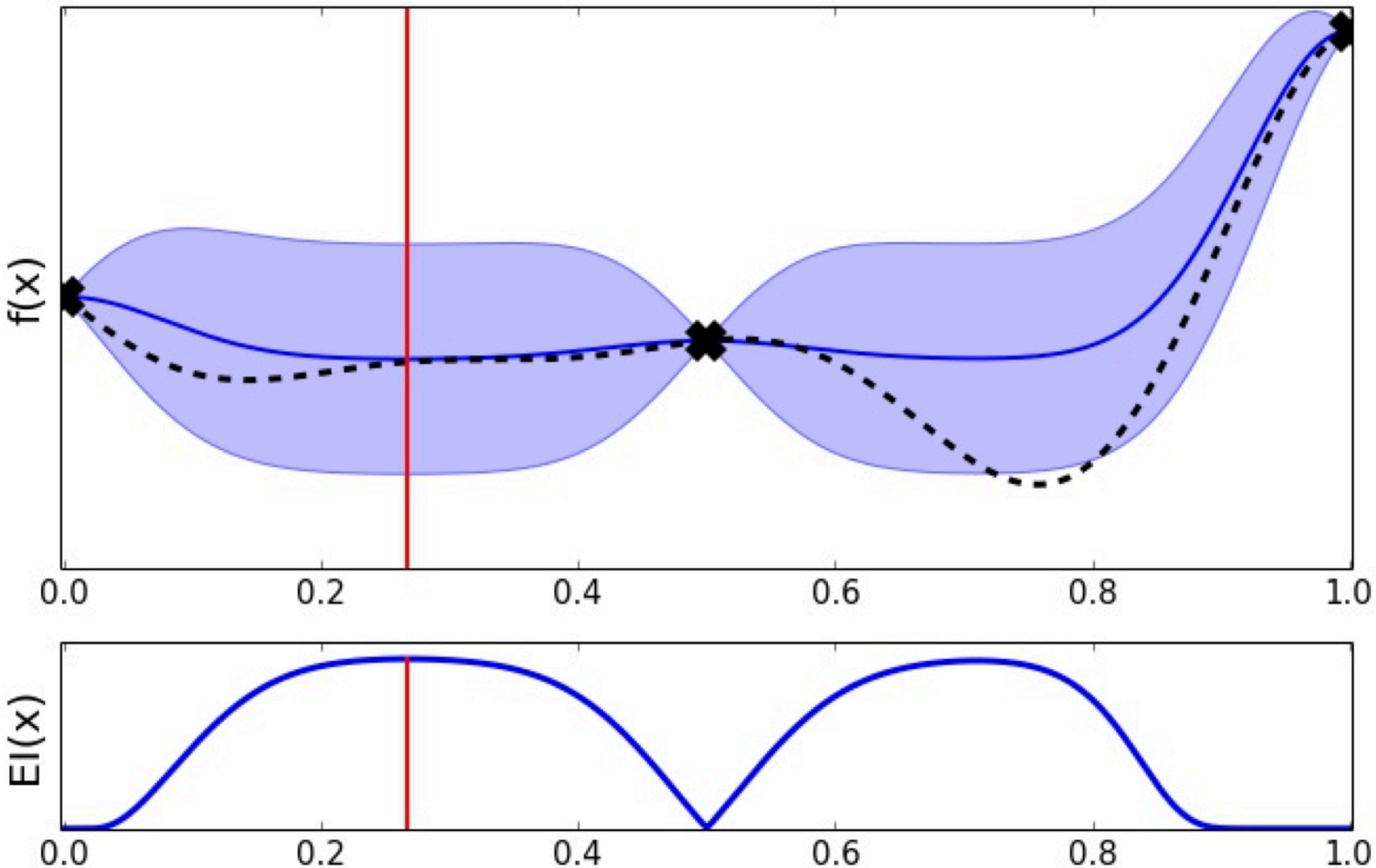
Expected Improvement: Toy Problem



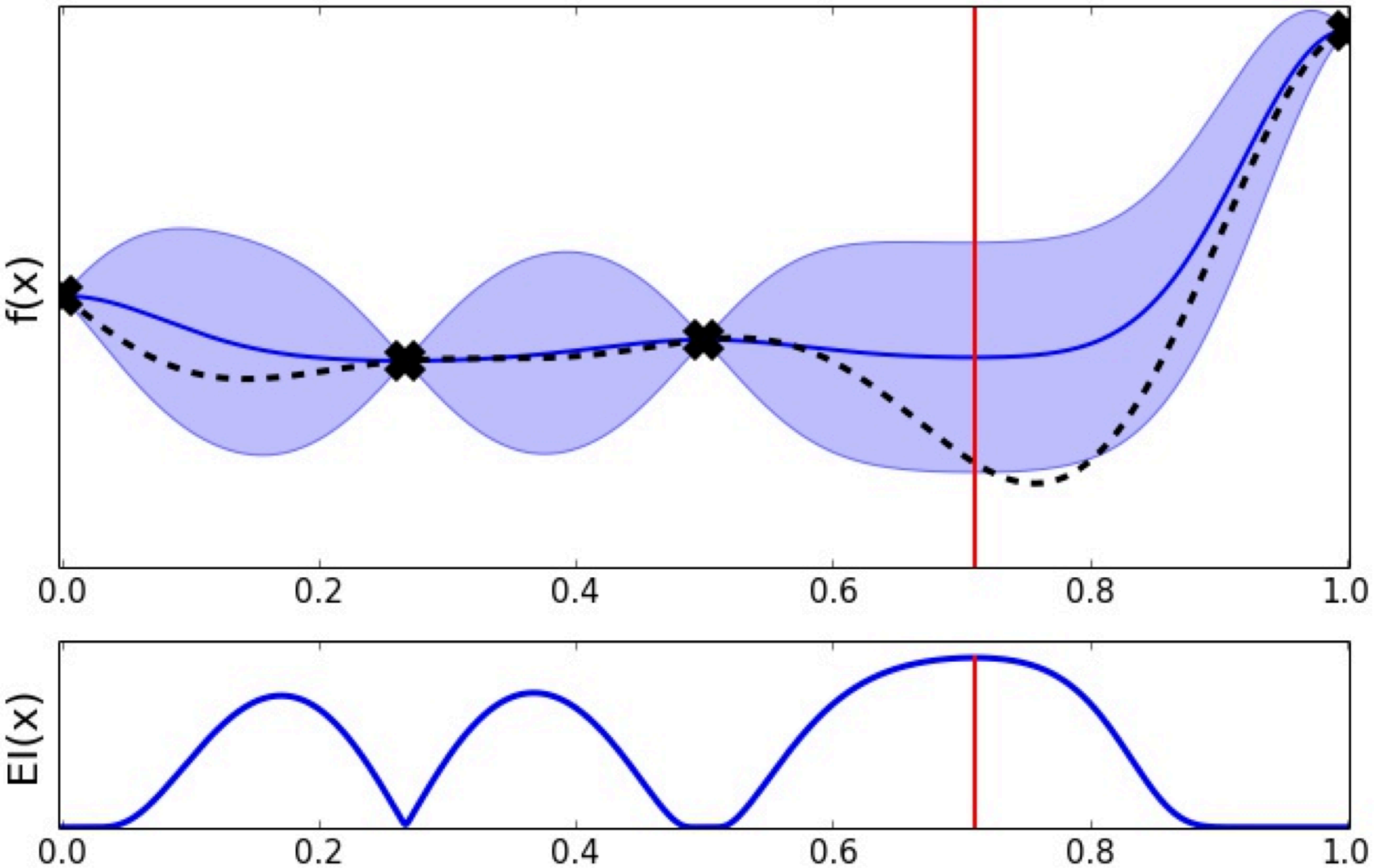
Expected Improvement: Toy Problem



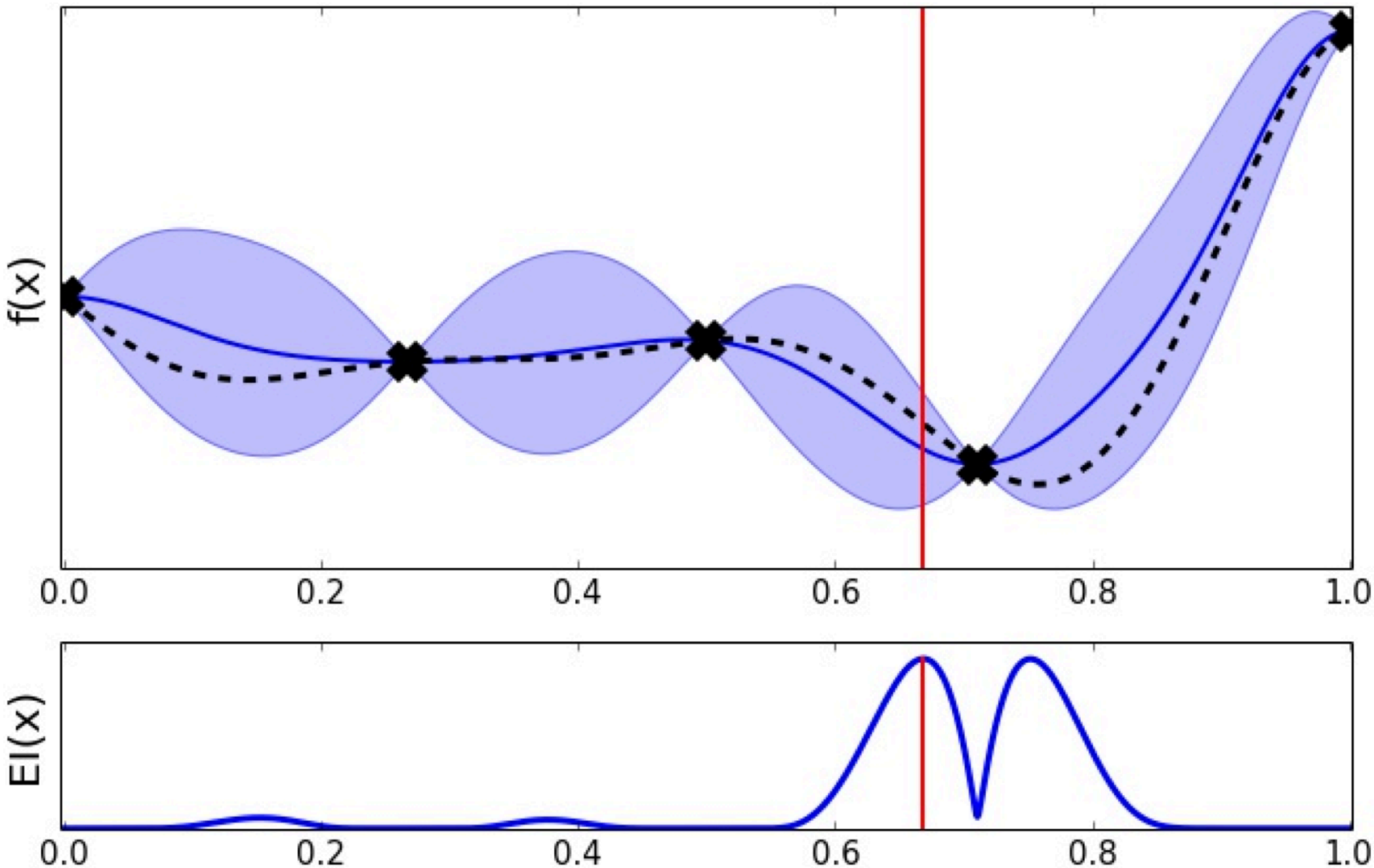
Expected Improvement: Toy Problem



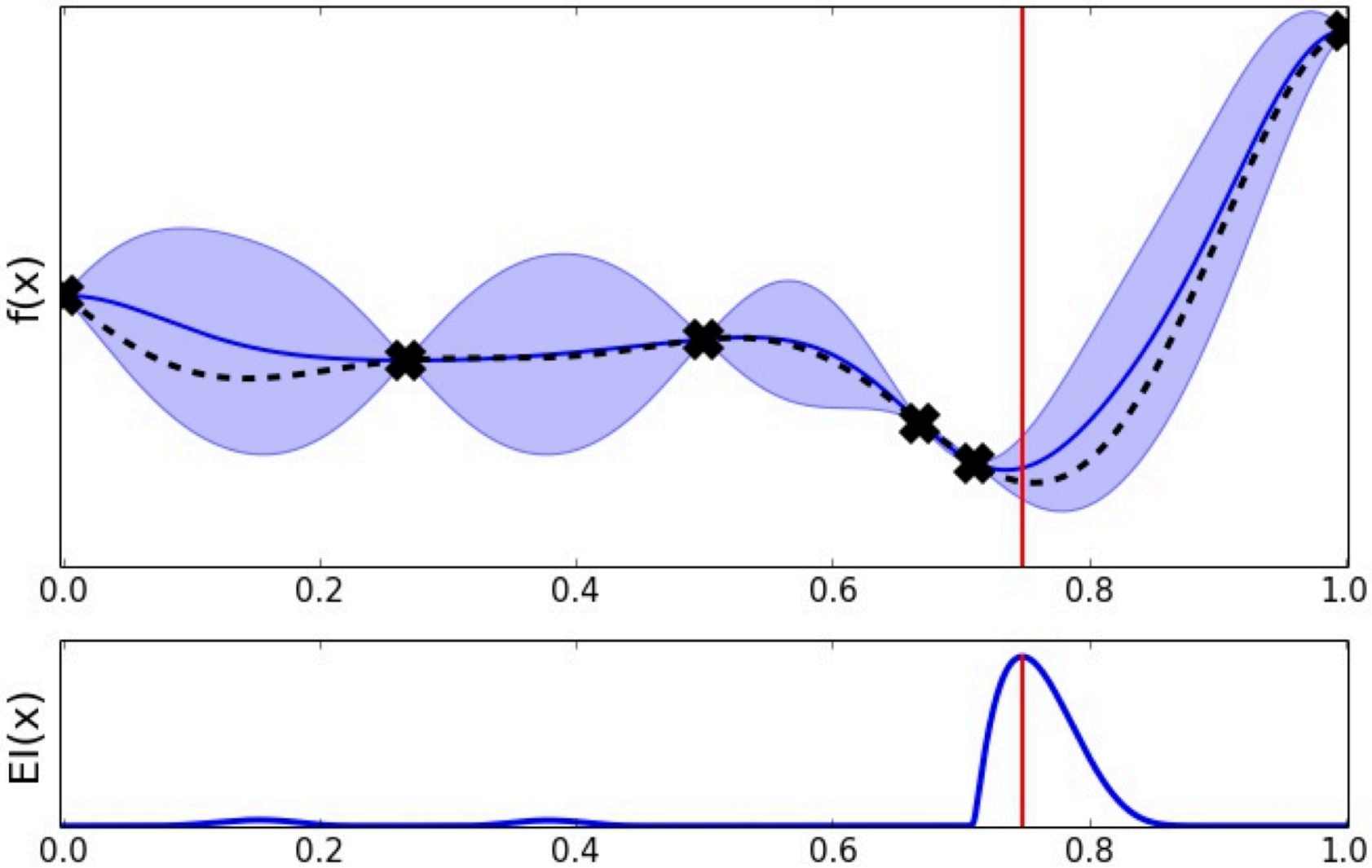
Expected Improvement: Toy Problem



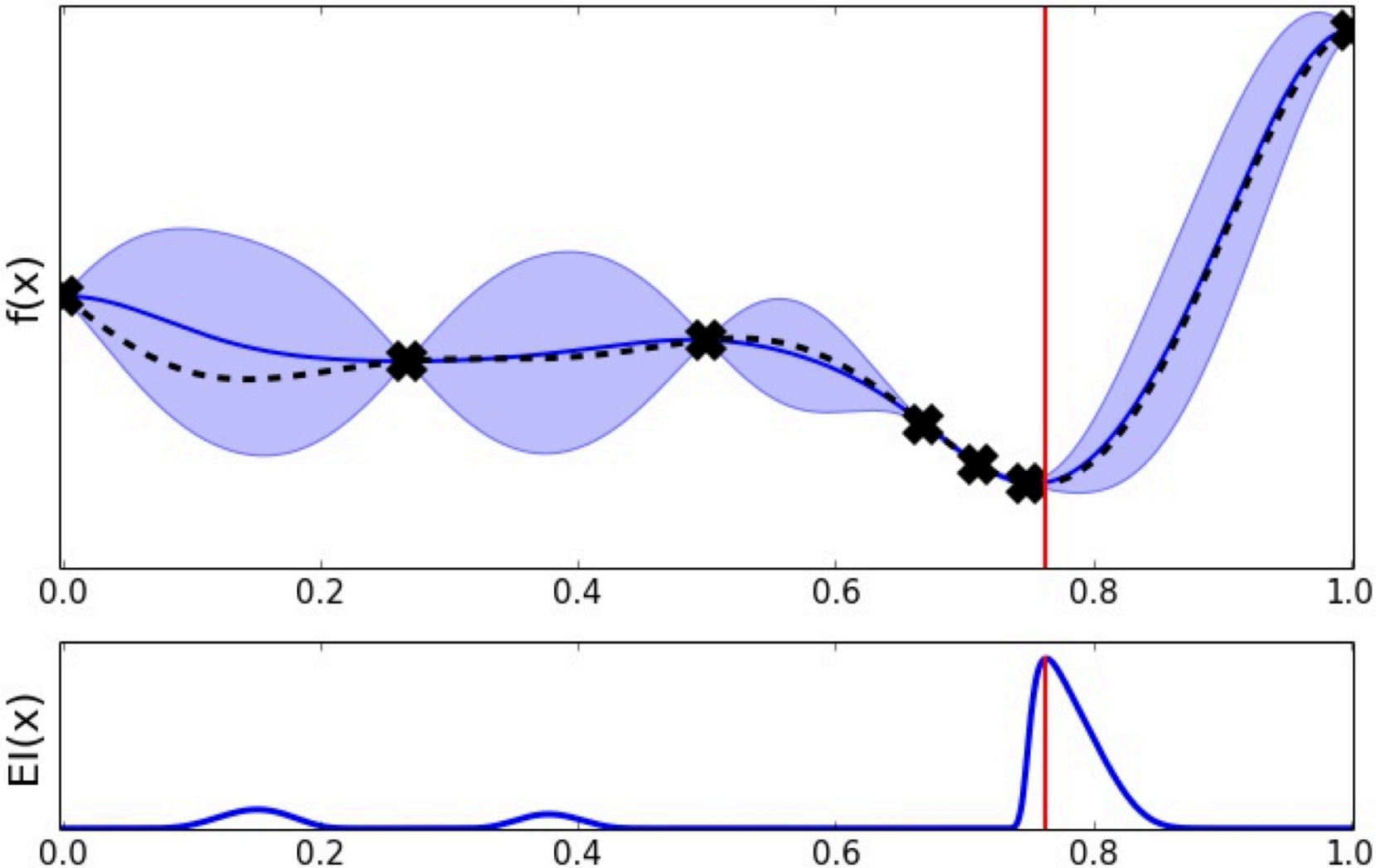
Expected Improvement: Toy Problem



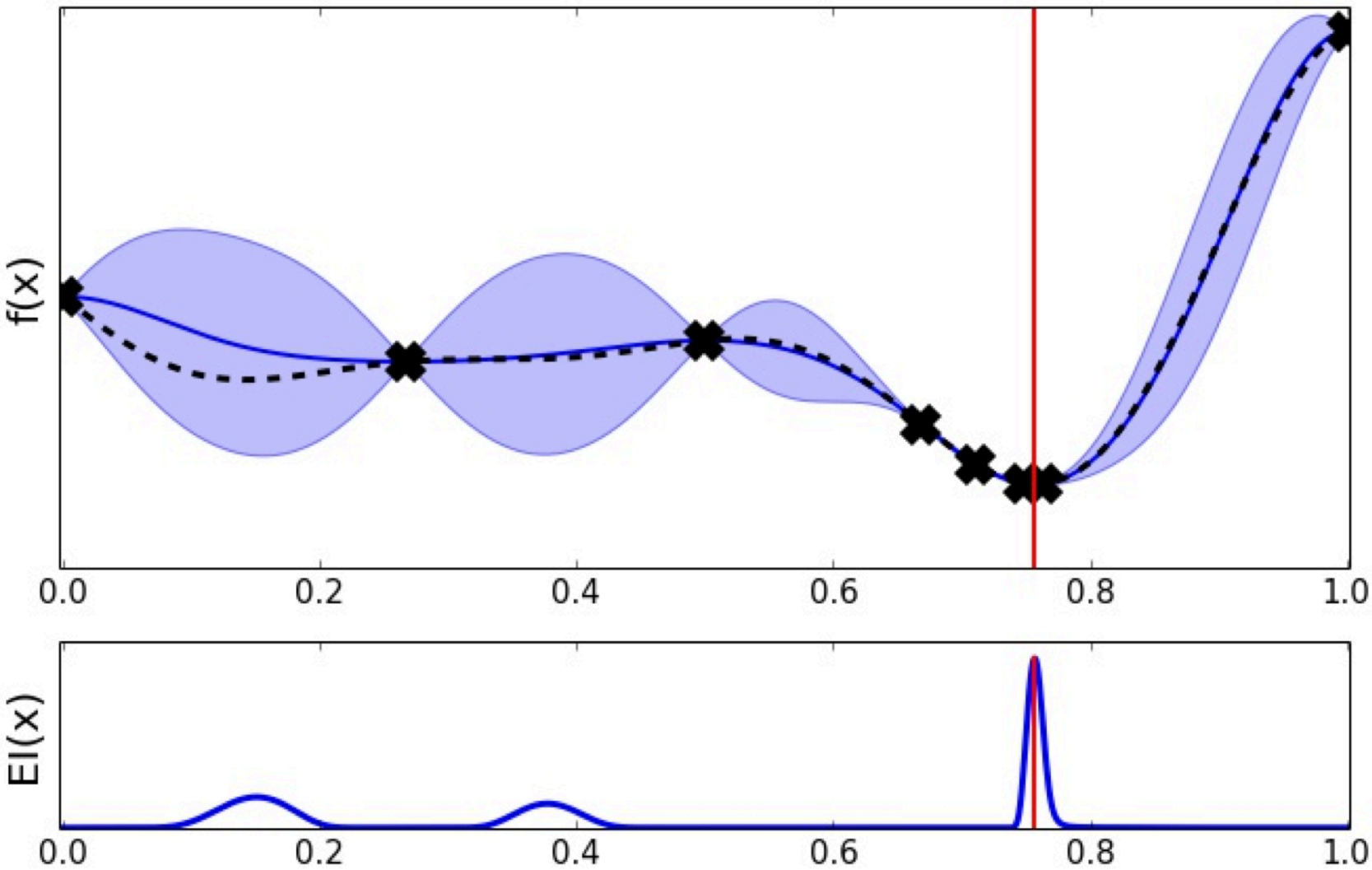
Expected Improvement: Toy Problem



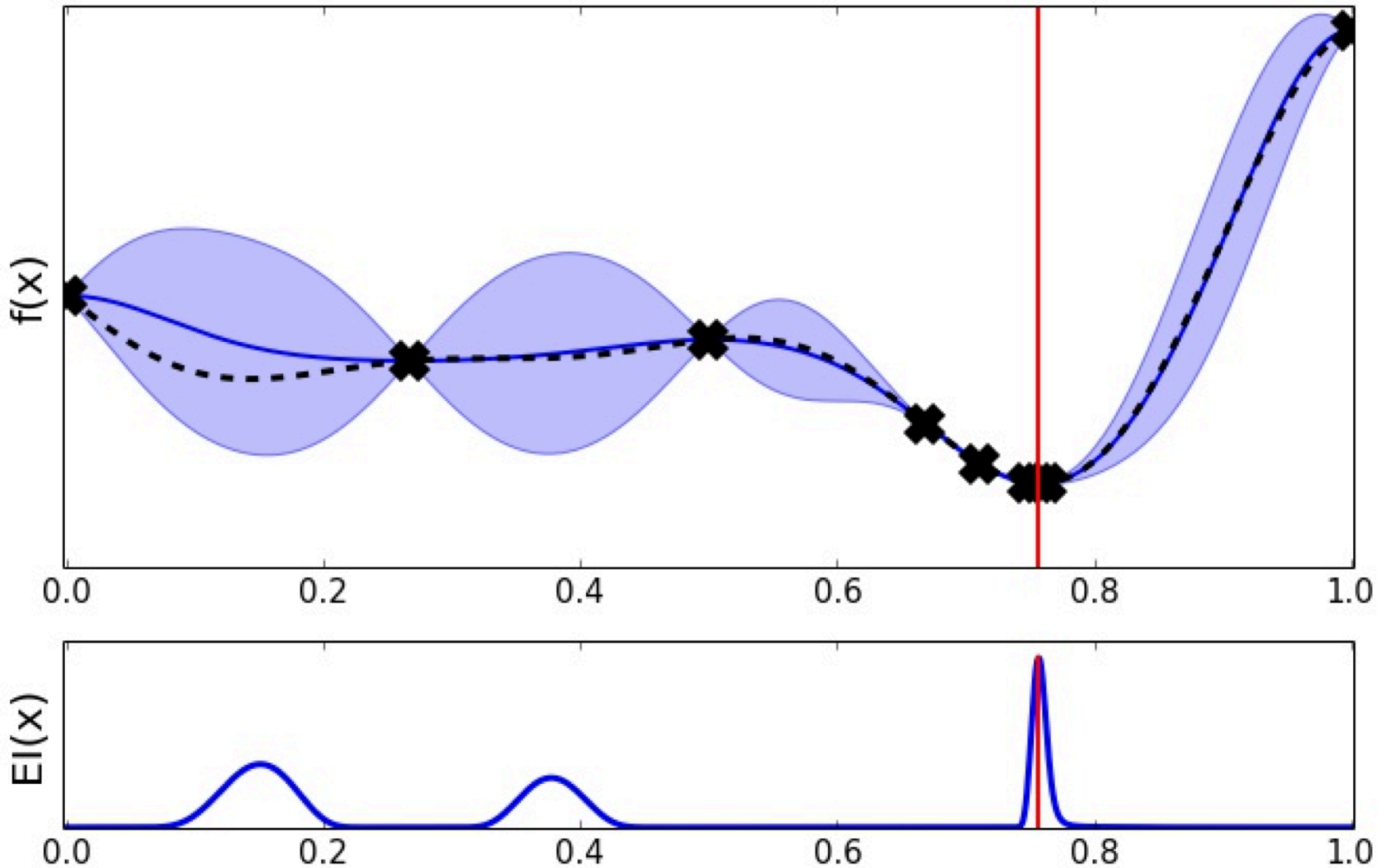
Expected Improvement: Toy Problem



Expected Improvement: Toy Problem



Expected Improvement: Toy Problem



- Scikit-learn dummy data for testing classifiers:

  n_samples=2500, n_features=45, n_informative=15, n_redundant=5

- Optimize w.r.t. $\mathrm{x} = (\mathrm{C}, \gamma)$, where C — penalization, and γ — kernel width
- The target function $\mathcal{L}(\mathrm{x})$ is the AUC based on three fold cross-validation

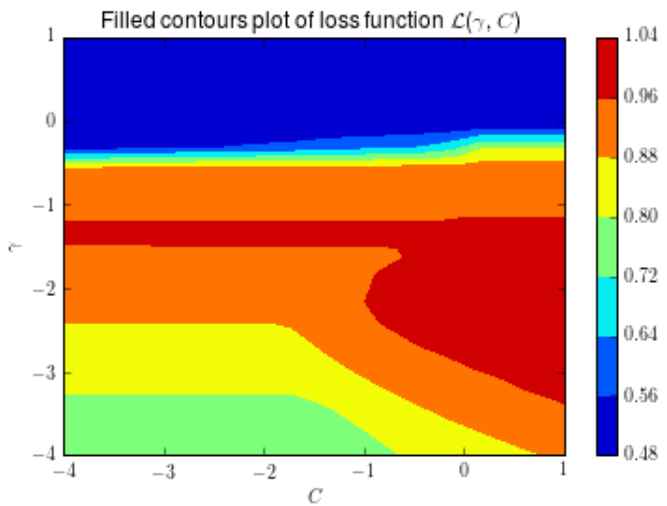
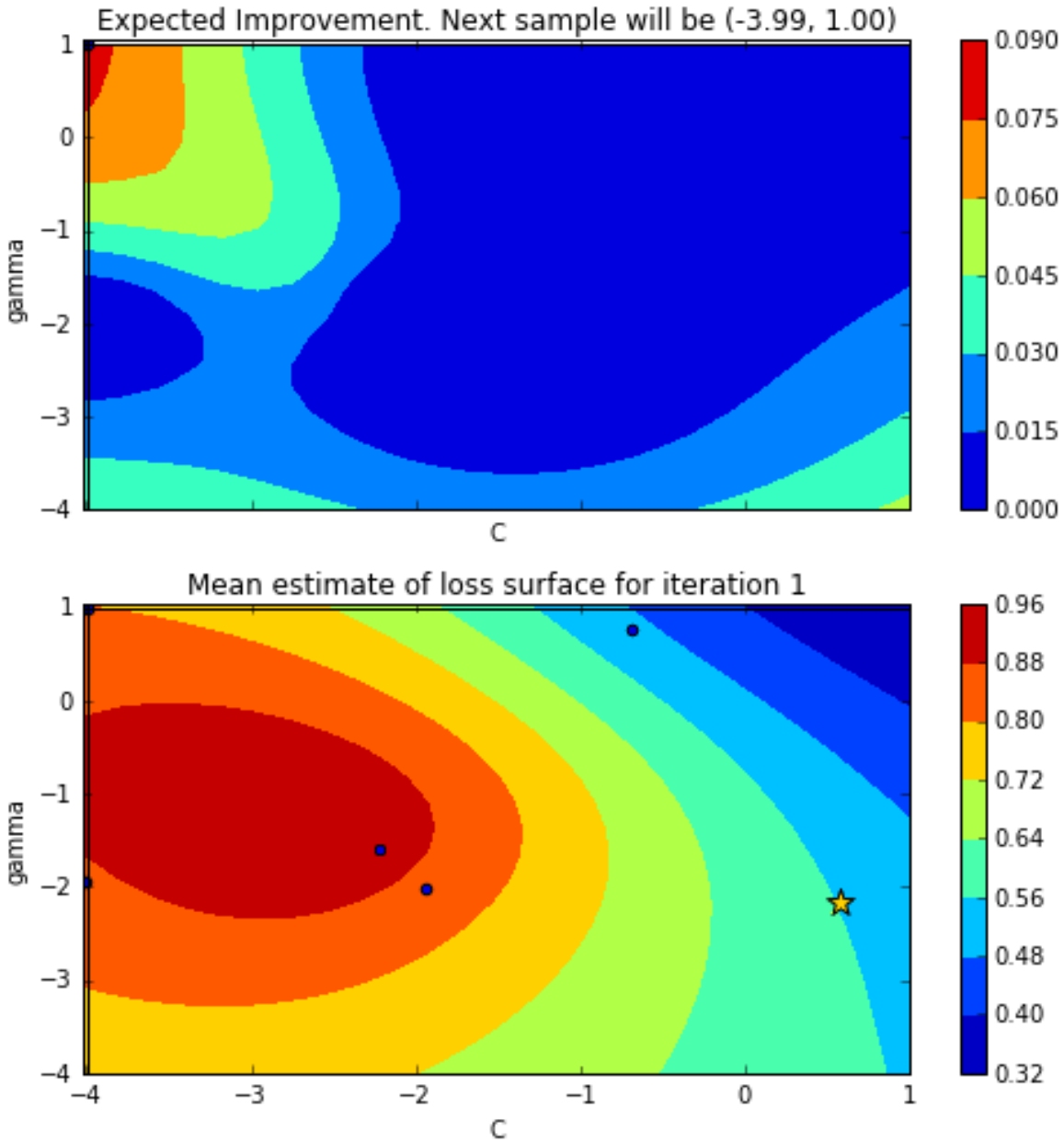
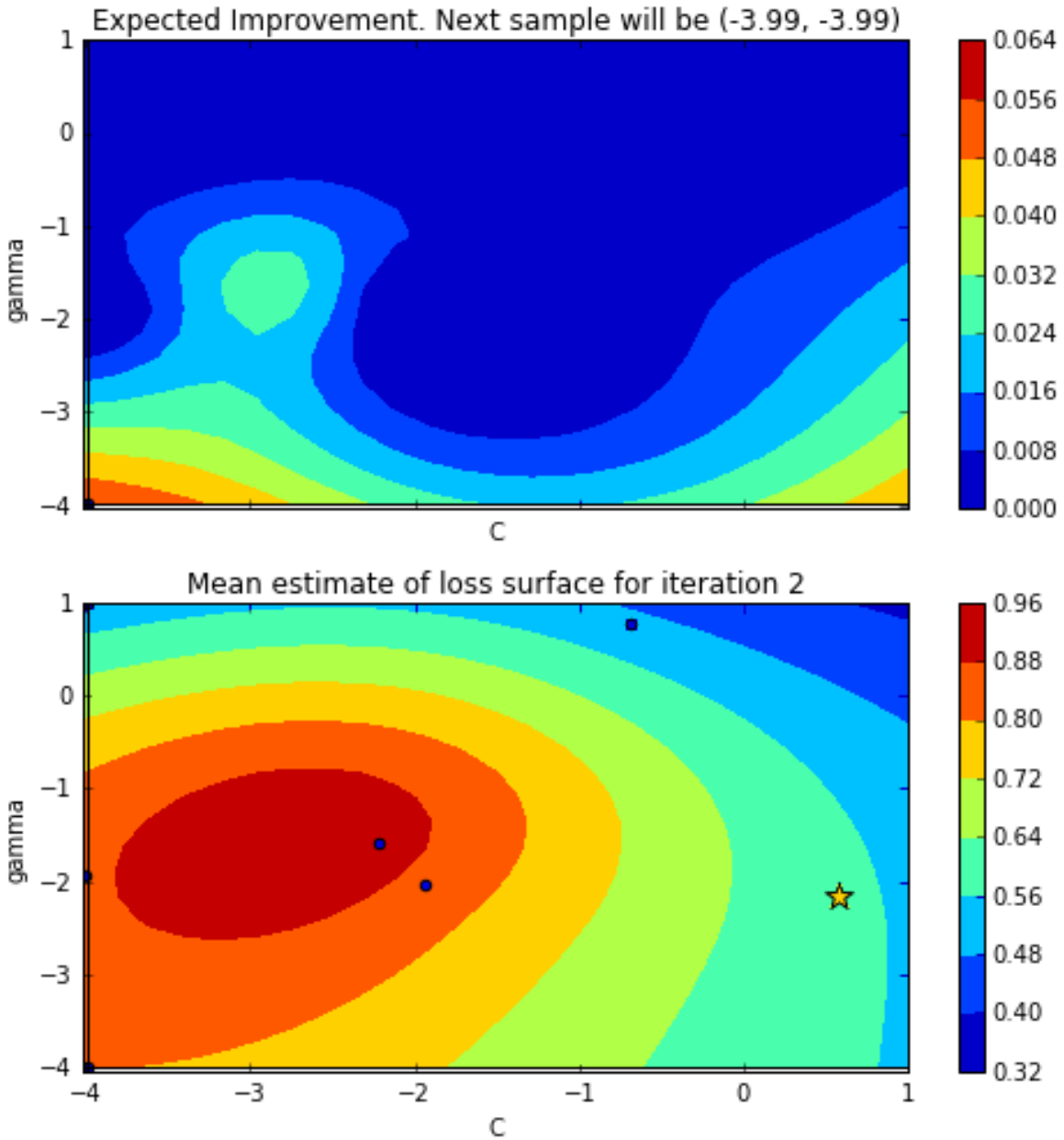


Figure – The target surface w.r.t. x to see where the true optimum is

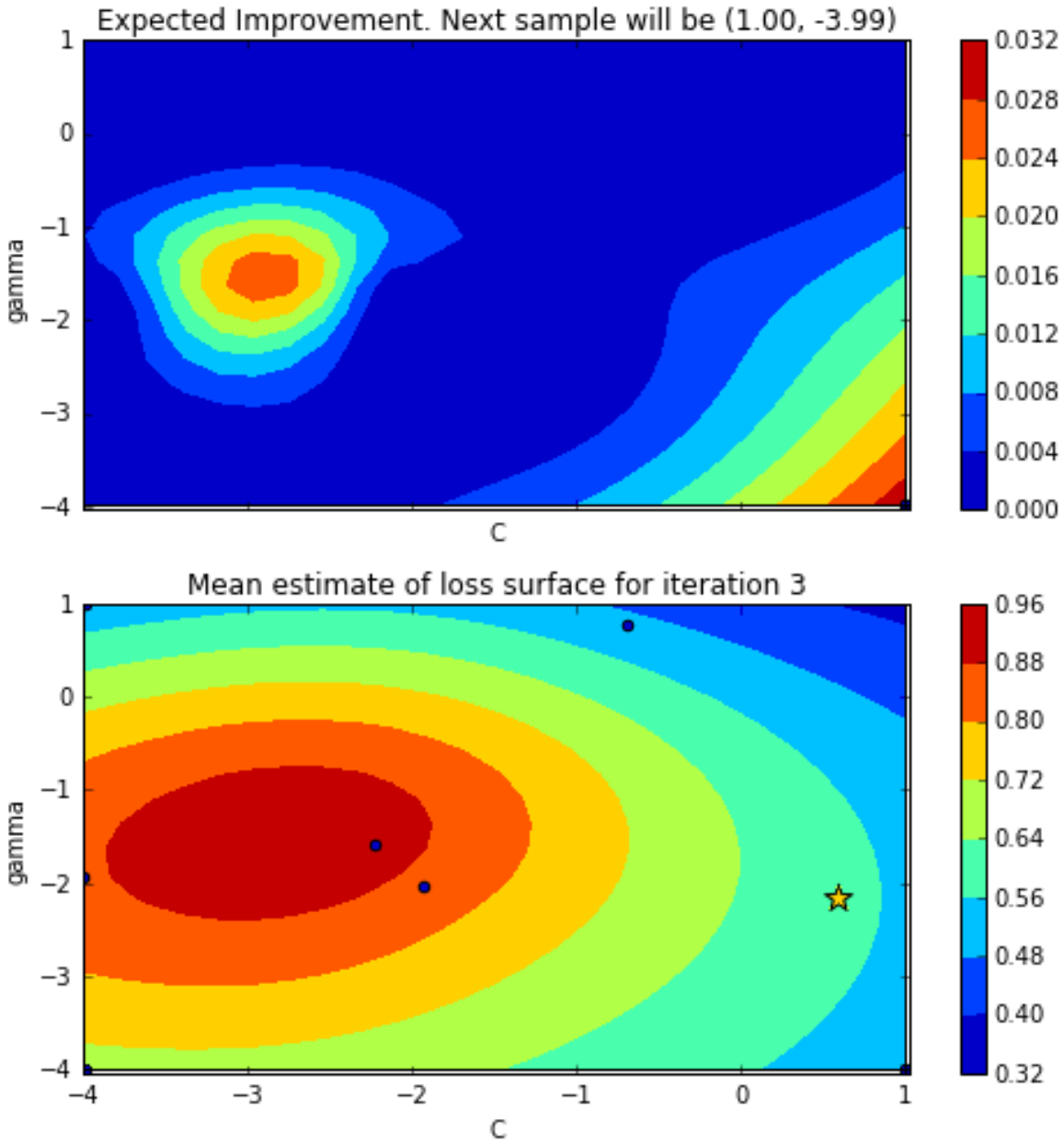
Iterations


Expected Improvement. Next sample will be (-3.99, 1.00)
gamma
C
1
0
−1
−2
−3
−4
0.090
0.075
0.060
0.045
0.030
0.015
0.000
Mean estimate of loss surface for iteration 1
gamma
C
−4
−3
−2
−1
0
1
0.96
0.88
0.80
0.72
0.64
0.56
0.48
0.40
0.32

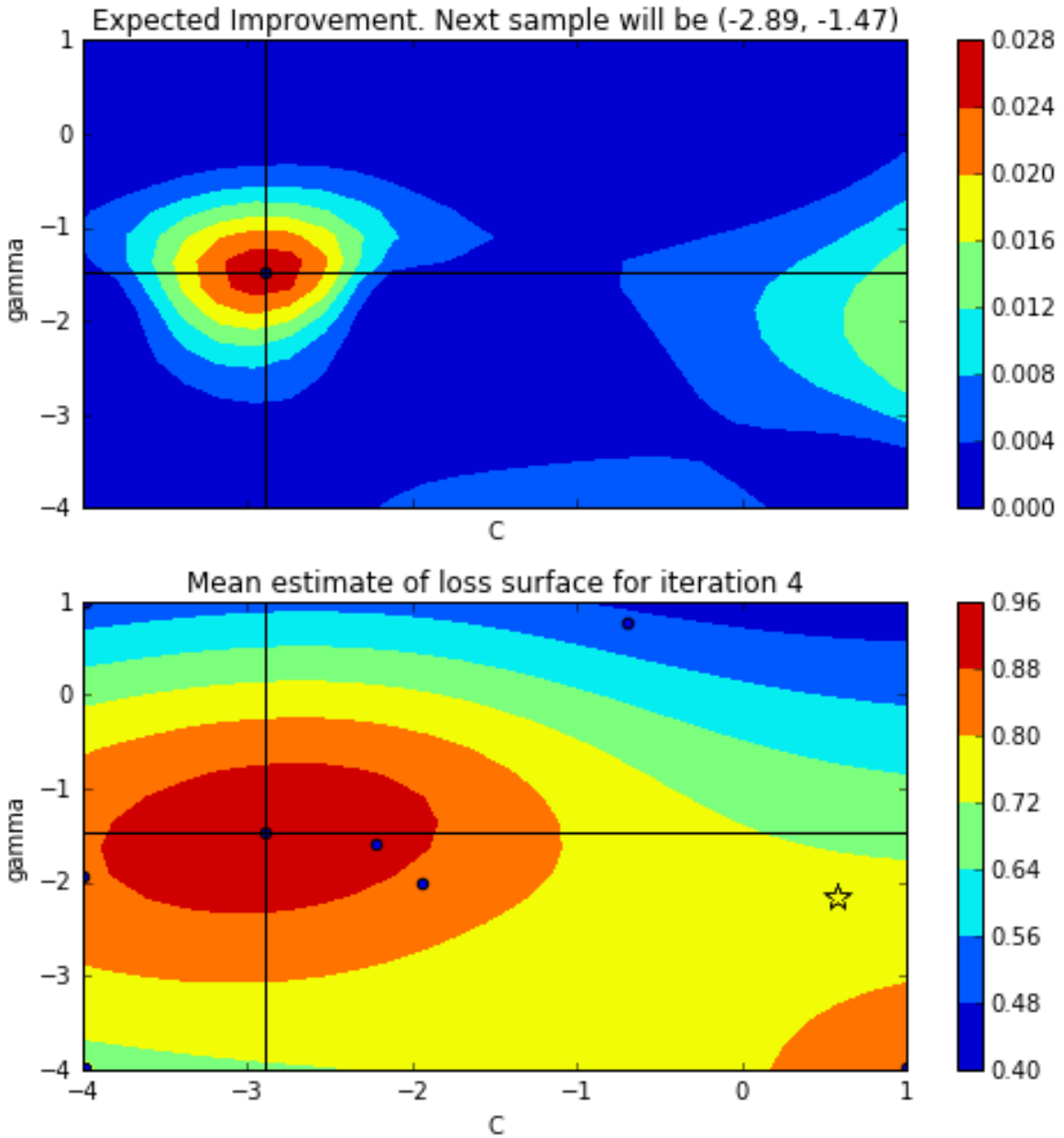
Iterations


Expected Improvement. Next sample will be (-3.99, -3.99)
gamma
C
1
0
−1
−2
−3
−4
0.064
0.056
0.048
0.040
0.032
0.024
0.016
0.008
0.000
Mean estimate of loss surface for iteration 2
gamma
C
−4
−3
−2
−1
0
1
0.96
0.88
0.80
0.72
0.64
0.56
0.48
0.40
0.32

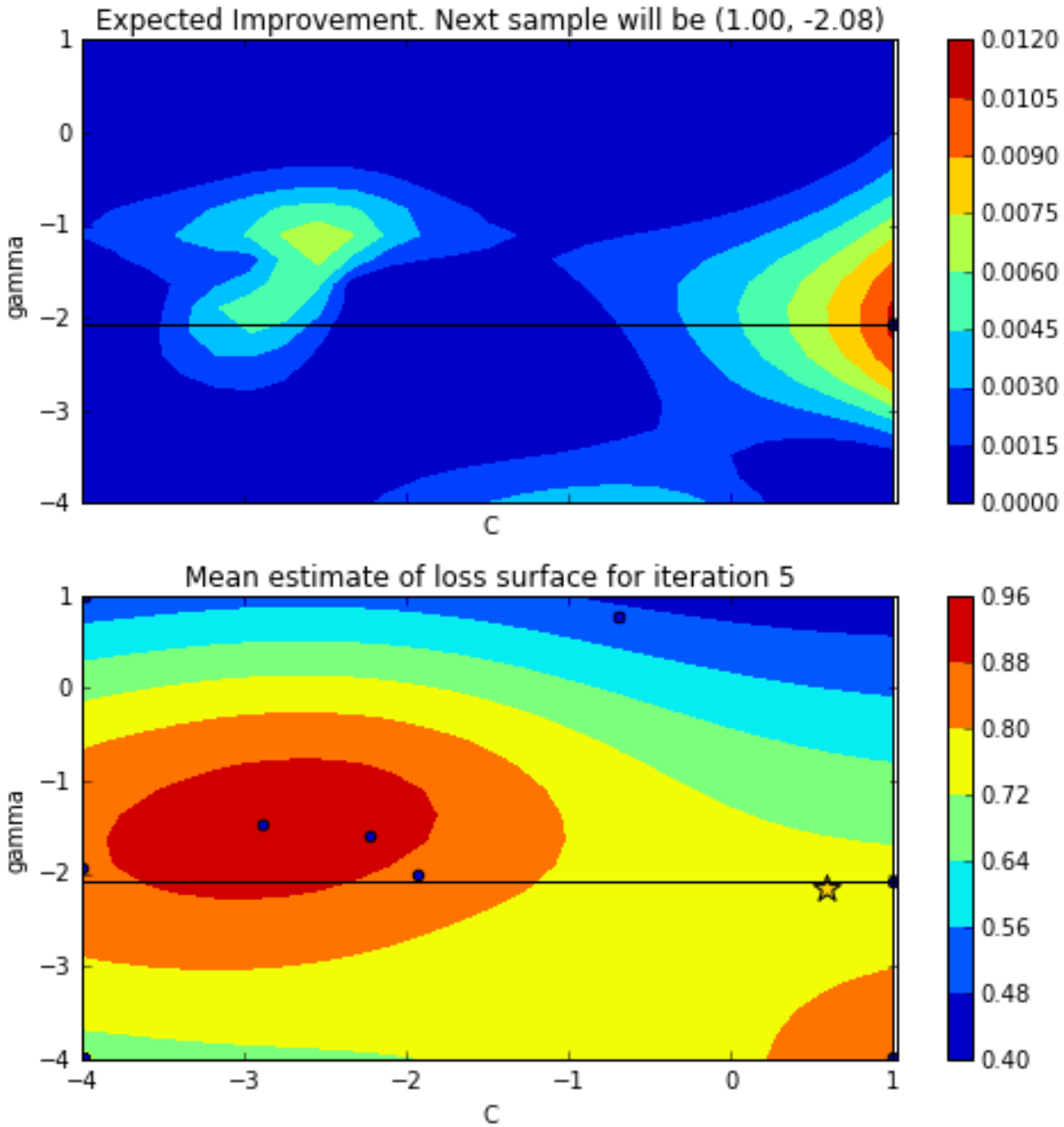
Iterations


Expected Improvement. Next sample will be (1.00, -3.99)
gamma
C
1
0
−1
−2
−3
−4
0.032
0.028
0.024
0.020
0.016
0.012
0.008
0.004
0.000
Mean estimate of loss surface for iteration 3
gamma
C
−4
−3
−2
−1
0
1
0.96
0.88
0.80
0.72
0.64
0.56
0.48
0.40
0.32

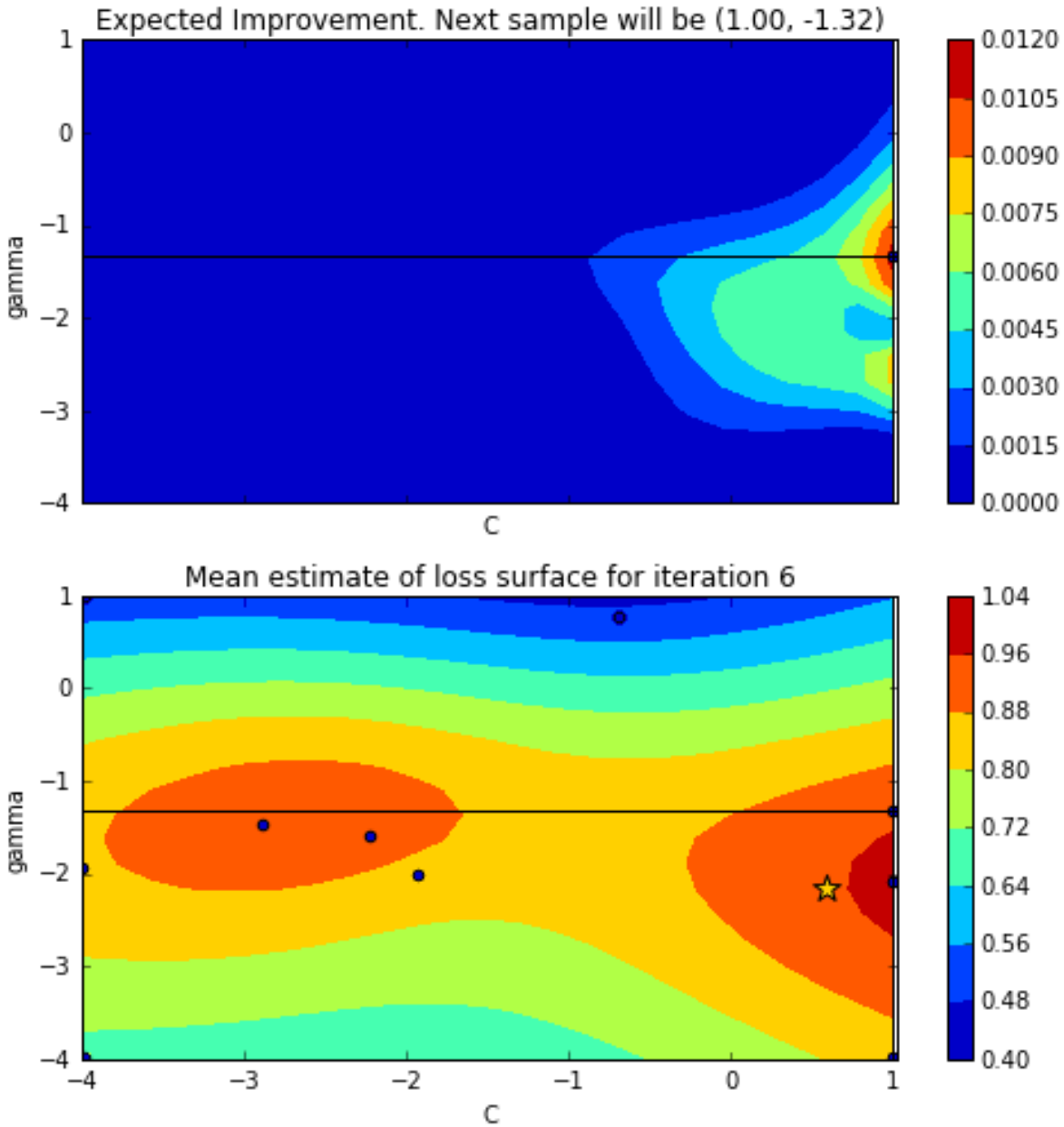
Iterations


Expected Improvement. Next sample will be (-2.89, -1.47)
gamma
C
0.028
0.024
0.020
0.016
0.012
0.008
0.004
0.000
Mean estimate of loss surface for iteration 4
gamma
C
0.96
0.88
0.80
0.72
0.64
0.56
0.48
0.40

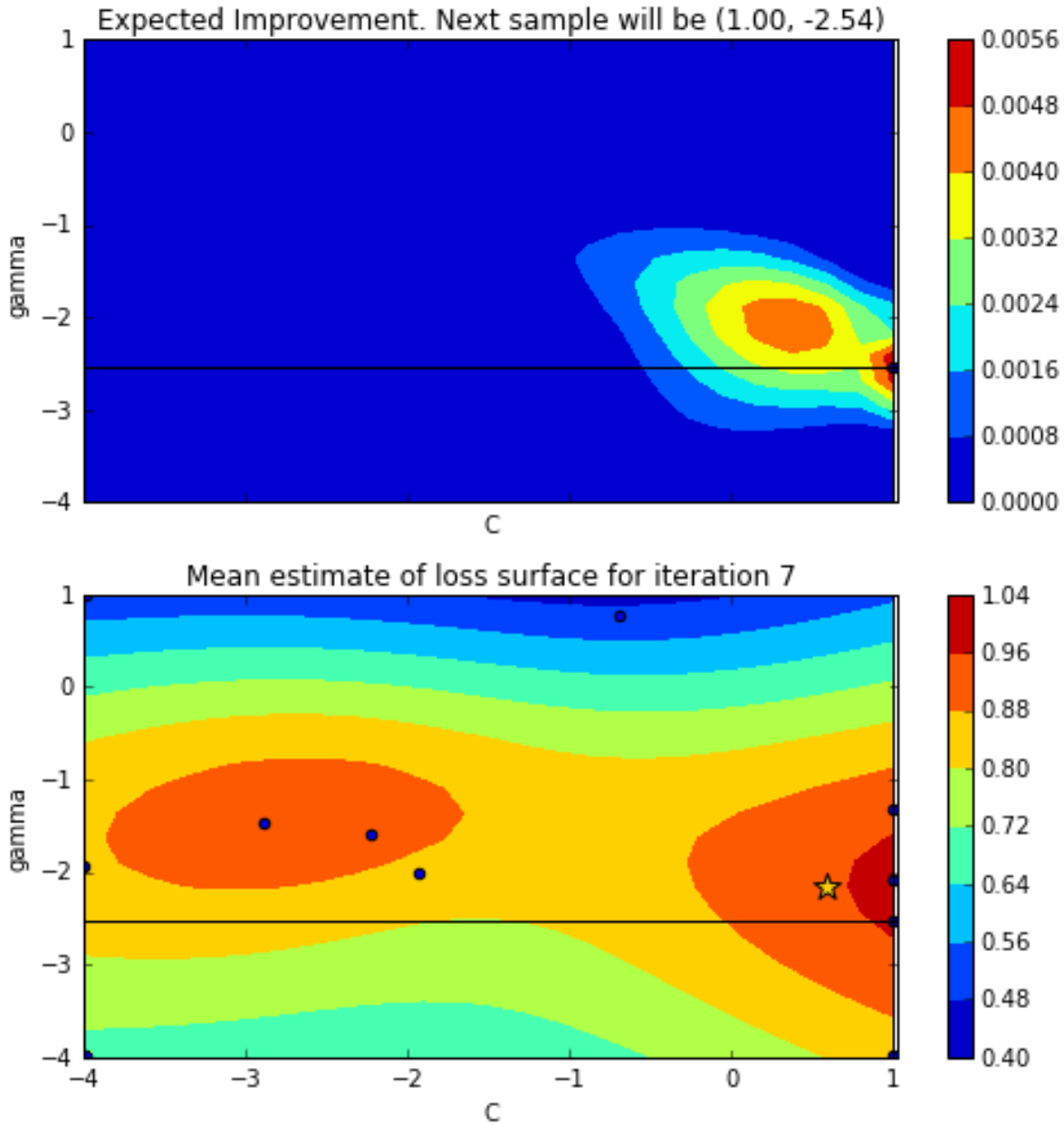
Iterations


Expected Improvement. Next sample will be (1.00, -2.08)
gamma
C
0.0120
0.0105
0.0090
0.0075
0.0060
0.0045
0.0030
0.0015
0.0000
Mean estimate of loss surface for iteration 5
gamma
C
0.96
0.88
0.80
0.72
0.64
0.56
0.48
0.40

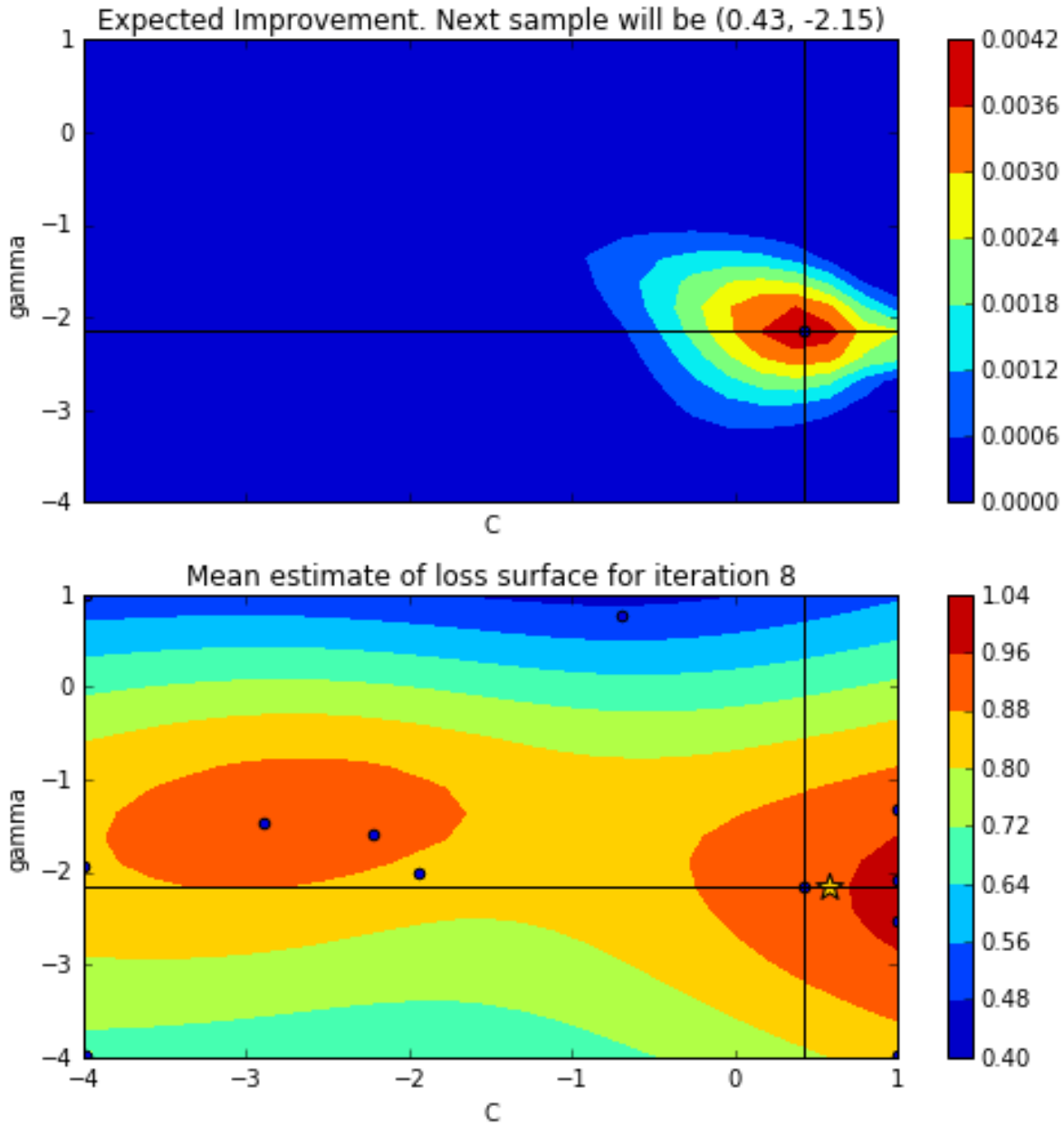
Iterations


Expected Improvement. Next sample will be (1.00, -1.32)
gamma
C
1
0
−1
−2
−3
−4
0.0120
0.0105
0.0090
0.0075
0.0060
0.0045
0.0030
0.0015
0.0000
Mean estimate of loss surface for iteration 6
gamma
C
−4
−3
−2
−1
0
1
1.04
0.96
0.88
0.80
0.72
0.64
0.56
0.48
0.40

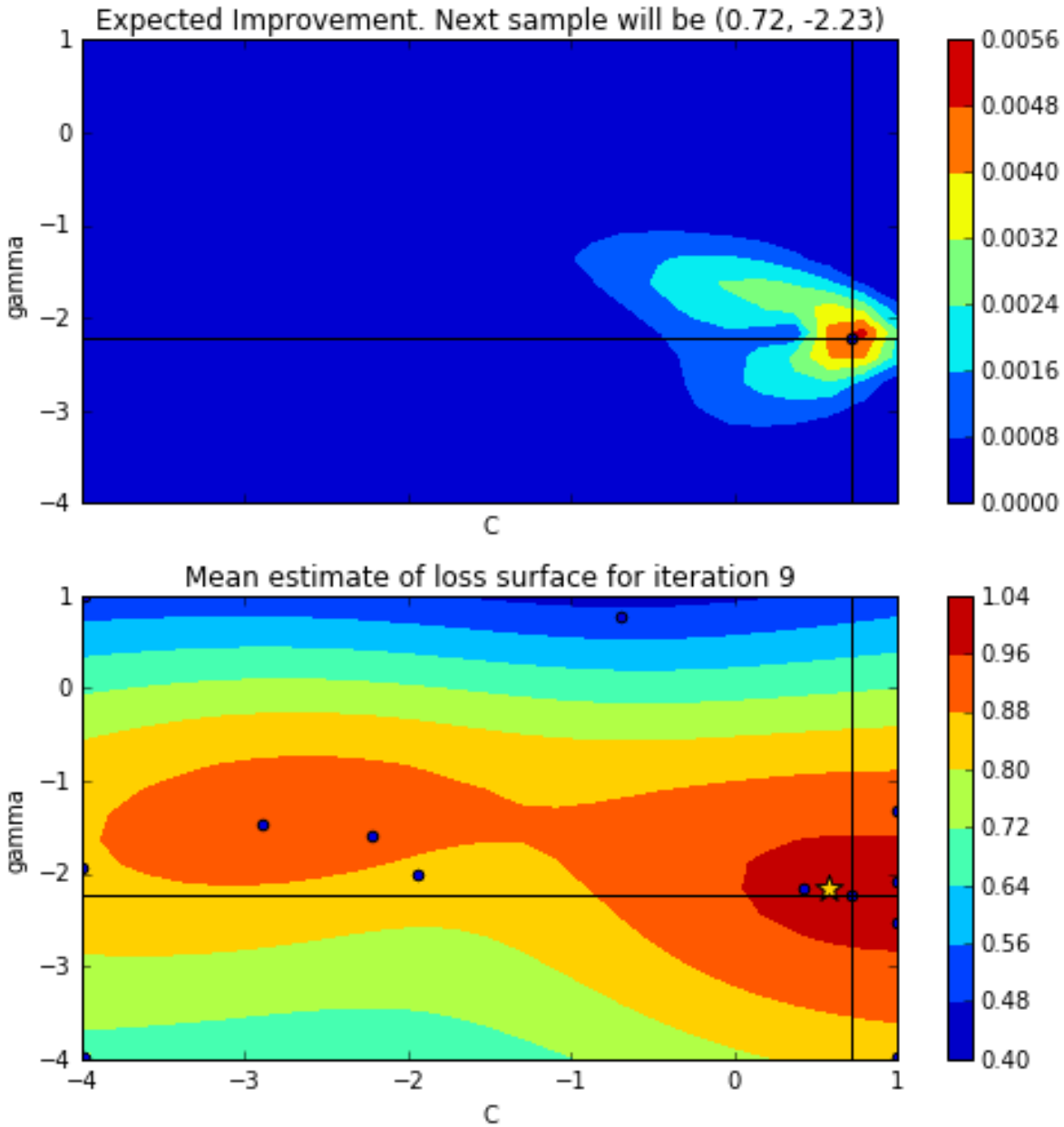
Iterations


Expected Improvement. Next sample will be (1.00, -2.54)
gamma
C
0.0056
0.0048
0.0040
0.0032
0.0024
0.0016
0.0008
0.0000
Mean estimate of loss surface for iteration 7
gamma
C
1.04
0.96
0.88
0.80
0.72
0.64
0.56
0.48
0.40

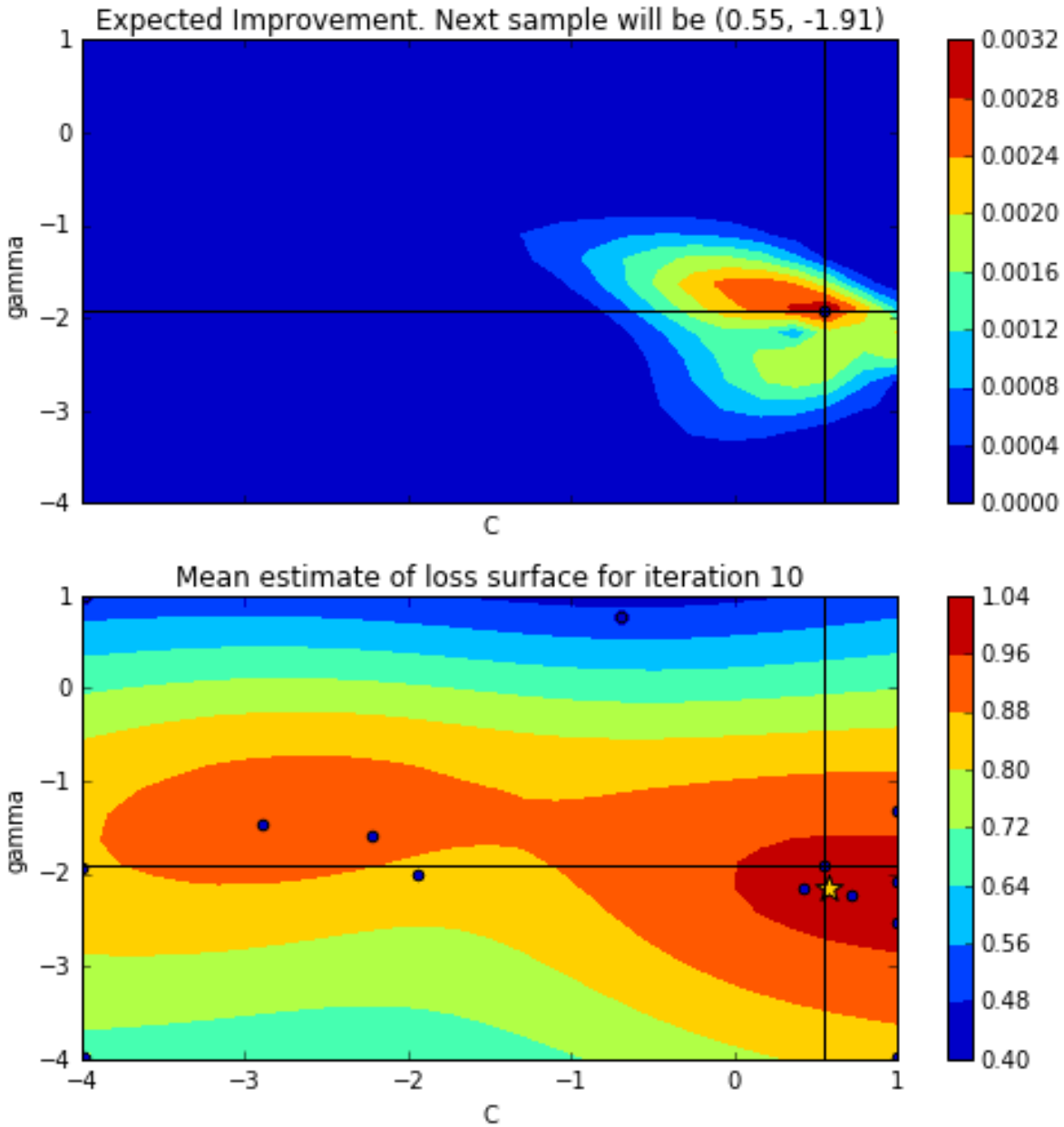
Iterations


Expected Improvement. Next sample will be (0.43, -2.15)
gamma
C
0.0042
0.0036
0.0030
0.0024
0.0018
0.0012
0.0006
0.0000
Mean estimate of loss surface for iteration 8
gamma
C
1.04
0.96
0.88
0.80
0.72
0.64
0.56
0.48
0.40

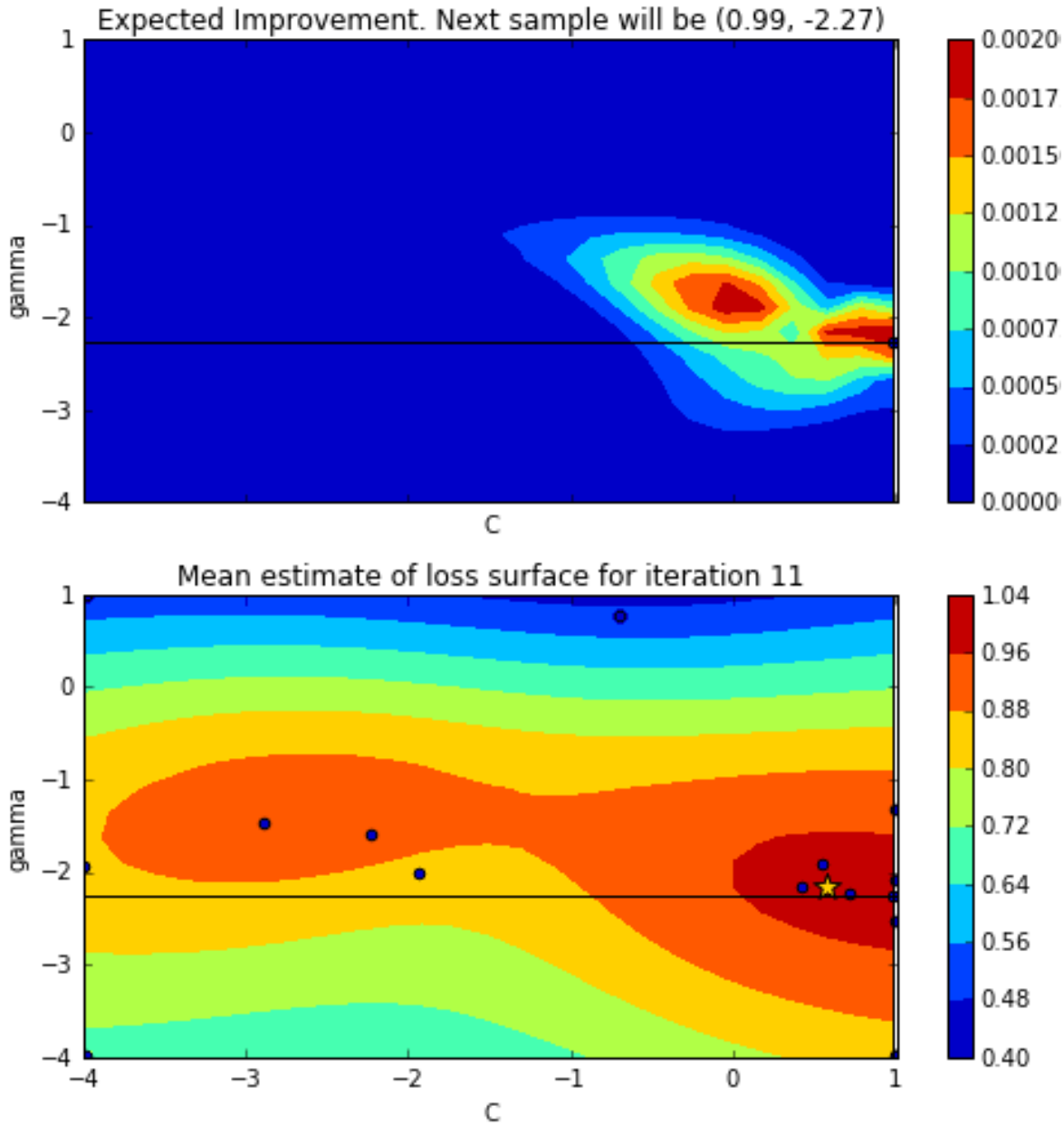
Iterations


Expected Improvement. Next sample will be (0.72, -2.23)
gamma
C
0.0056
0.0048
0.0040
0.0032
0.0024
0.0016
0.0008
0.0000
Mean estimate of loss surface for iteration 9
gamma
C
1.04
0.96
0.88
0.80
0.72
0.64
0.56
0.48
0.40

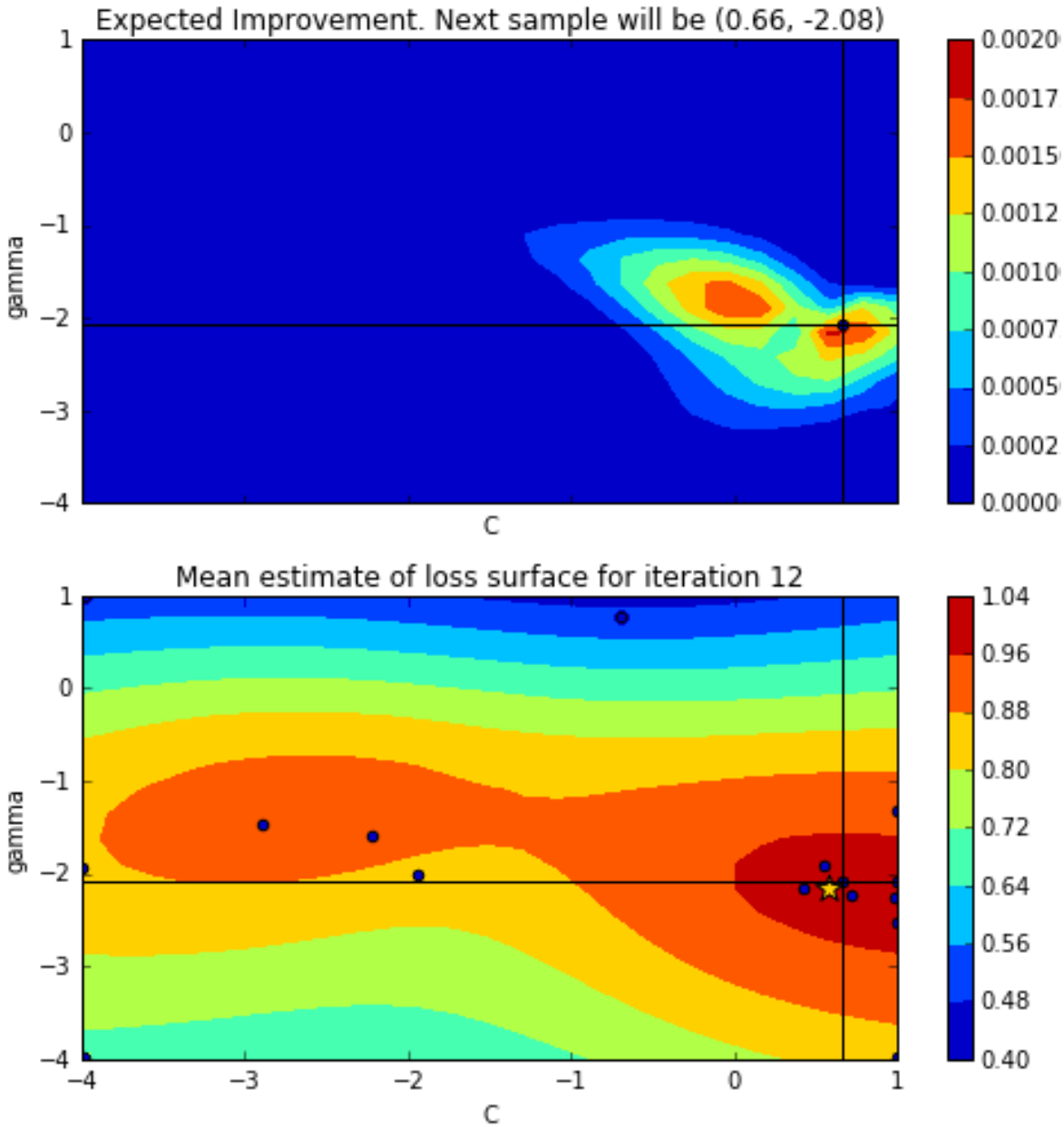
Iterations


Expected Improvement. Next sample will be (0.55, -1.91)
gamma
C
1
0
−1
−2
−3
−4
0.0032
0.0028
0.0024
0.0020
0.0016
0.0012
0.0008
0.0004
0.0000
Mean estimate of loss surface for iteration 10
gamma
C
−4
−3
−2
−1
0
1
1.04
0.96
0.88
0.80
0.72
0.64
0.56
0.48
0.40

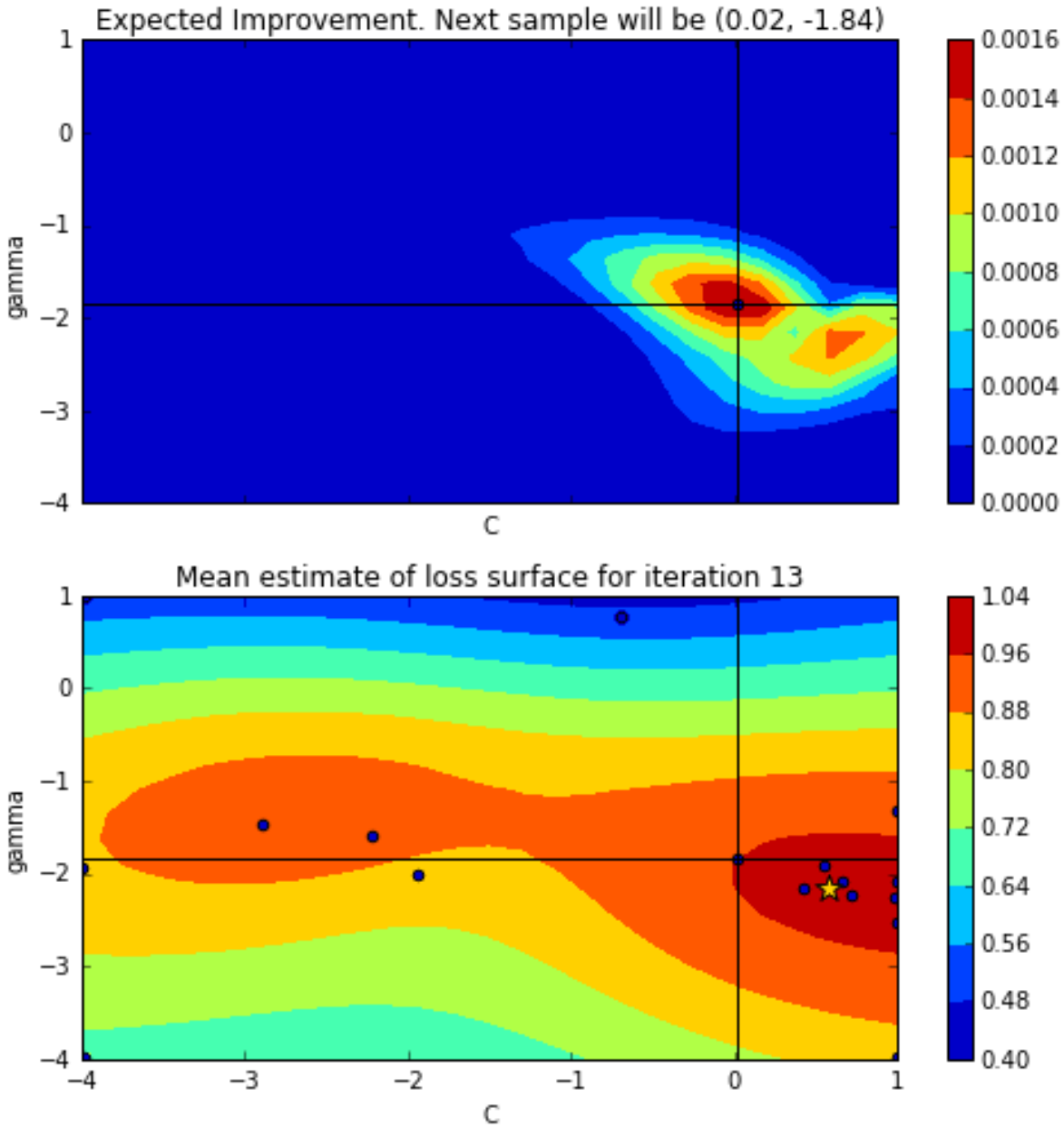
Iterations


Expected Improvement. Next sample will be (0.99, -2.27)
gamma
C
0.0020
0.0017
0.0015
0.0012
0.0010
0.0007
0.0005
0.0002
0.0000
Mean estimate of loss surface for iteration 11
gamma
C
1.04
0.96
0.88
0.80
0.72
0.64
0.56
0.48
0.40

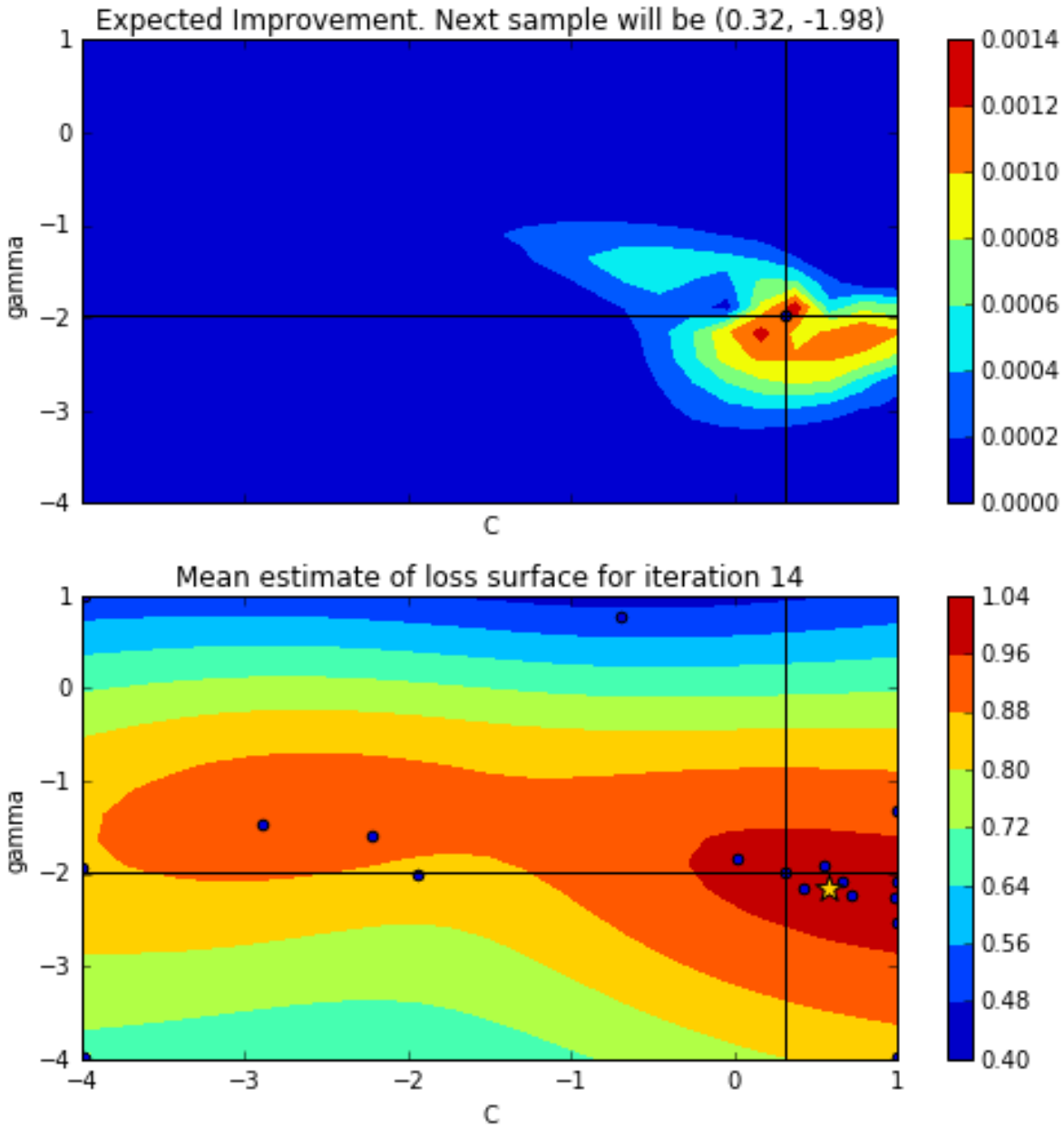
Iterations


Expected Improvement. Next sample will be (0.66, -2.08)
gamma
C
1
0
−1
−2
−3
−4
0.0020
0.0017
0.0015
0.0012
0.0010
0.0007
0.0005
0.0002
0.0000
Mean estimate of loss surface for iteration 12
gamma
C
−4
−3
−2
−1
0
1
1.04
0.96
0.88
0.80
0.72
0.64
0.56
0.48
0.40

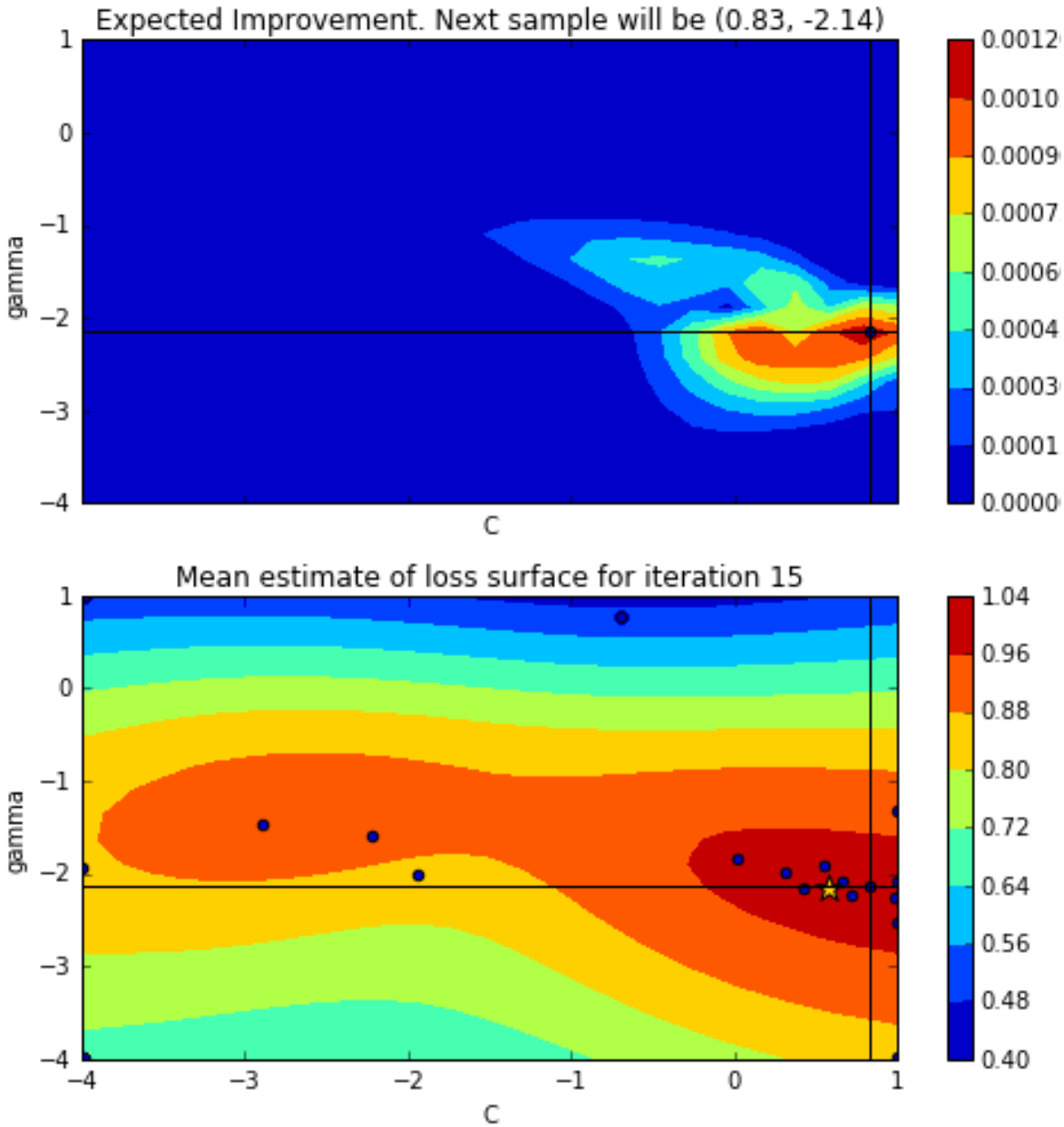
Iterations


Expected Improvement. Next sample will be (0.02, -1.84)
gamma
C
0.0016
0.0014
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
Mean estimate of loss surface for iteration 13
gamma
C
1.04
0.96
0.88
0.80
0.72
0.64
0.56
0.48
0.40

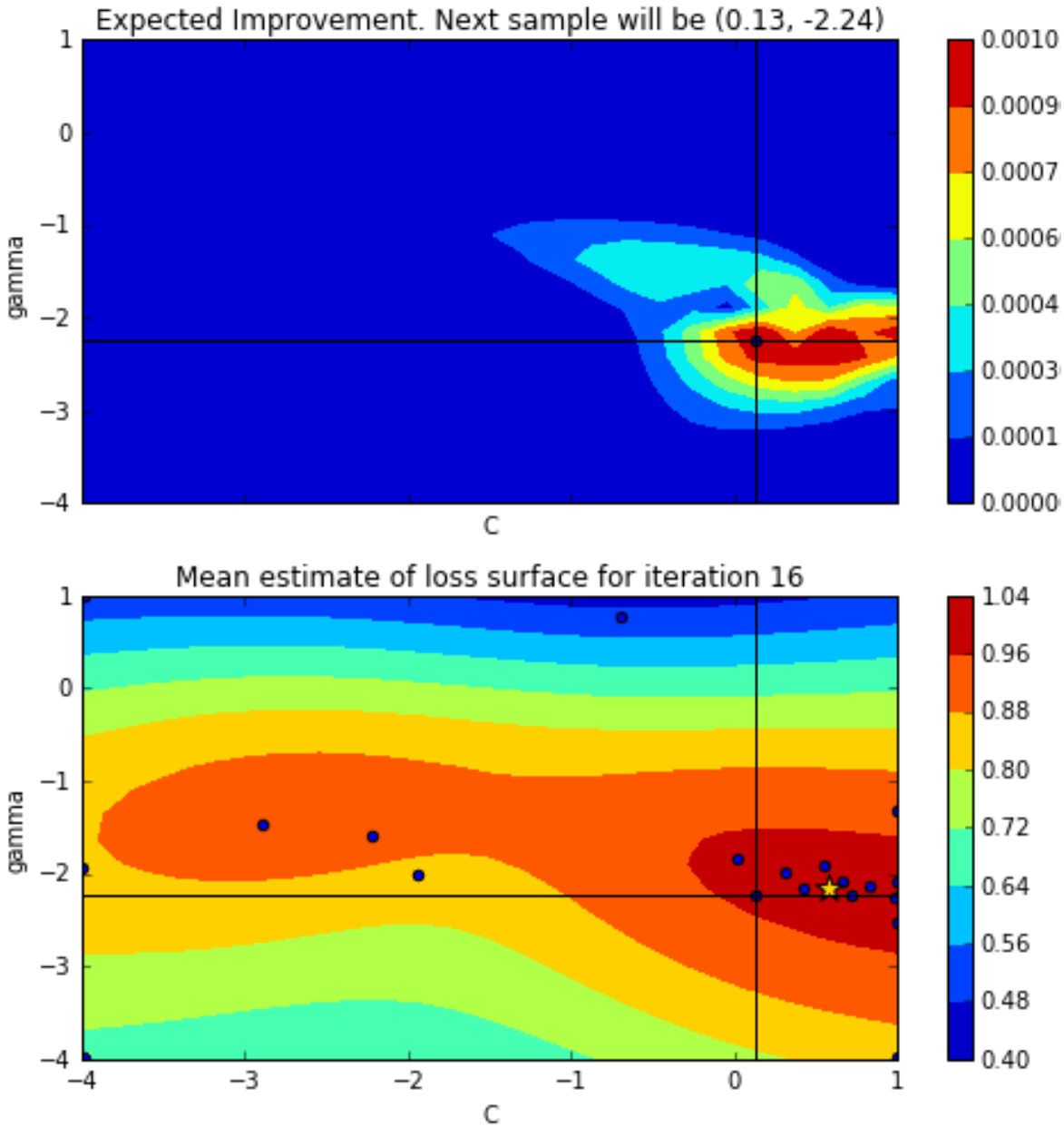
Iterations


Expected Improvement. Next sample will be (0.32, -1.98)
gamma
C
0.0014
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
Mean estimate of loss surface for iteration 14
gamma
C
1.04
0.96
0.88
0.80
0.72
0.64
0.56
0.48
0.40

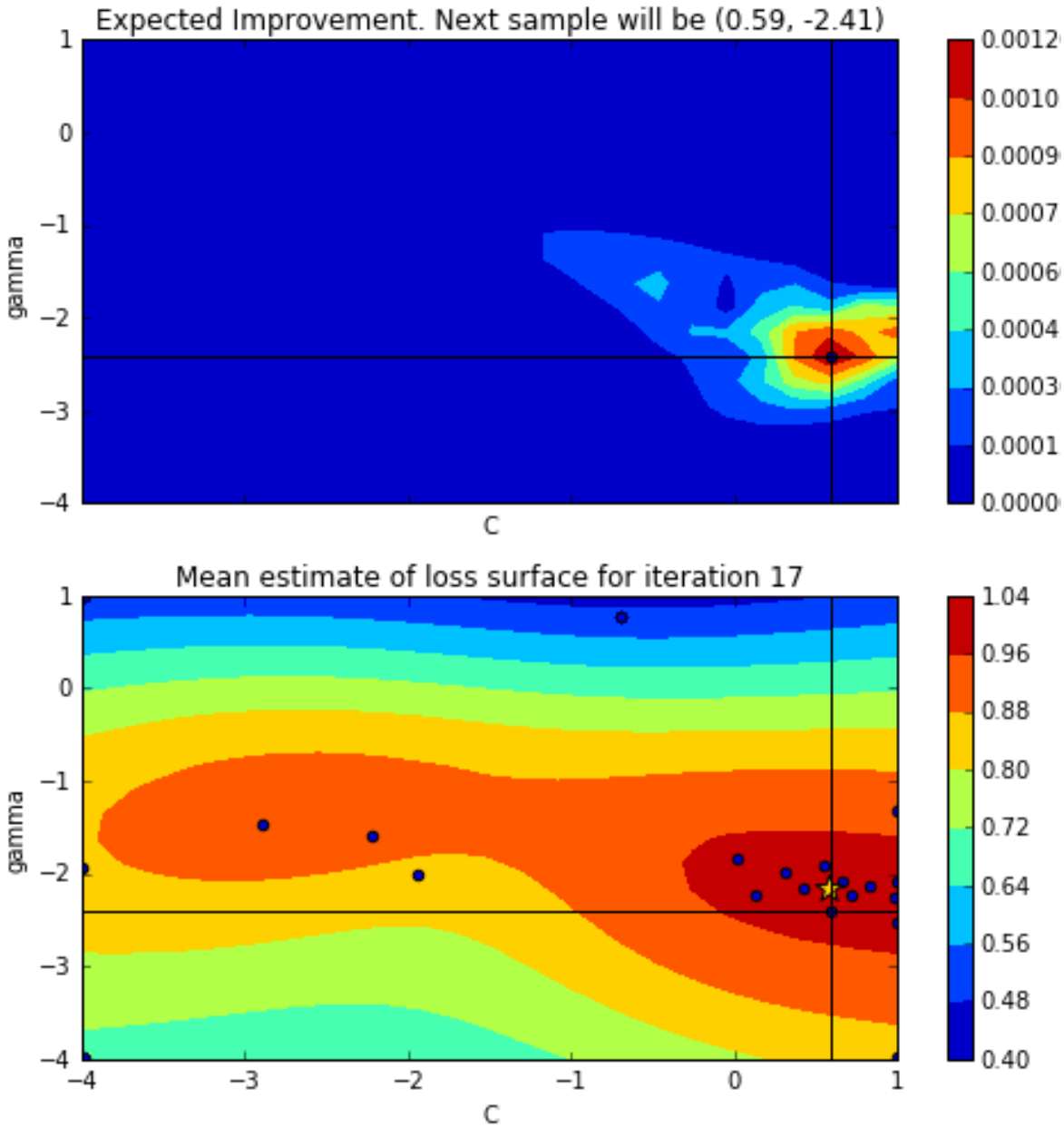
Iterations


Expected Improvement. Next sample will be (0.83, -2.14)
gamma
C
0.0012
0.0010
0.0009
0.0007
0.0006
0.0004
0.0003
0.0001
0.0000
Mean estimate of loss surface for iteration 15
gamma
C
1.04
0.96
0.88
0.80
0.72
0.64
0.56
0.48
0.40

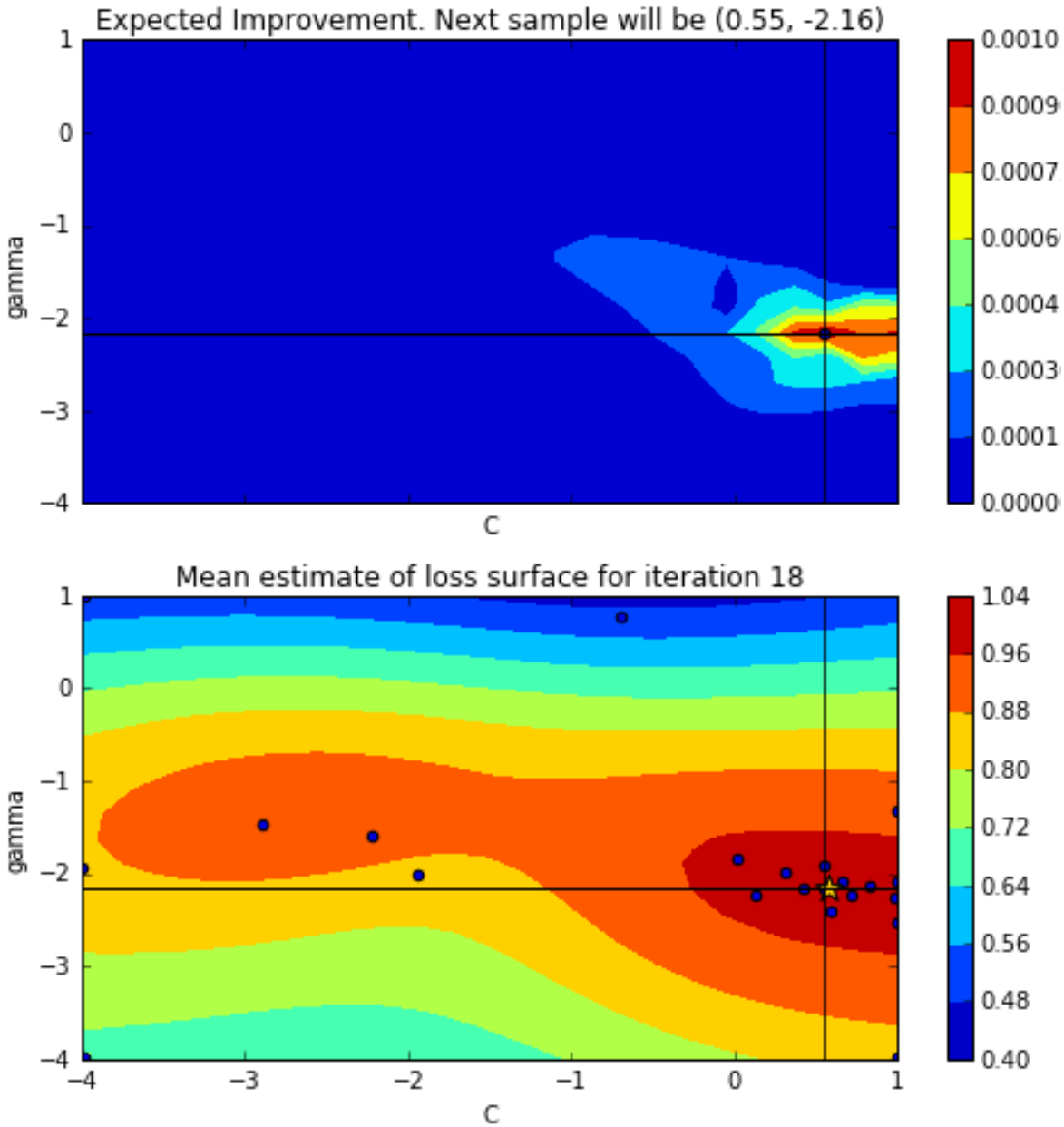
Iterations


Expected Improvement. Next sample will be (0.13, -2.24)
gamma
C
0.0010
0.0009
0.0007
0.0006
0.0004
0.0003
0.0001
0.0000
Mean estimate of loss surface for iteration 16
gamma
C
1.04
0.96
0.88
0.80
0.72
0.64
0.56
0.48
0.40

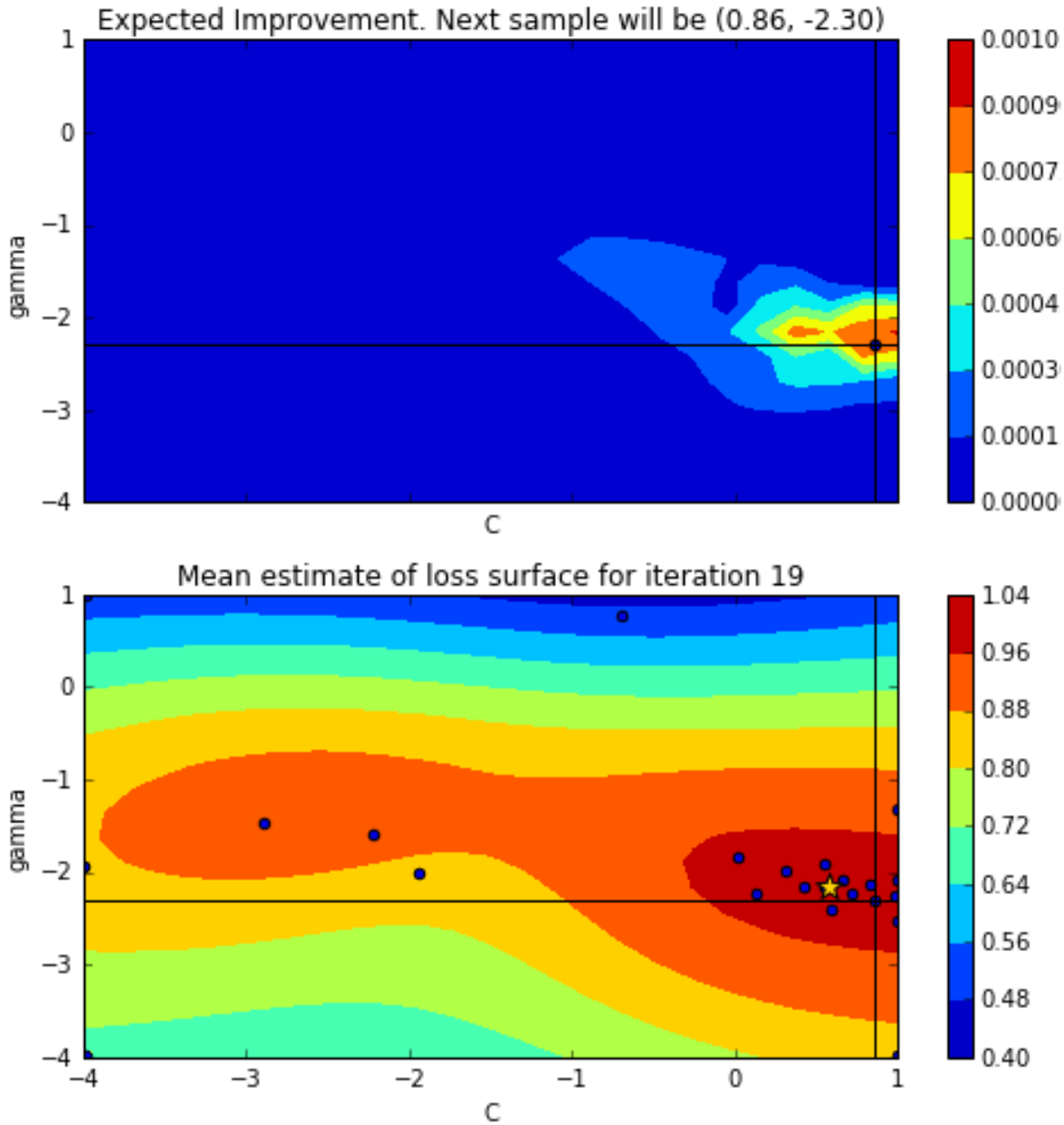
Iterations


Expected Improvement. Next sample will be (0.59, -2.41)
gamma
C
0.0012
0.0010
0.0009
0.0007
0.0006
0.0004
0.0003
0.0001
0.0000
Mean estimate of loss surface for iteration 17
gamma
C
1.04
0.96
0.88
0.80
0.72
0.64
0.56
0.48
0.40

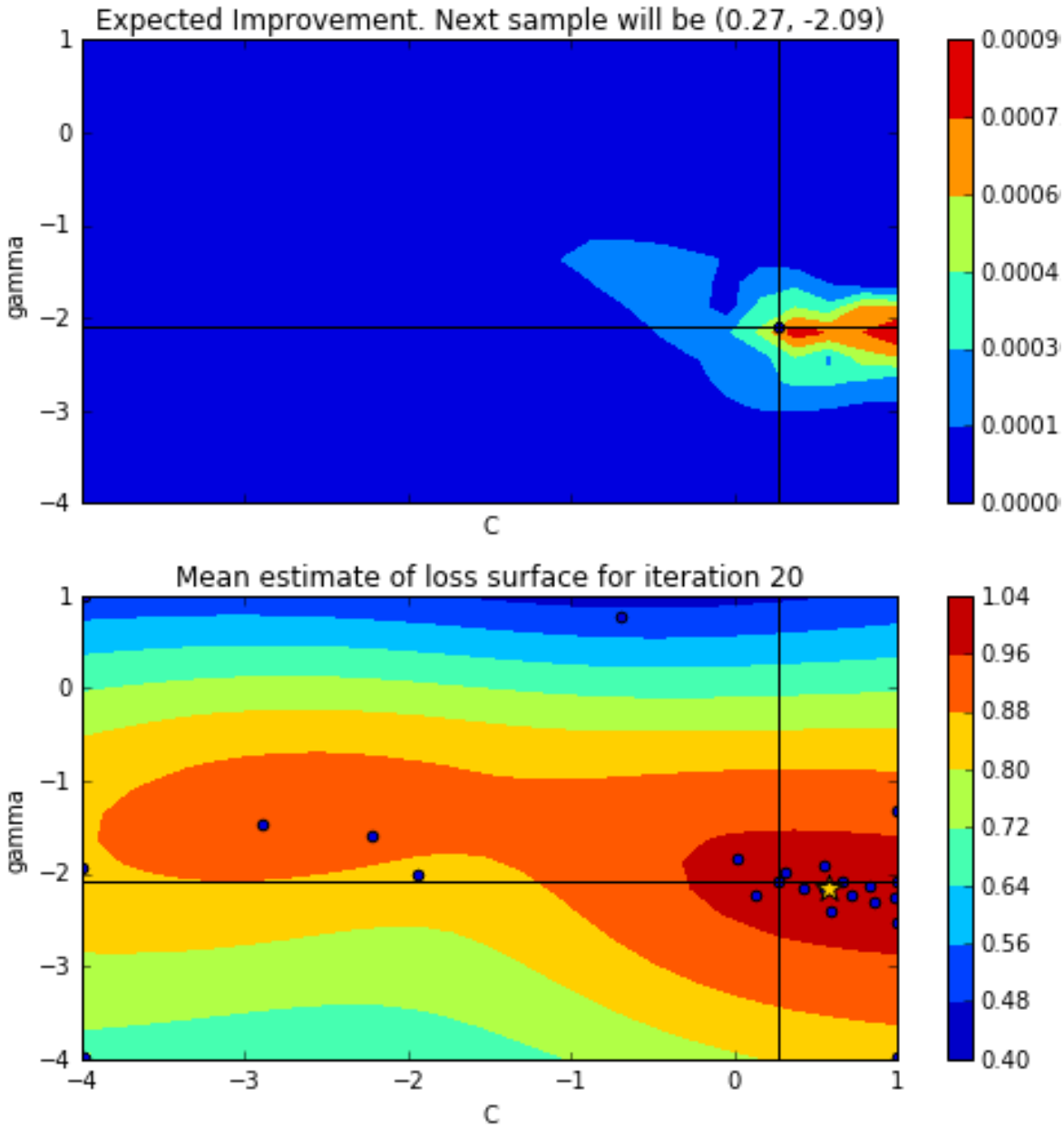
Iterations


Expected Improvement. Next sample will be (0.55, -2.16)
gamma
C
0.0010
0.0009
0.0007
0.0006
0.0004
0.0003
0.0001
0.0000
Mean estimate of loss surface for iteration 18
gamma
C
1.04
0.96
0.88
0.80
0.72
0.64
0.56
0.48
0.40

Iterations


Expected Improvement. Next sample will be (0.86, -2.30)
gamma
C
0.0010
0.0009
0.0007
0.0006
0.0004
0.0003
0.0001
0.0000
Mean estimate of loss surface for iteration 19
gamma
C
1.04
0.96
0.88
0.80
0.72
0.64
0.56
0.48
0.40

Iterations


Expected Improvement. Next sample will be (0.27, -2.09)
gamma
C
1
0
−1
−2
−3
−4
0.0009
0.0007
0.0006
0.0004
0.0003
0.0001
0.0000
Mean estimate of loss surface for iteration 20
gamma
C
−4
−3
−2
−1
0
1
1.04
0.96
0.88
0.80
0.72
0.64
0.56
0.48
0.40