Bayesian Optimization

Shu-Yang Ye

University of British Columbia

January 14, 2021

Mining Gold

Suppose you are a gold digger.

Figure 1: Gold Digger

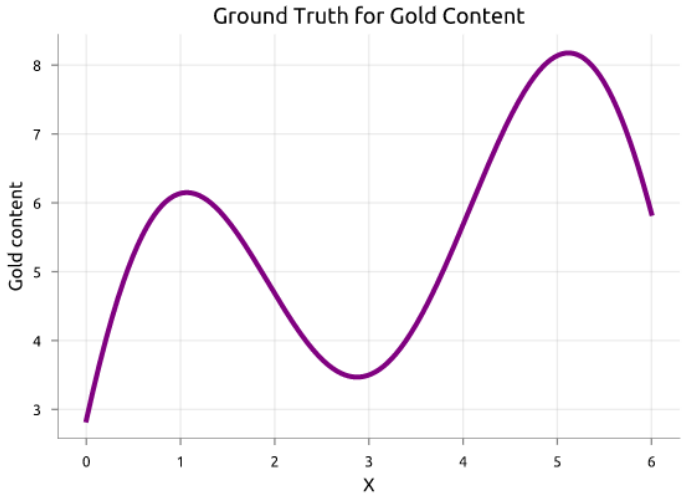


Figure 2: Gold Content

You have no idea how does the gold distributed under the ground. You can learn the gold distribution by drilling at different locations. However, this drilling is costly.

Mining Gold

Problem:Location of Maximum Gold (Bayesian Optimization)

- We want to find the location of the maximum gold content.
- We can not drill at every location. Instead, we should drill at locations showing high promise about the gold content.
- This problem is akin to Bayesian Optimization

Beyesian Optimization

The Bayesian Optimization has two major components:

- A surrogate model to learn the true function.
- An acquisition function to decide the next sample point.

Beyesian Optimization

Gaussian Process:

$$f(\mathbf{x}) \sim \mathrm{GP}[m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')] \tag{1}$$

$$\mathbb{E}[f(\mathbf{x})] = m(\mathbf{x}) \tag{2}$$

$$\mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))] = k(\mathbf{x}, \mathbf{x}') \tag{3}$$

For example:

$$m(\mathbf{x}) = 0 \tag{4}$$

$$k(\mathbf{x}, \mathbf{x}^{'}) = \exp\left[-\frac{1}{2}(\mathbf{x} - \mathbf{x}')^T(\mathbf{x} - \mathbf{x}')\right] \tag{5}$$

Gaussian Process

Given observations $\mathbf{f} = [f(\mathbf{x}_1), f(\mathbf{x}_2), ..., f(\mathbf{x}_t)]$ at t points, a new function value $f^* = f(\mathbf{x}^*)$ is jointly normally distributed with the observations $\mathbf{f}$,

$$Pr\left(\begin{bmatrix} \mathbf{f} \\ f^* \end{bmatrix}\right) = \text{Norm}\left[0, \begin{bmatrix} \mathbf{K}(\mathbf{X}, \mathbf{X}) & \mathbf{K}(\mathbf{X}, \mathbf{x}^*) \\ \mathbf{K}(\mathbf{x}^*, \mathbf{X}) & \mathbf{K}(\mathbf{x}^*, \mathbf{x}^*) \end{bmatrix}\right],$$

where $\mathbf{K}(\mathbf{X}, \mathbf{X})$ is a $t \times t$ matrix where element (i, j) is given by $k(\mathbf{x_i}, \mathbf{x_j})$, $\mathbf{K}(\mathbf{X}, \mathbf{x}^*)$ is a $t \times 1$ vector where element i is given by $k(\mathbf{x_i}, \mathbf{x}^*)$ and so on. The conditional distribution $Pr(f^*|\mathbf{f})$ must also be normal,

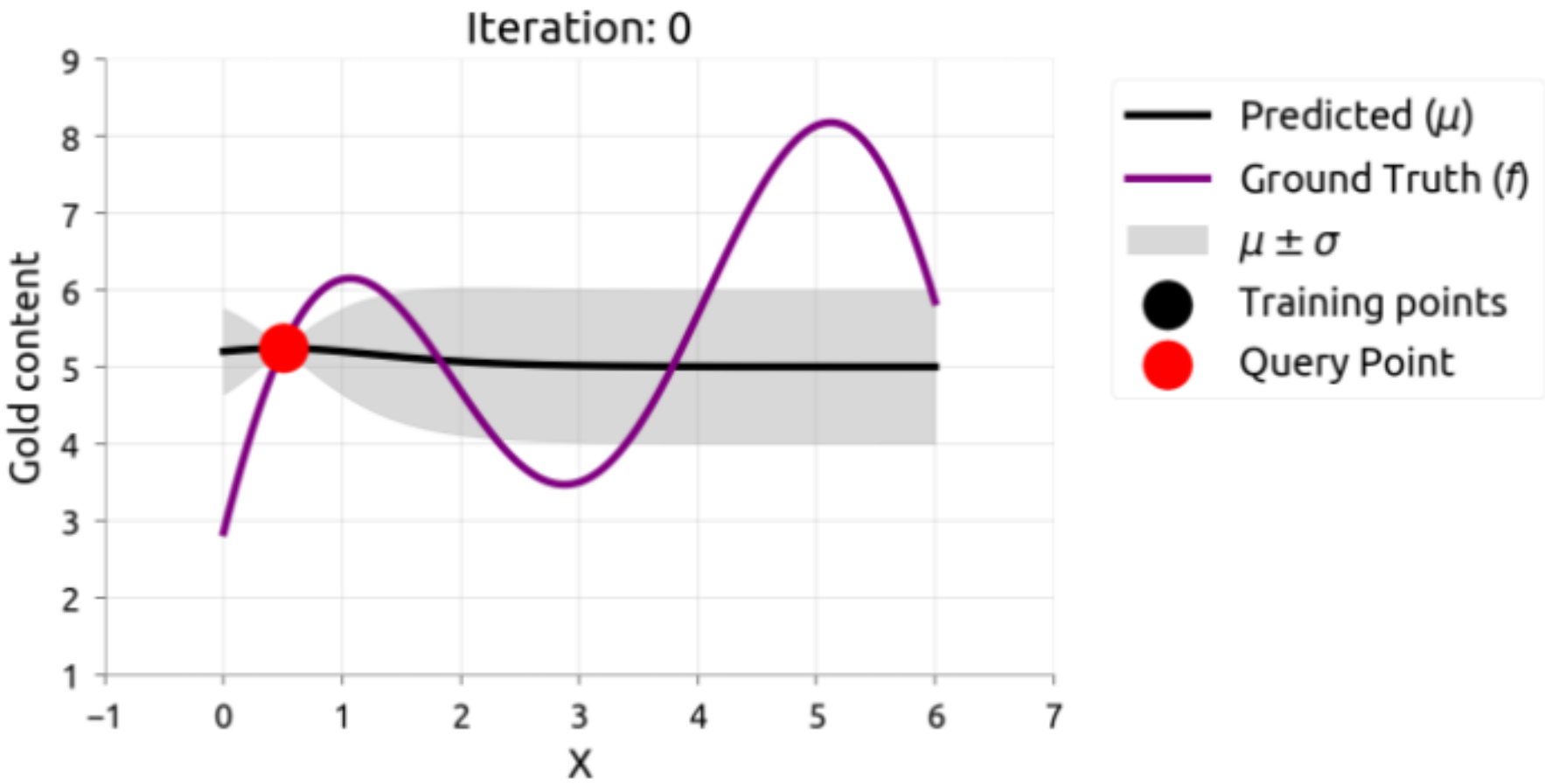
$$Pr(f^*|\mathbf{f}) = \text{Norm}[\mu(\mathbf{x}^*), \sigma^2(x^*)], \tag{6}$$

where

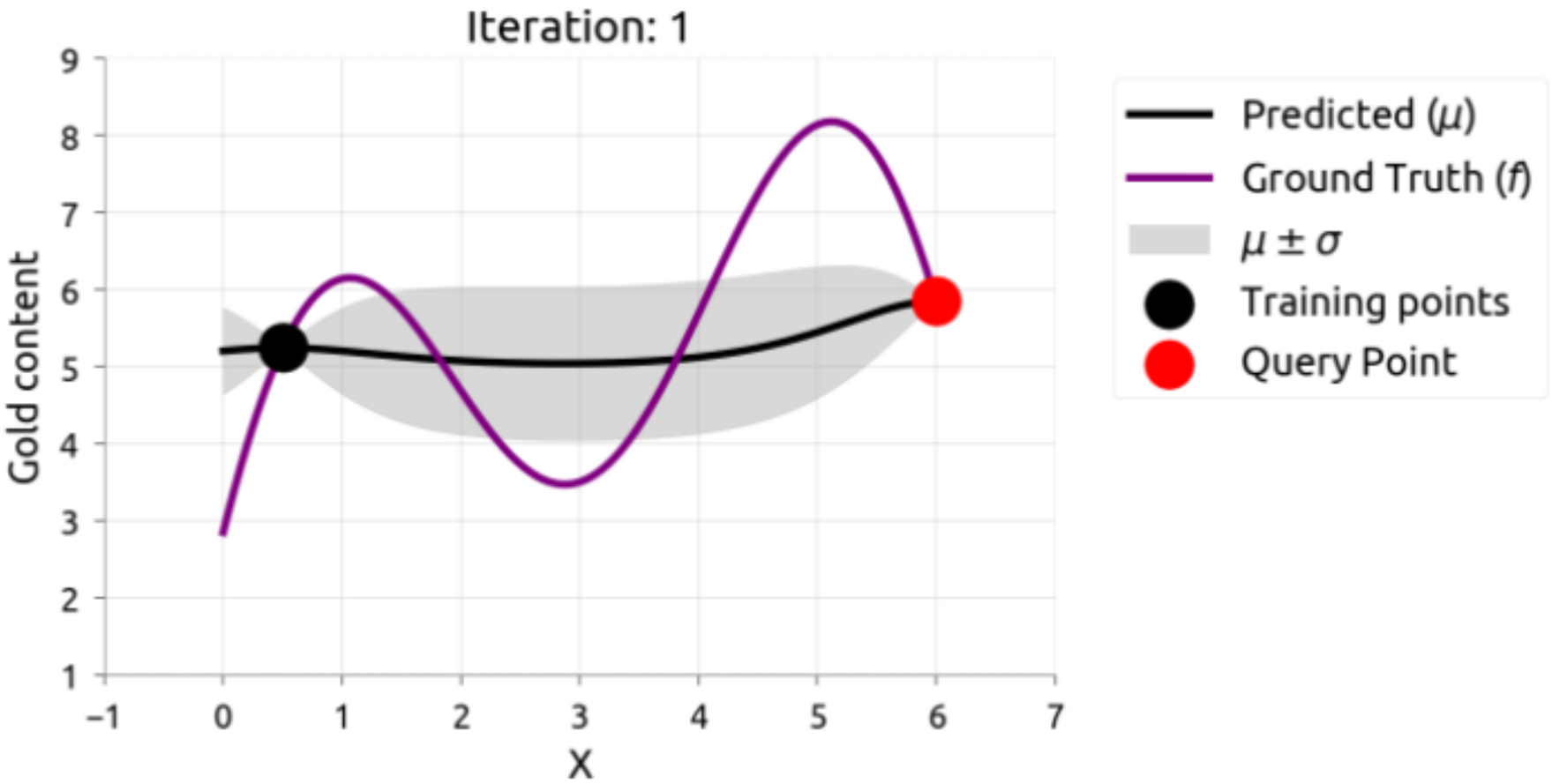
$$\mu(\mathbf{x}^*) = \mathbf{K}(\mathbf{x}^*, \mathbf{X})\mathbf{K}(\mathbf{X}, \mathbf{X})^{-1}\mathbf{f} \tag{7}$$

$$\sigma^2(x^*) = \mathbf{K}(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{K}(\mathbf{x}^*, \mathbf{X})\mathbf{K}(\mathbf{X}, \mathbf{X})^{-1}\mathbf{K}(\mathbf{X}, \mathbf{x}^*) \tag{8}$$

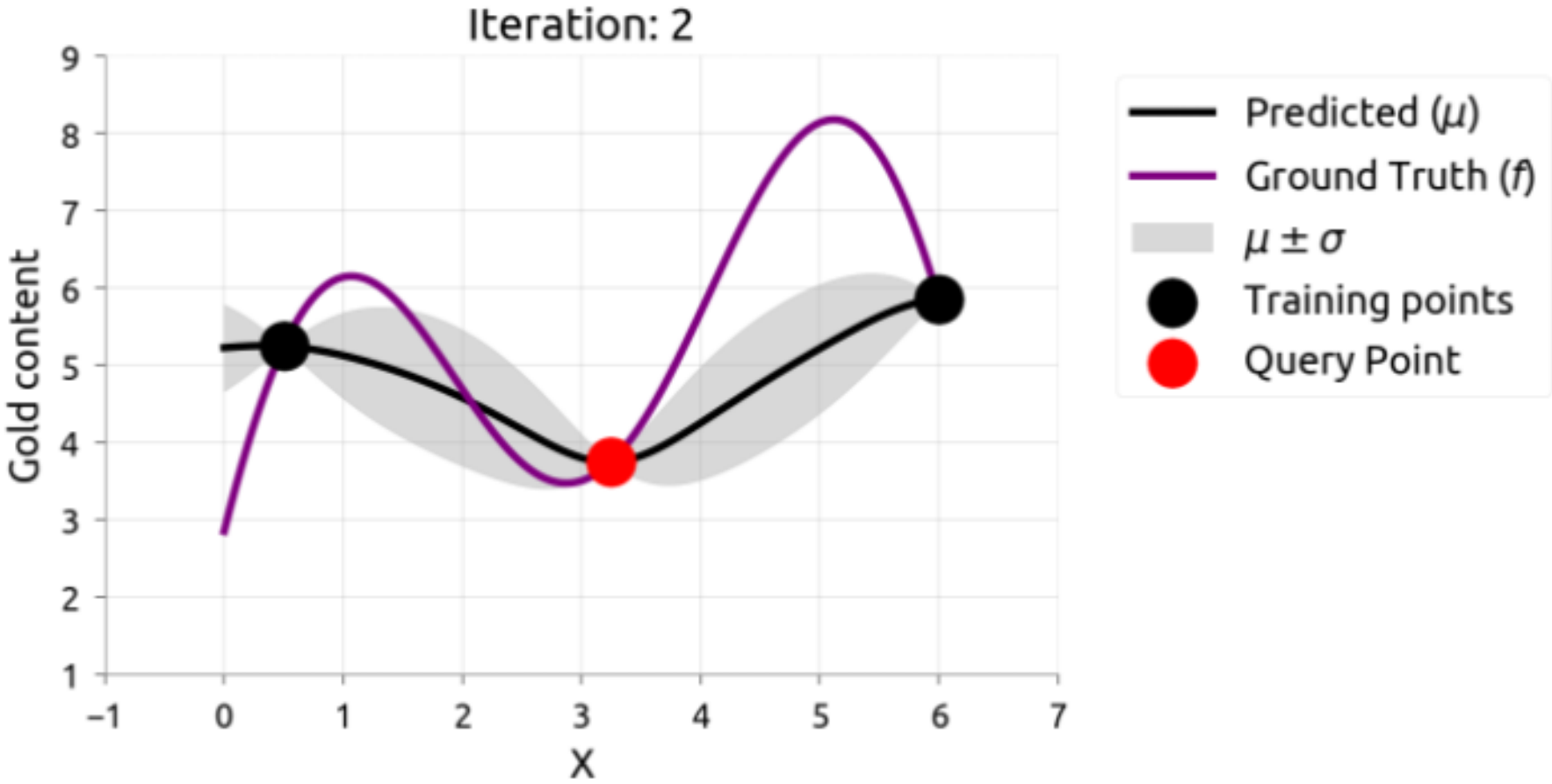
Gaussian Process



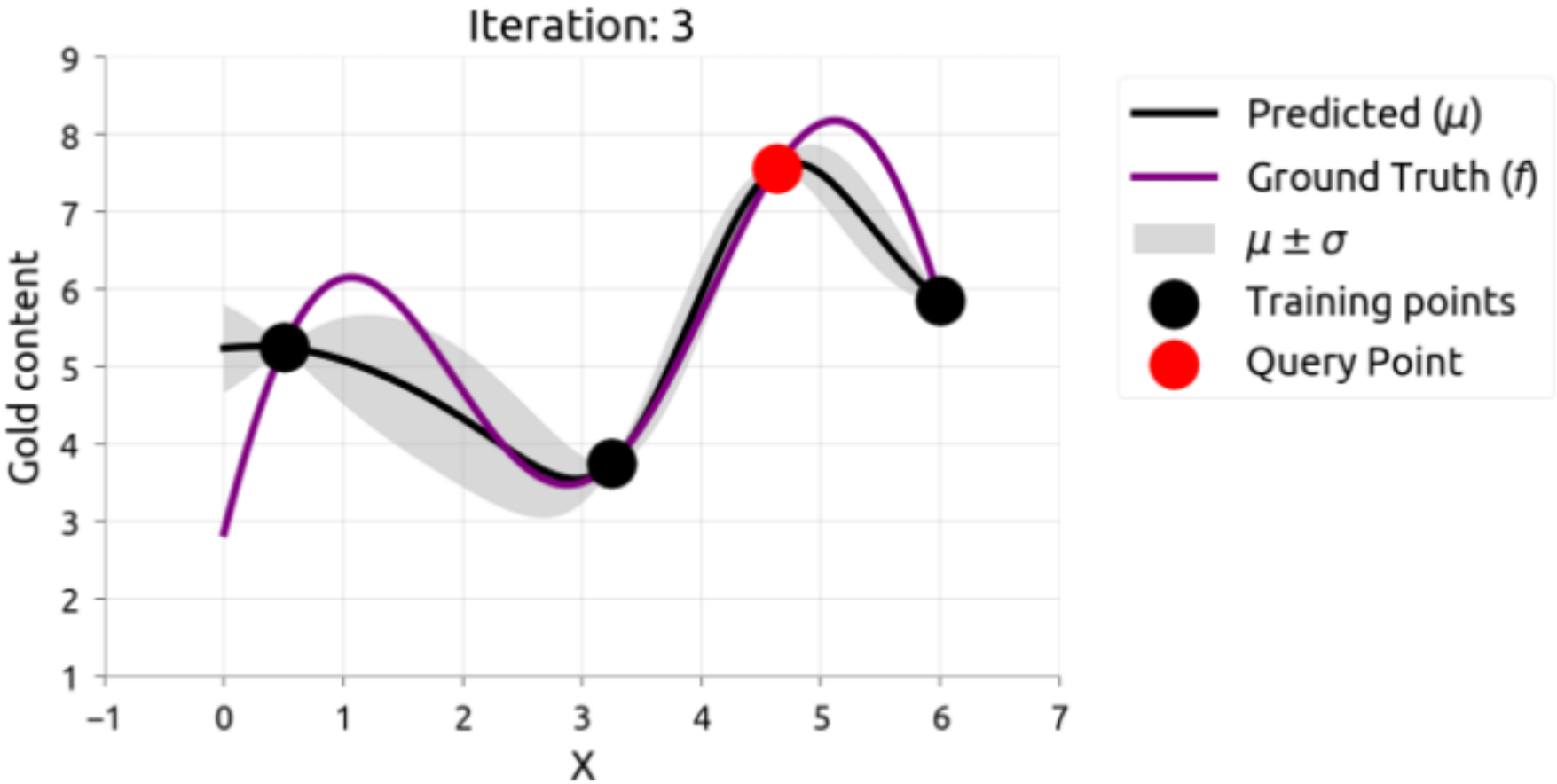
Gaussian Process



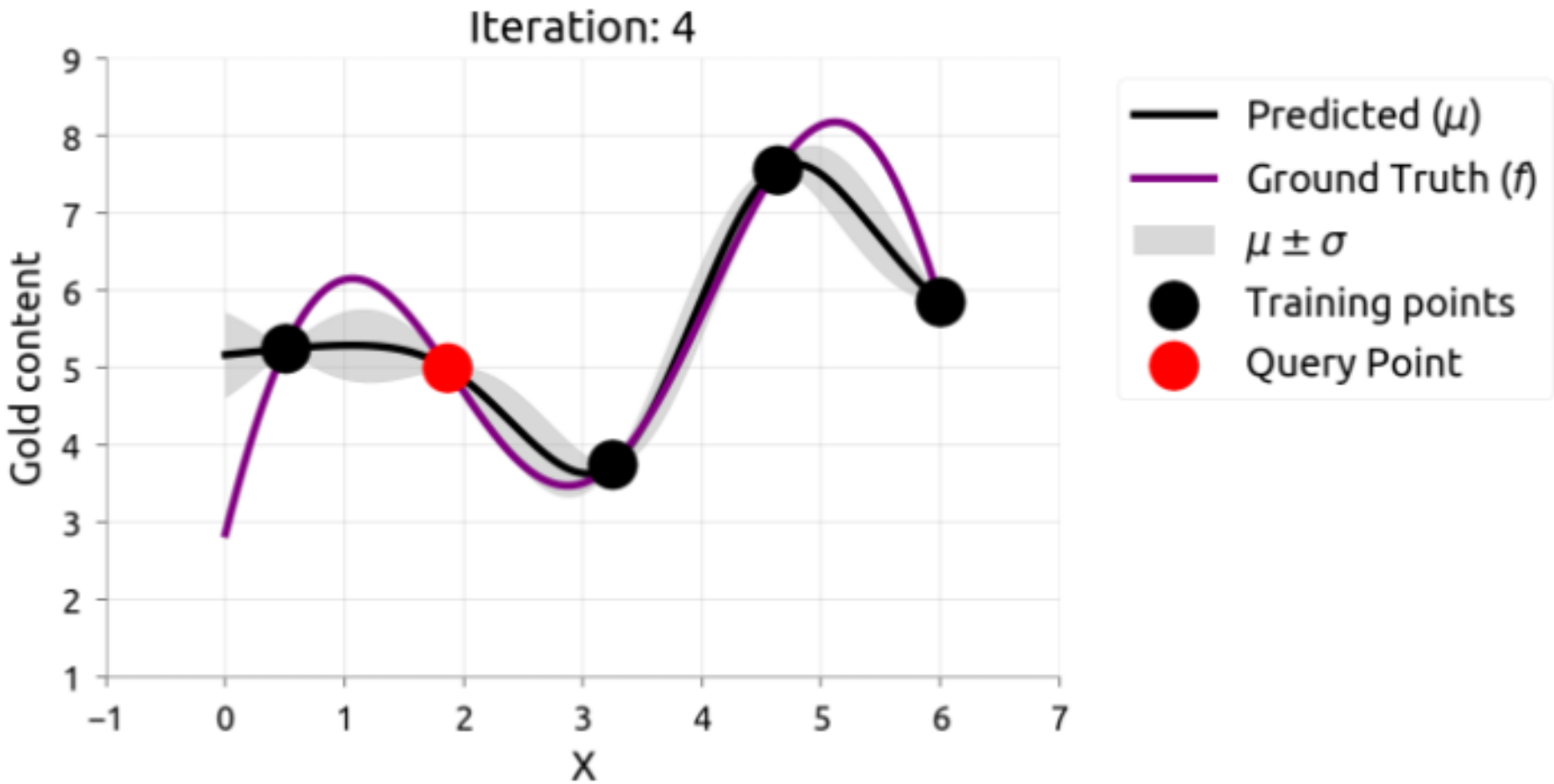
Gaussian Process



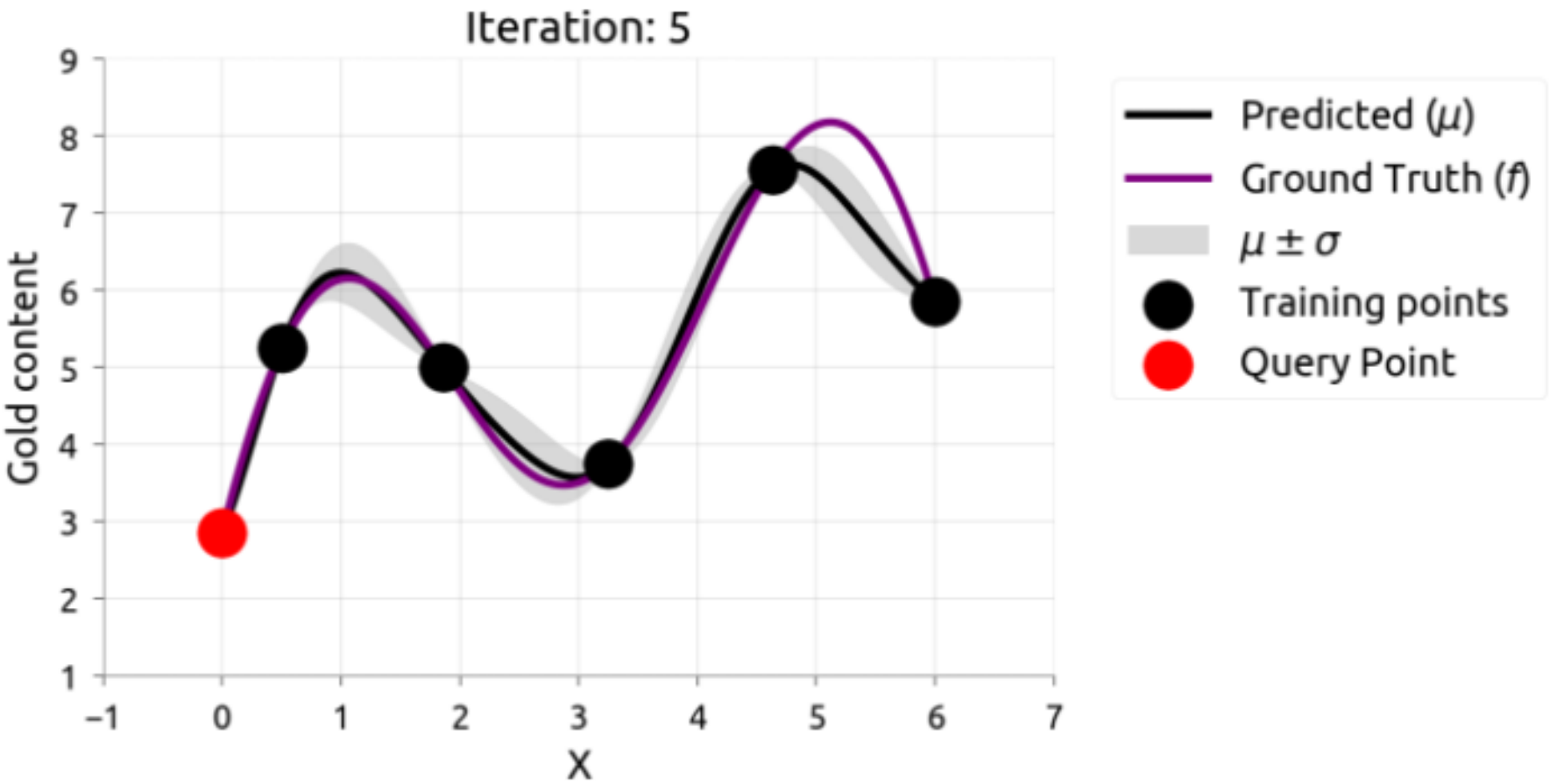
Gaussian Process



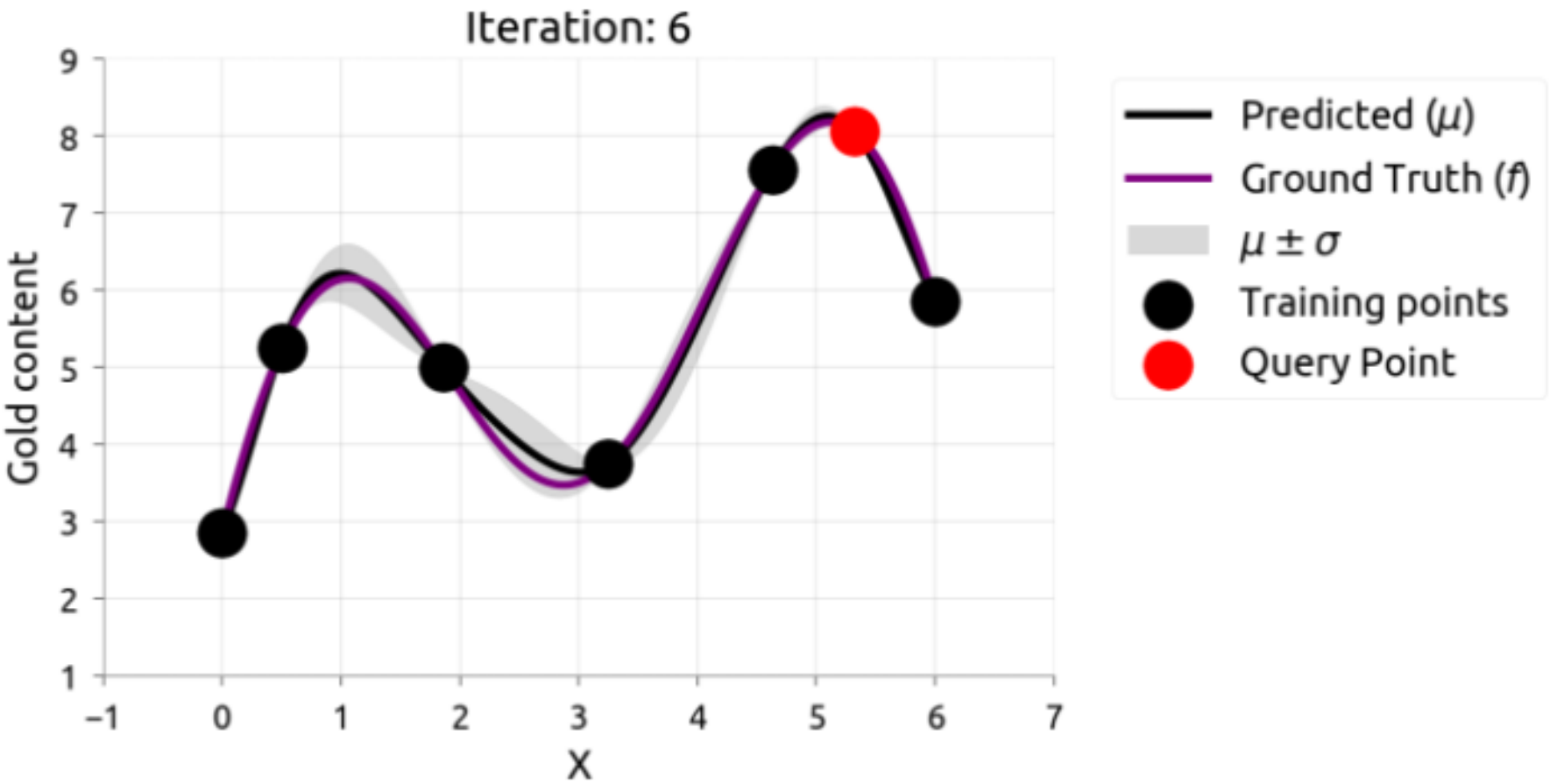
Gaussian Process



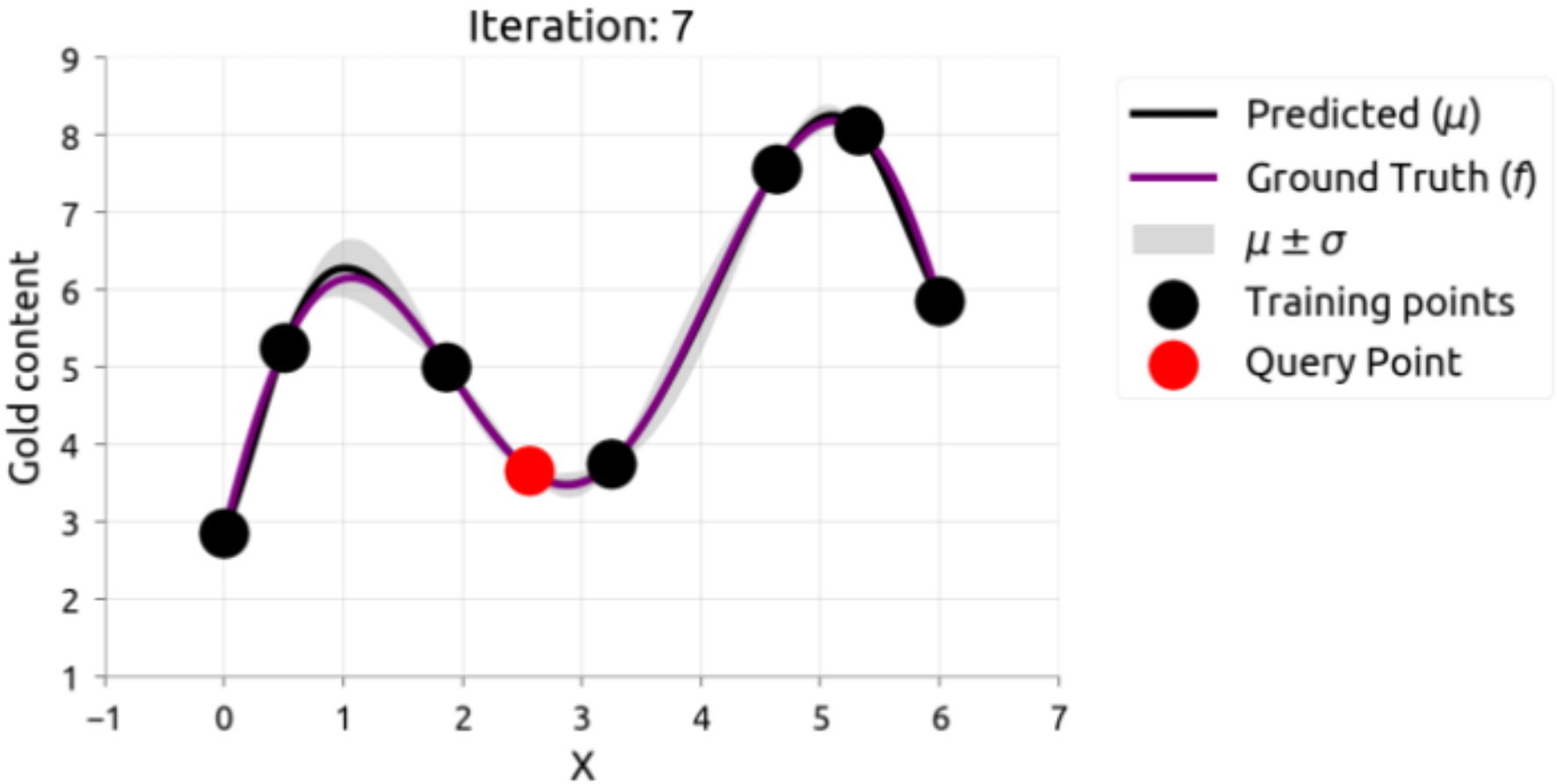
Gaussian Process



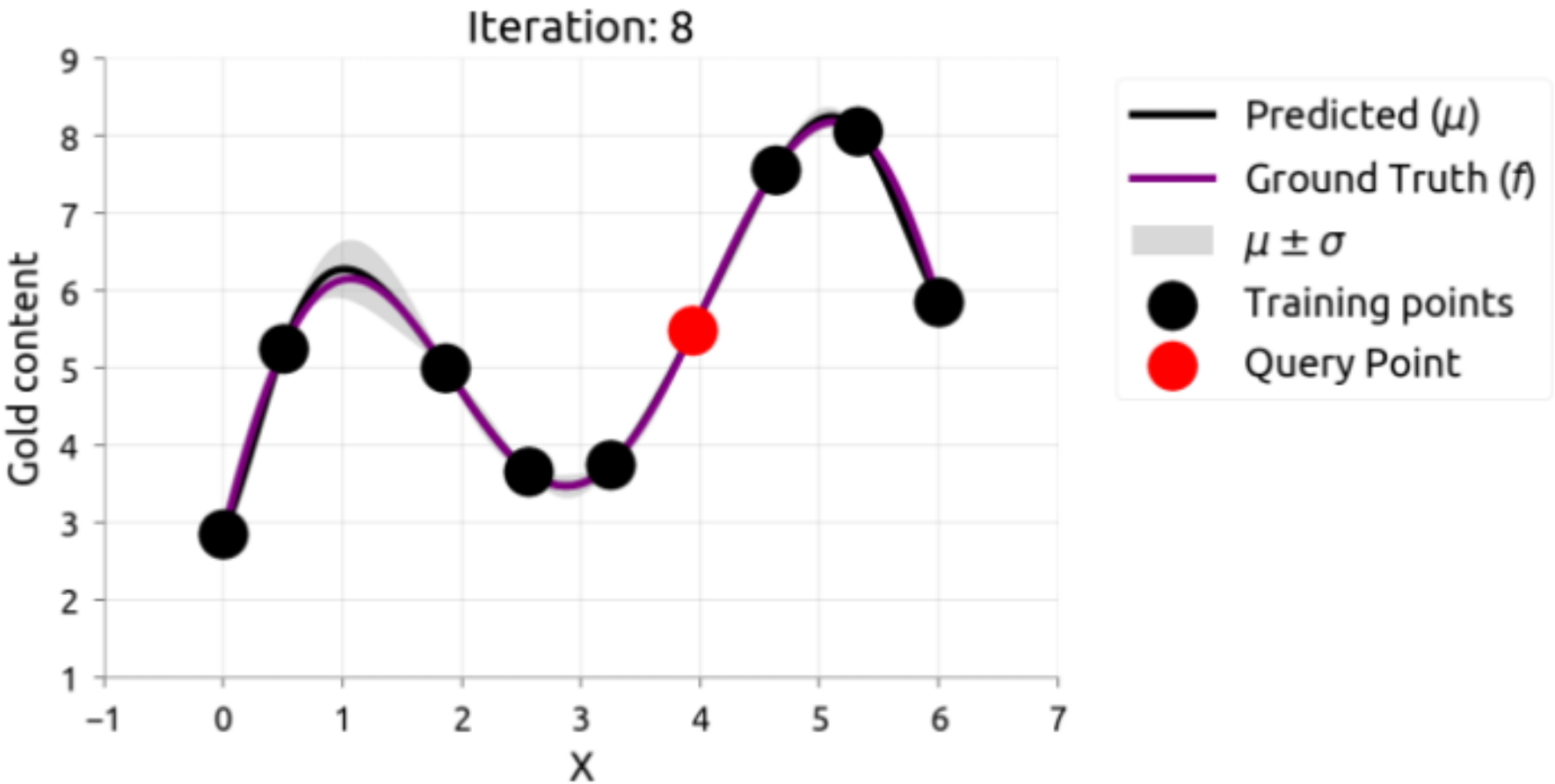
Gaussian Process



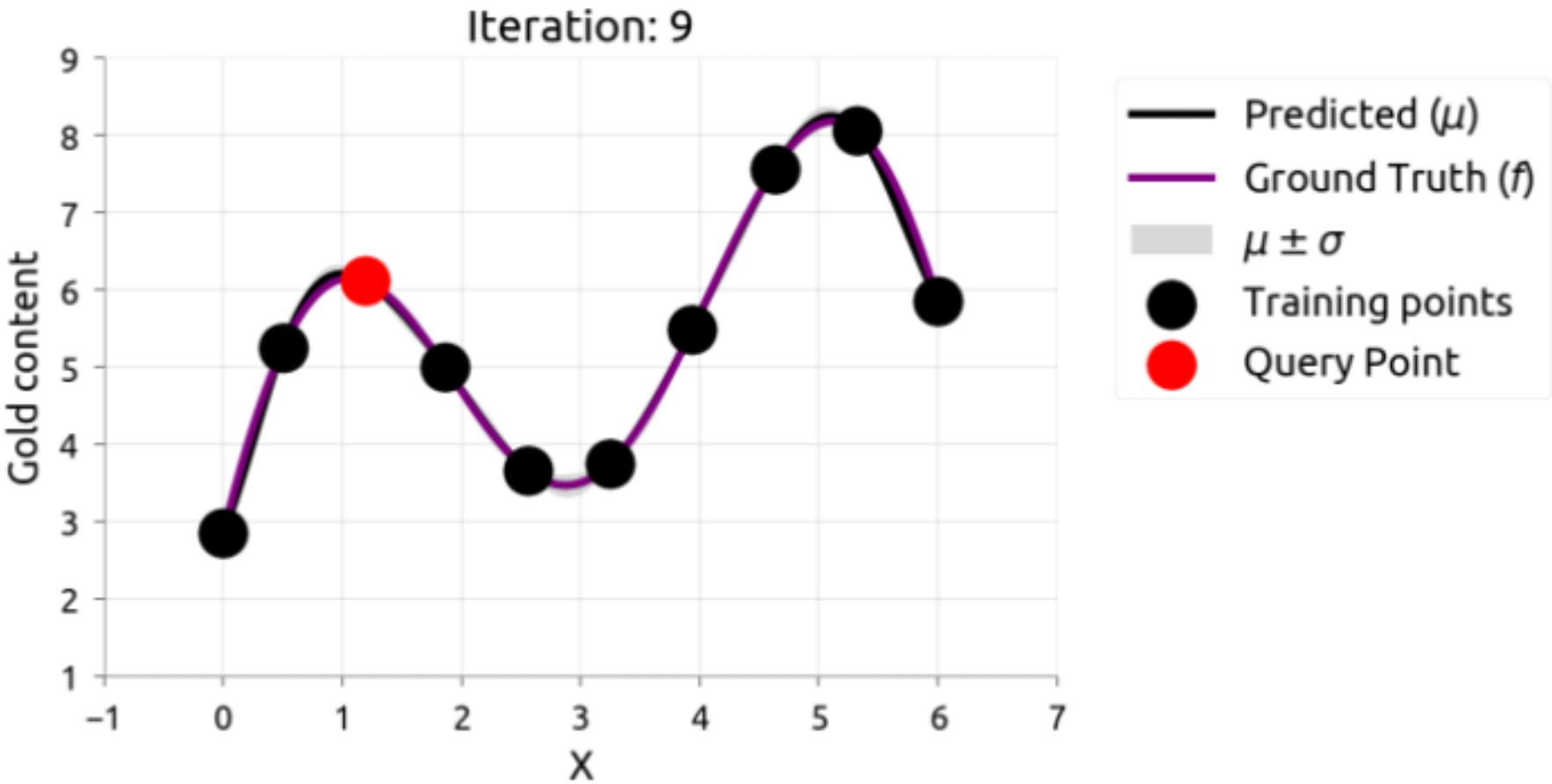
Gaussian Process



Gaussian Process



Gaussian Process

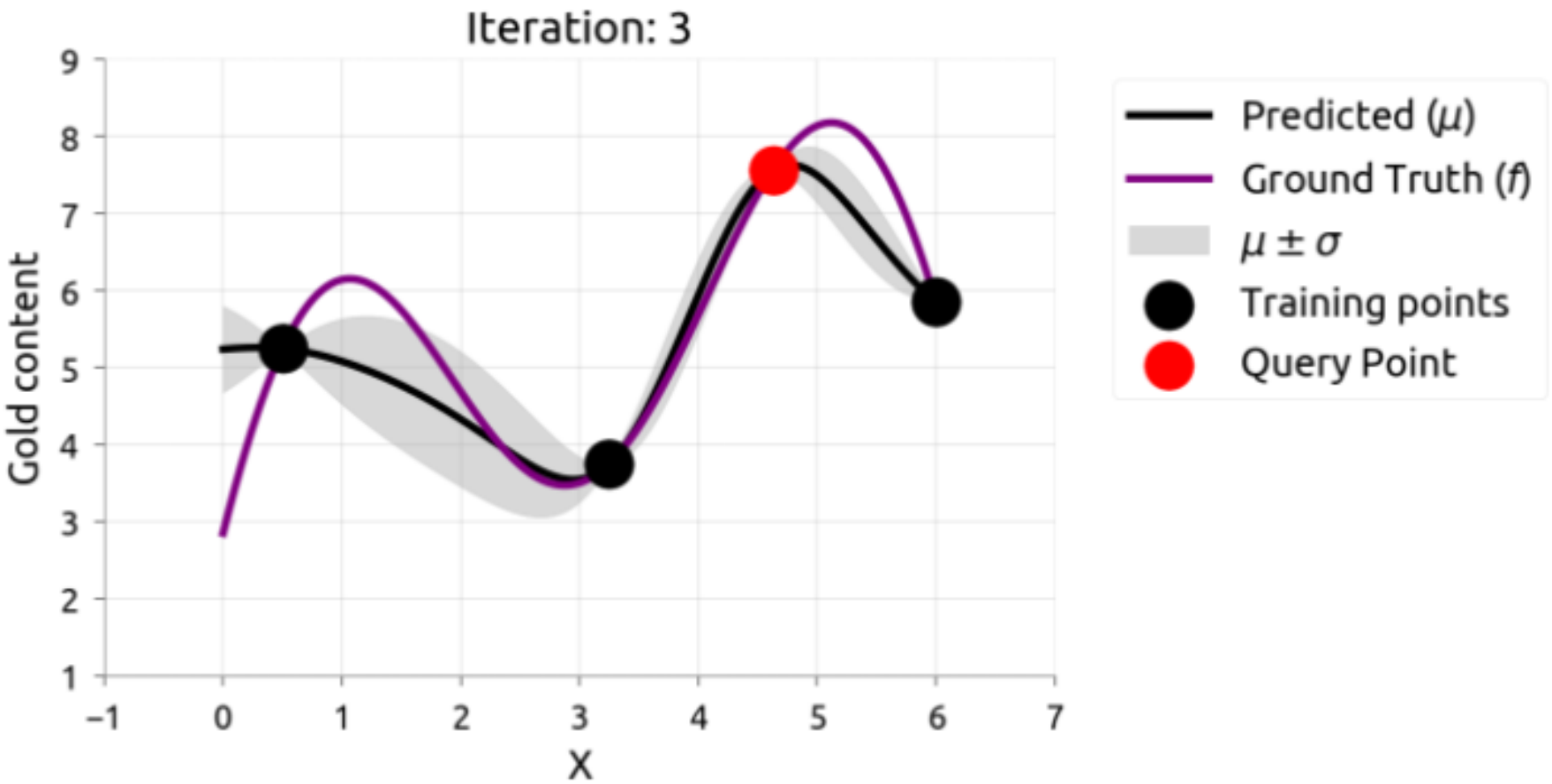


Acquisition Function

Our goal is simply to find the location of maximum gold content, why should we waste our precious time to query the location of low gold content?
The core question in Bayesian Optimization:"Based on what we know so far, which point should we evaluate next?" The decision is made with the help of acquisition function $\alpha(x)$.

Acquisition Function

Intuitively, there are two types of location deserving investigating: a location has high expectation value or high variance.



Acquisition Function

Examples of acquisition function :

- Thompson sampling (Thompson, 1933)
- Probability of improvement (Kushner, 1964)
- Expected improvement (Mokus, 1975)
- Upper confidence bound (Srinivas et al., 2010)
- Entropy search (Villemonteix et al., 2009, Hennig and Schuler, 2012)
- Knowledge gradient (Wu et al., 2017)

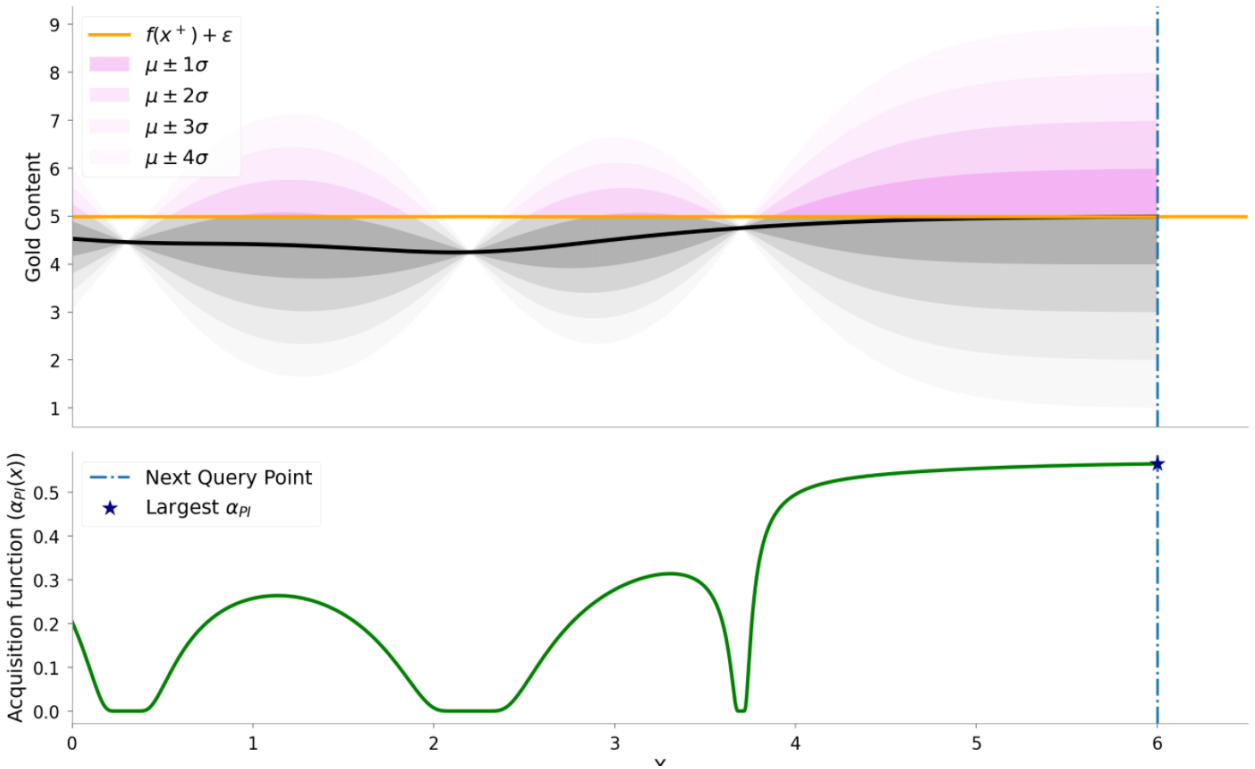
These acquisition functions help us to trade off the exploration and exploitation.

Probability of improvement (PI)

Mathematically, PI can be formulated as

$$\mathbf{x}_{t+1} = \text{argmax}[\text{PI}(\mathbf{X})] = \text{argmax}[Pr(f(x) \geq (f(x^+) + \epsilon))], \qquad (9)$$

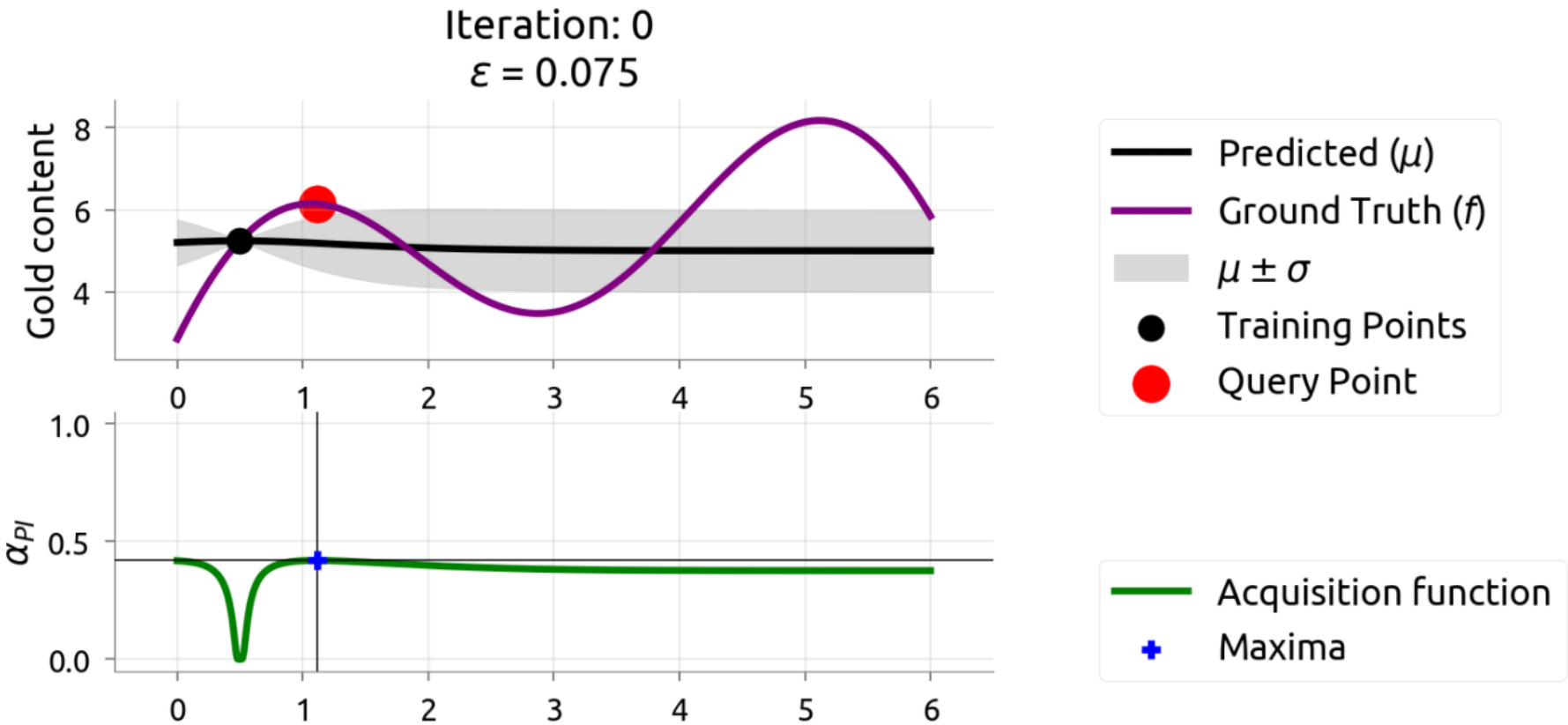
where ϵ is a small positive number, and $x^+ = \text{argmax}_{x_i \in x_{1:t}} f(x_i)$ where x_i is the location queried at i^{th} time step.



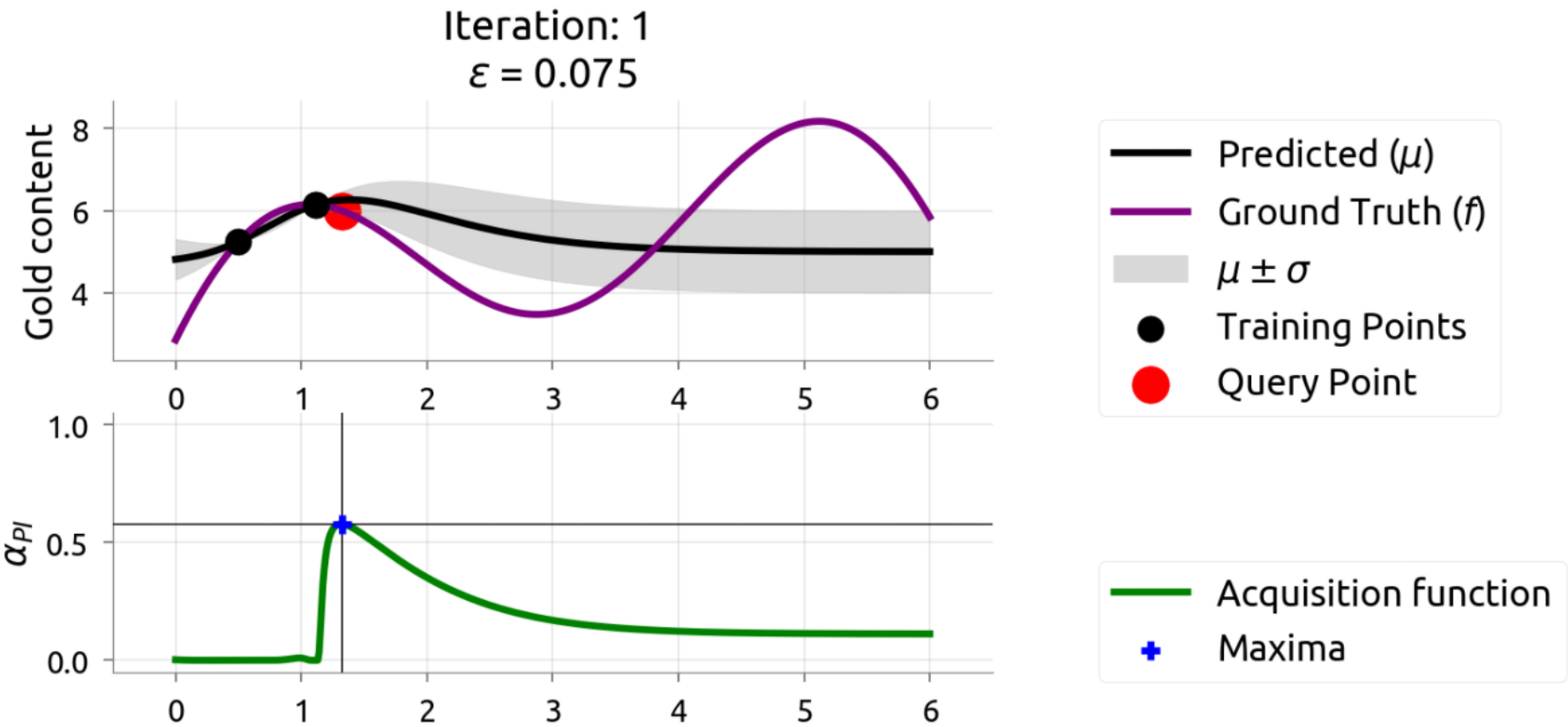
Effect of ϵ

PI uses ϵ to strike a balance between exploration and exploitation.
Increasing ϵ results in querying locations with a larger σ.

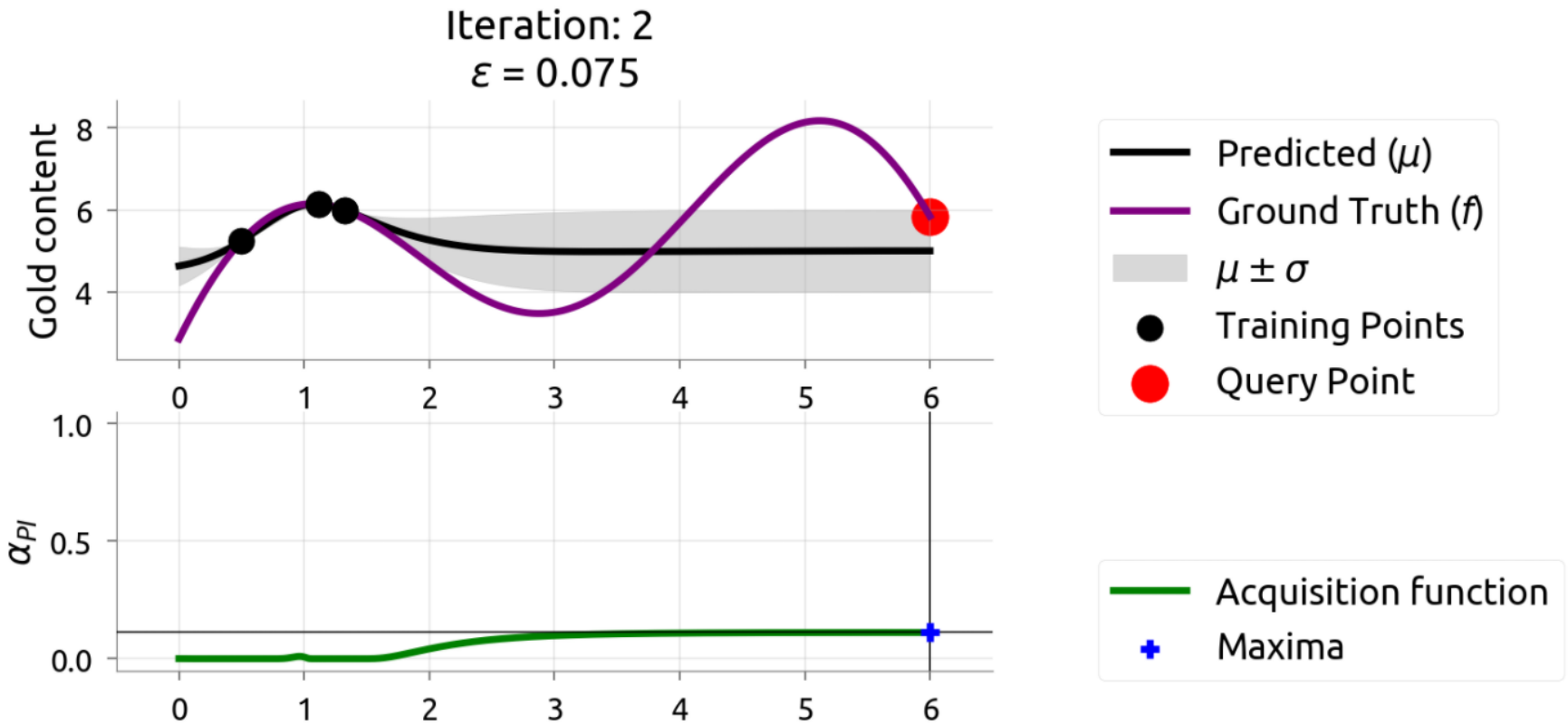
$\epsilon = 0.075$



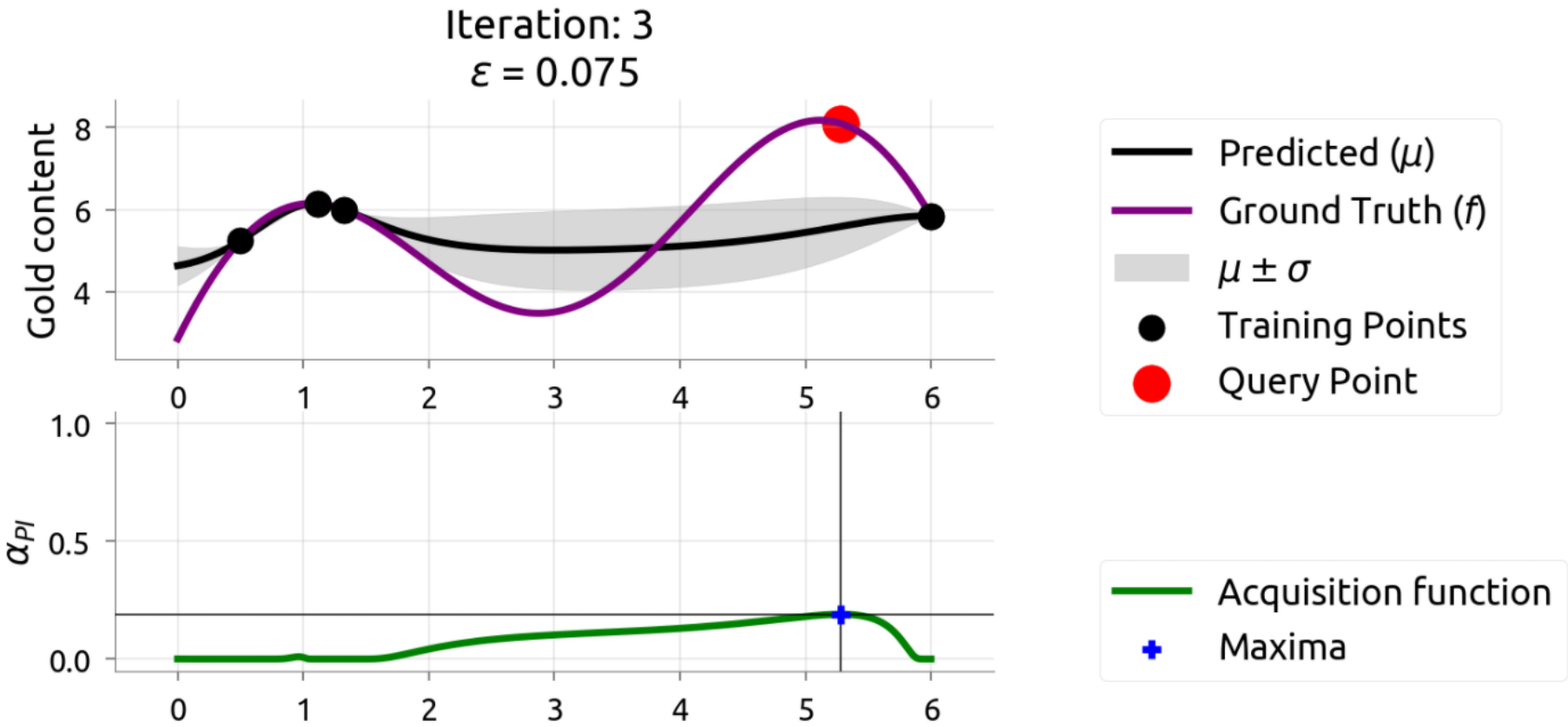
$\epsilon = 0.075$



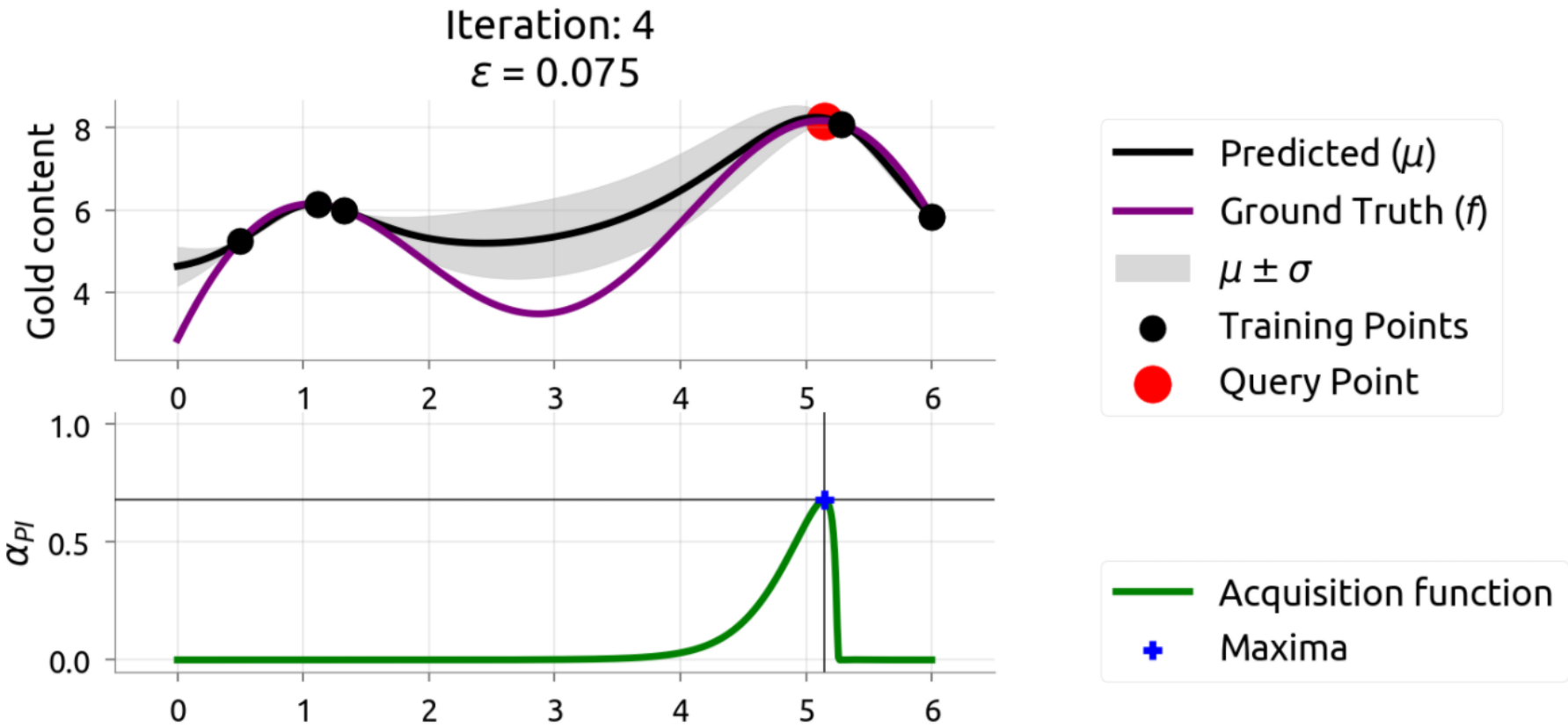
$\epsilon = 0.075$



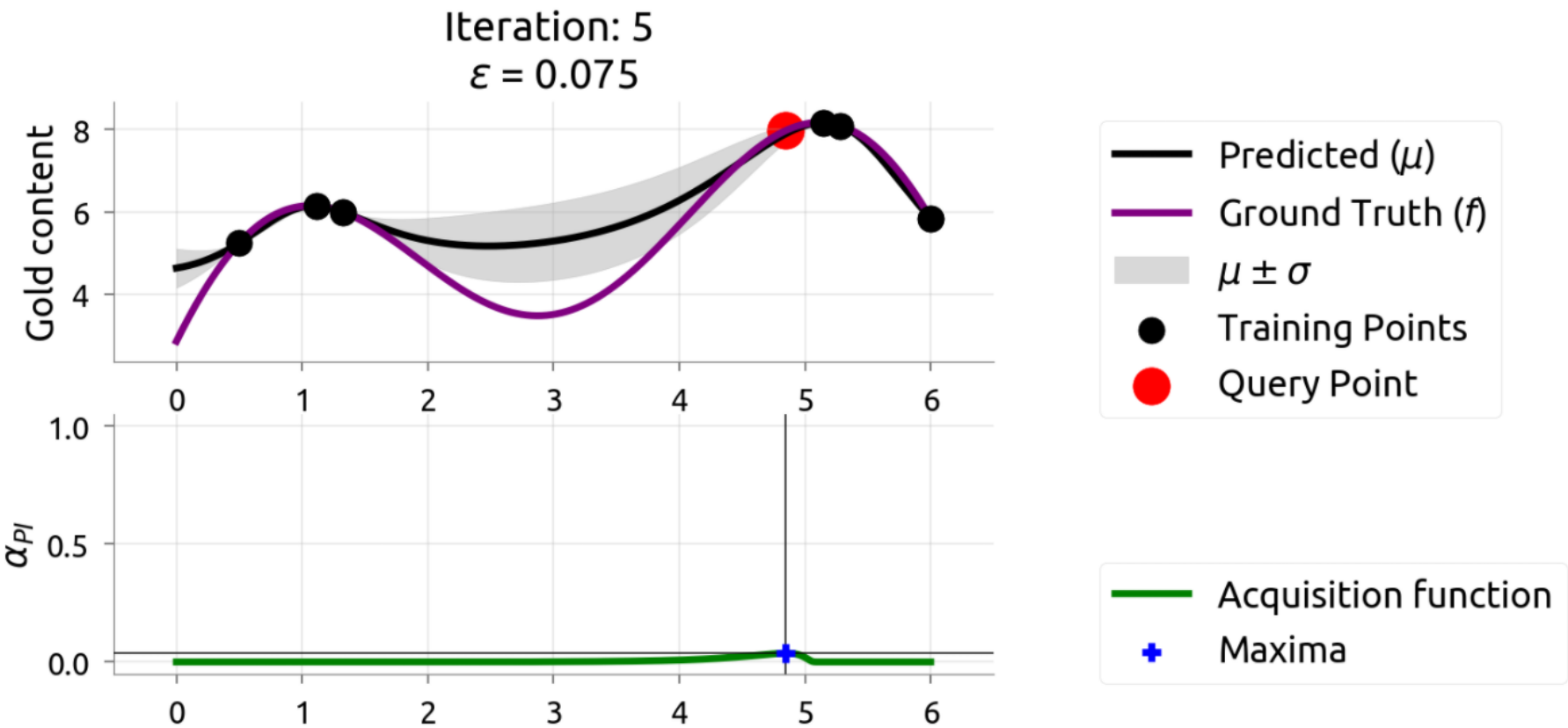
$\epsilon = 0.075$



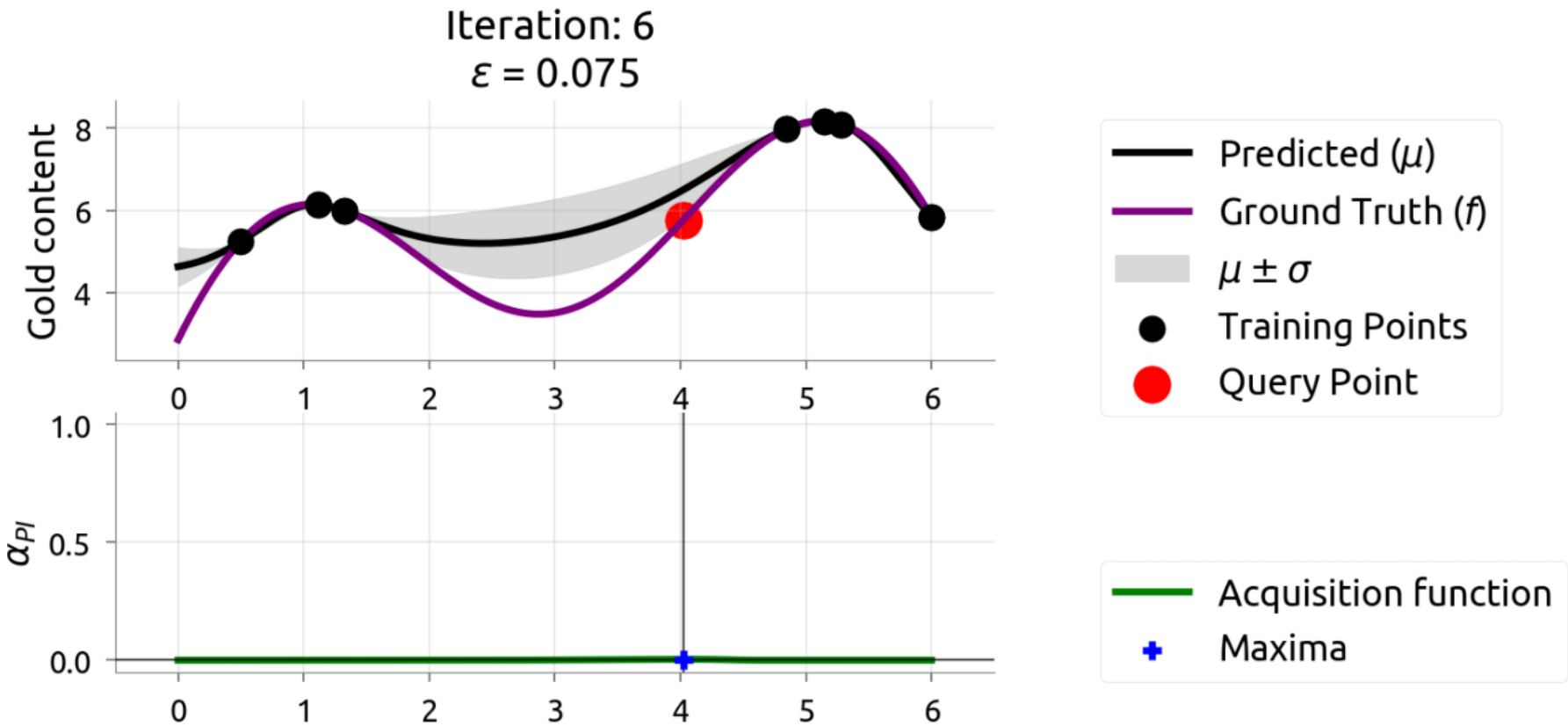
$\epsilon = 0.075$



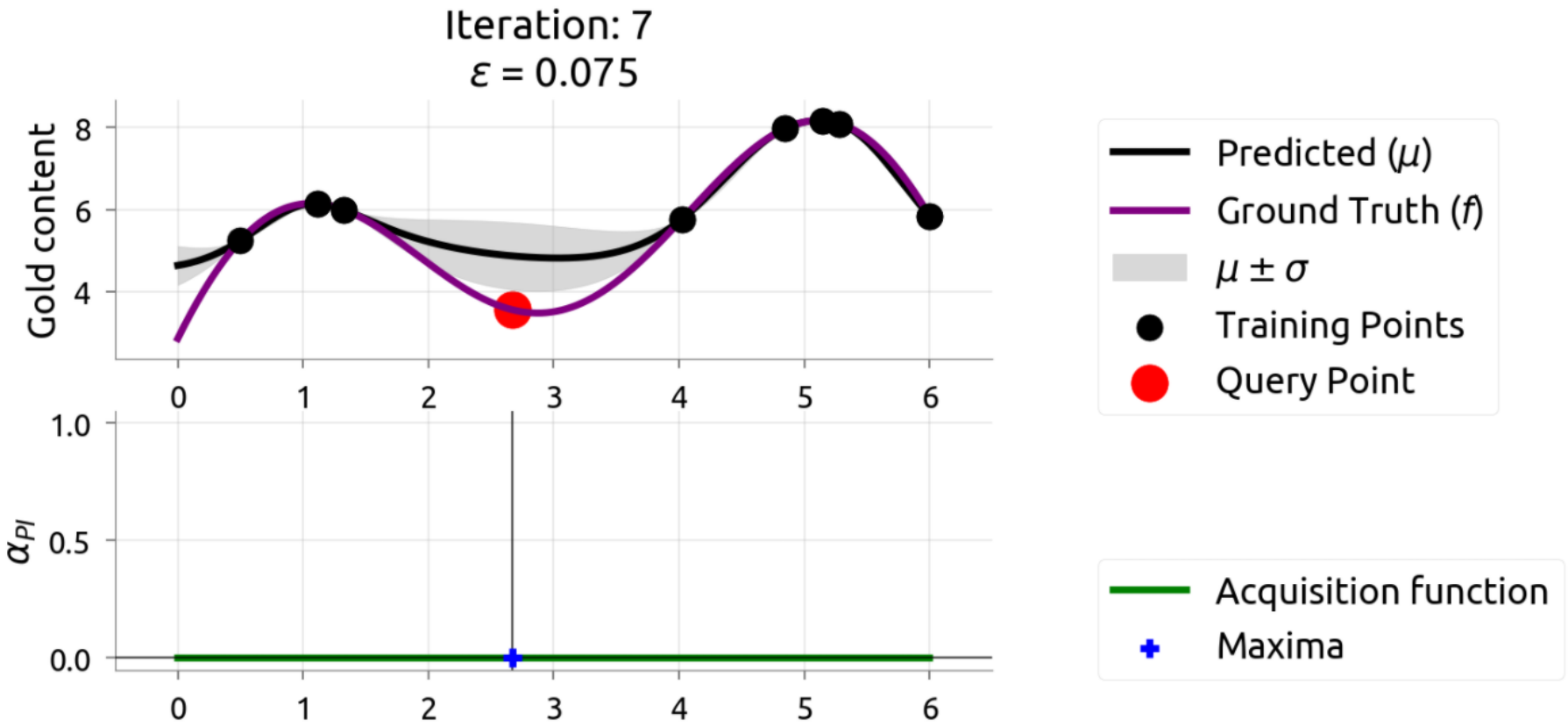
$\epsilon = 0.075$



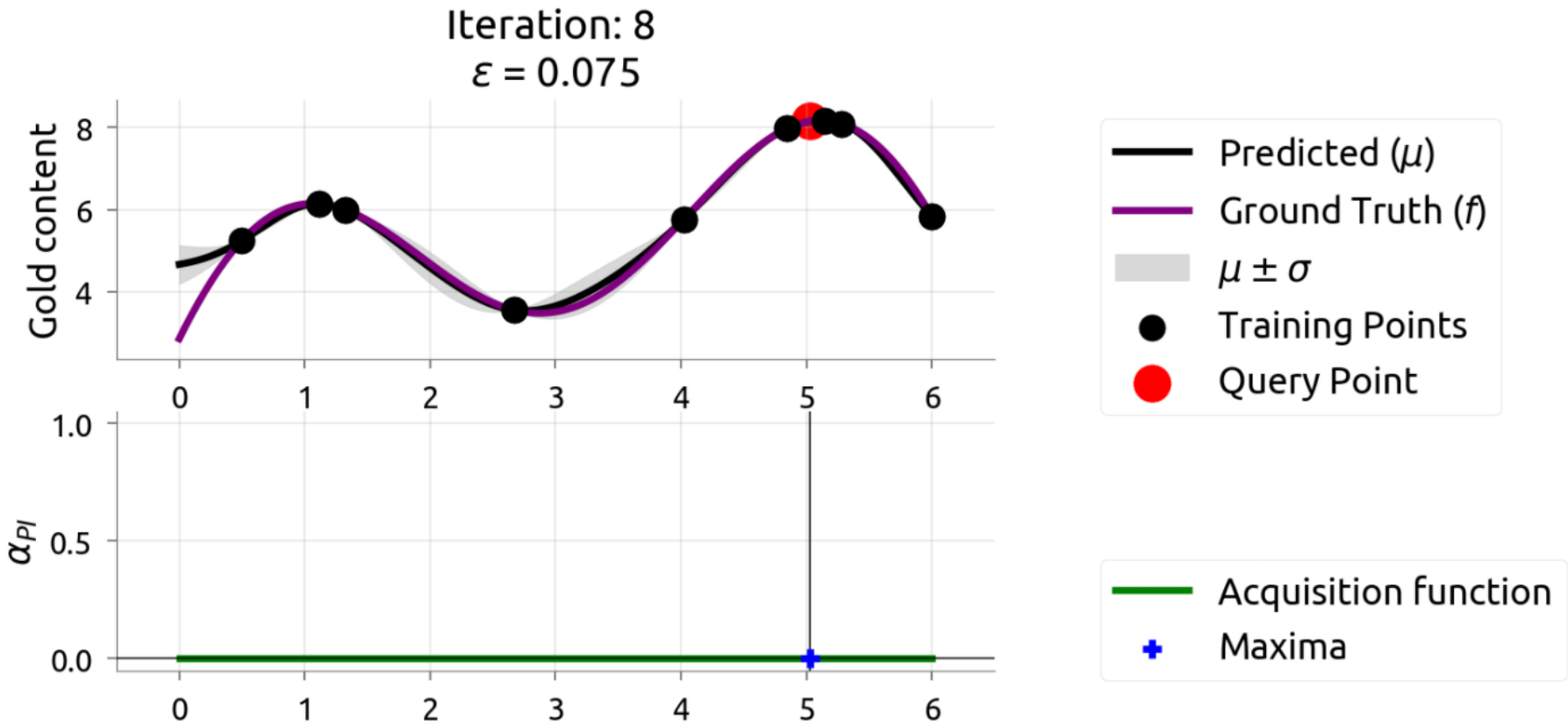
$\epsilon = 0.075$



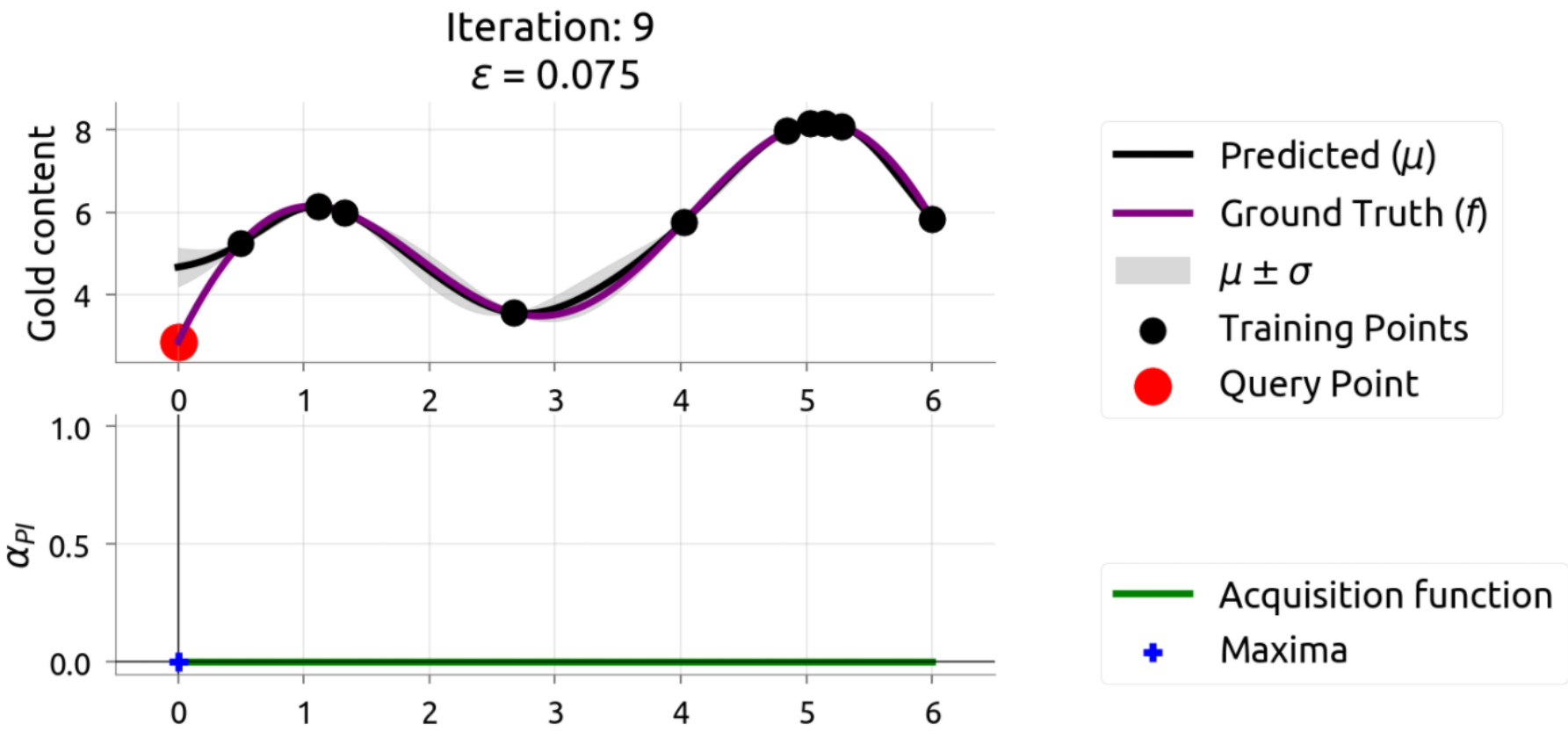
$\epsilon = 0.075$



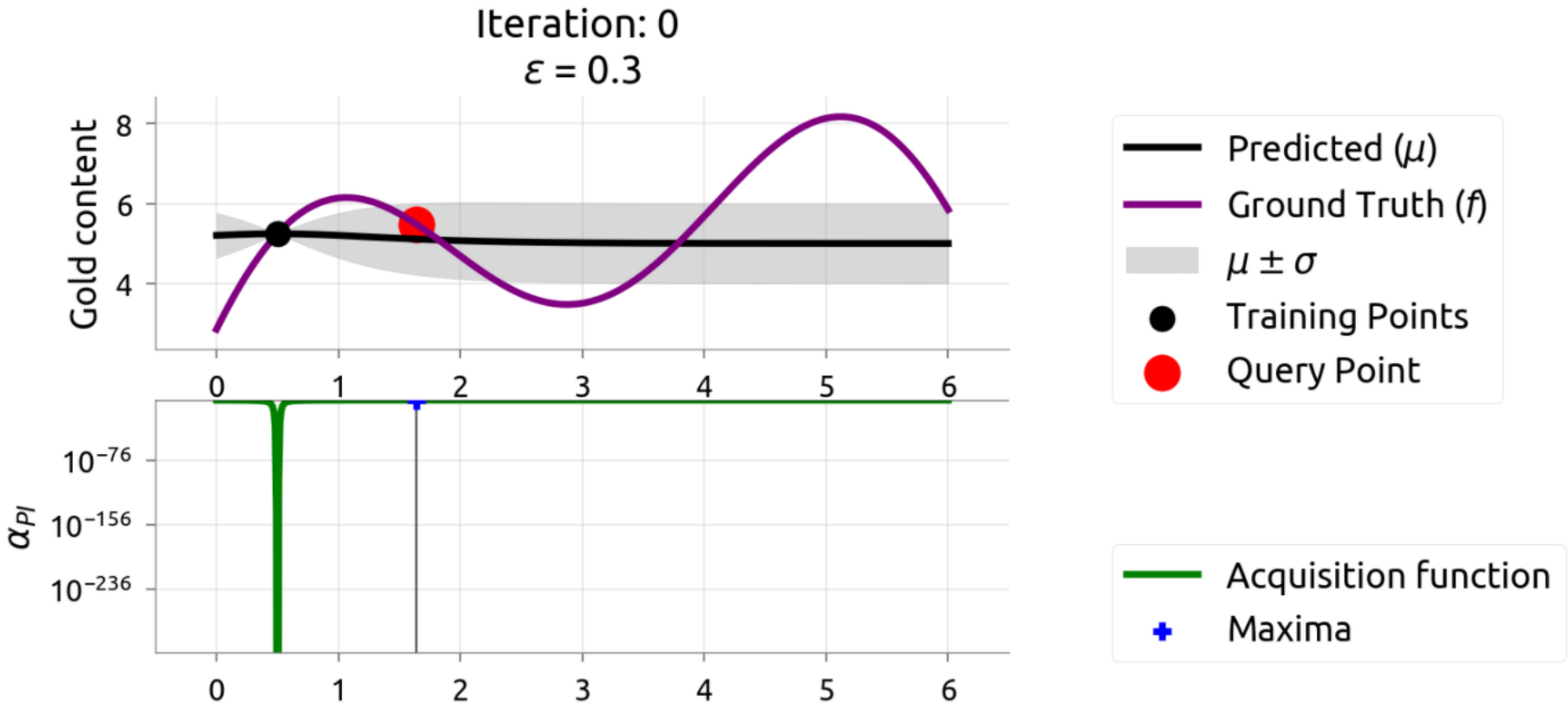
$\epsilon = 0.075$



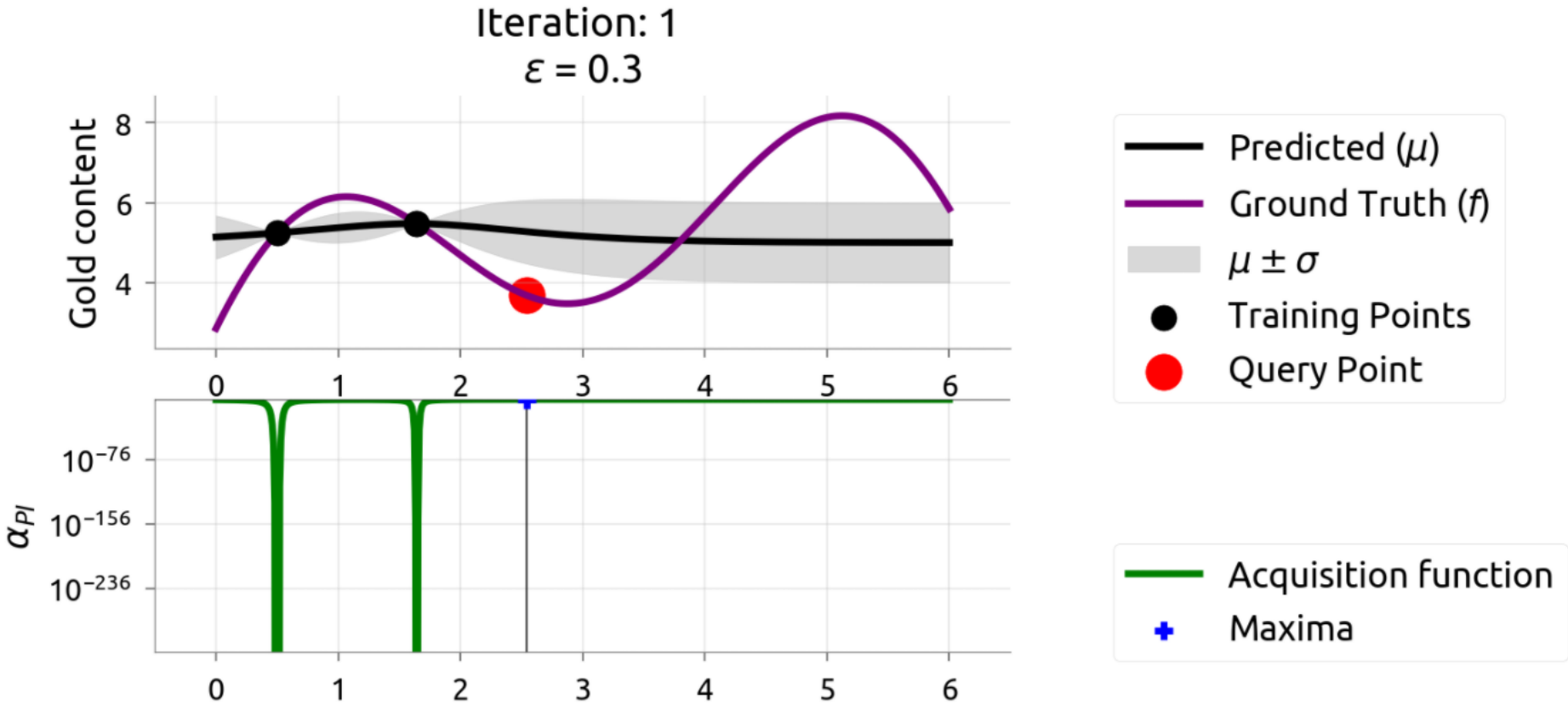
$\epsilon = 0.075$



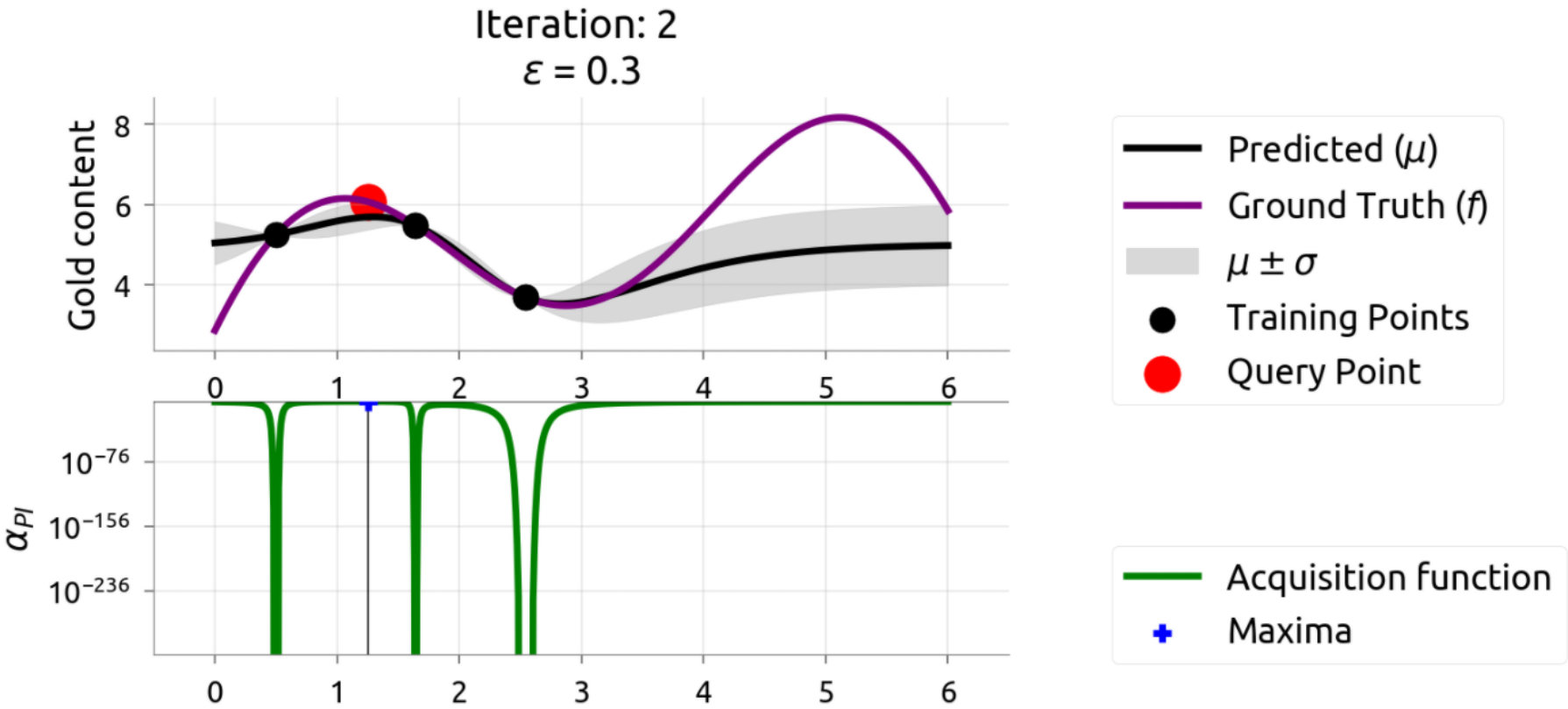
$\epsilon = 0.3$


Iteration: 0
ϵ = 0.3
Gold content
8
6
4
0
1
2
3
4
5
6
α_{PI}
10^{-76}
10^{-156}
10^{-236}
0
1
2
3
4
5
6
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

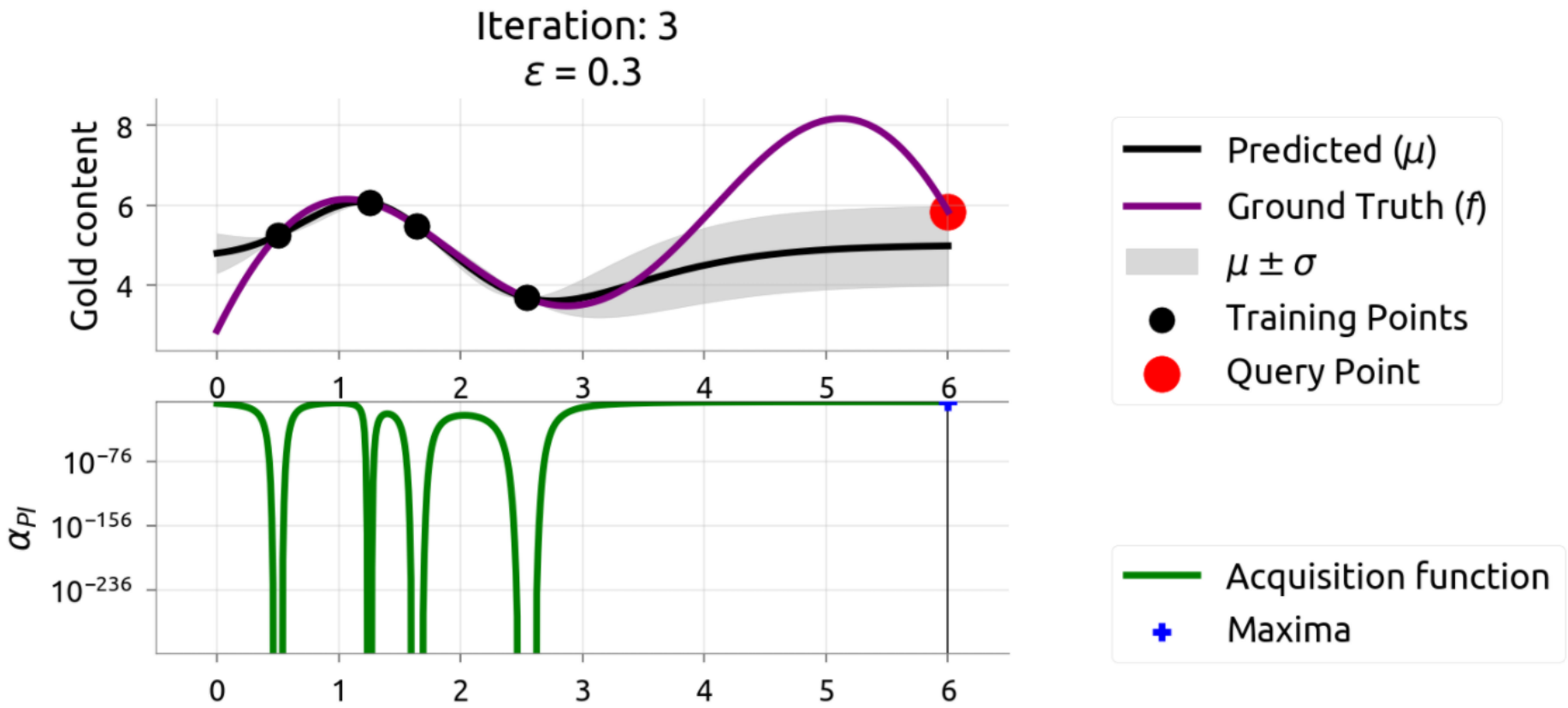
$\epsilon = 0.3$


Iteration: 1
ϵ = 0.3
Gold content
8
6
4
0
1
2
3
4
5
6
α_{PI}
10^{-76}
10^{-156}
10^{-236}
0
1
2
3
4
5
6
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

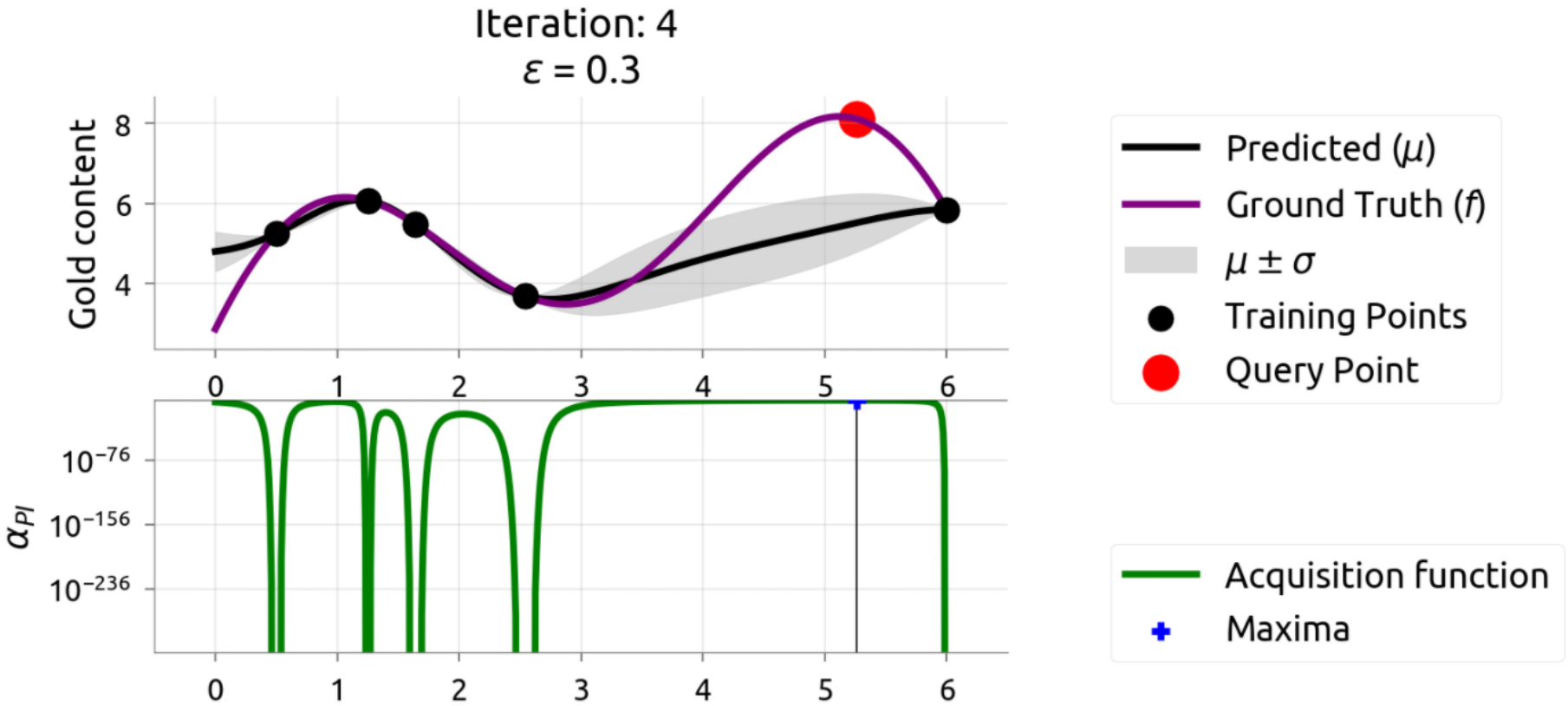
$\epsilon = 0.3$


Iteration: 2
ϵ = 0.3
Gold content
8
6
4
0
1
2
3
4
5
6
α_{PI}
10^{-76}
10^{-156}
10^{-236}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

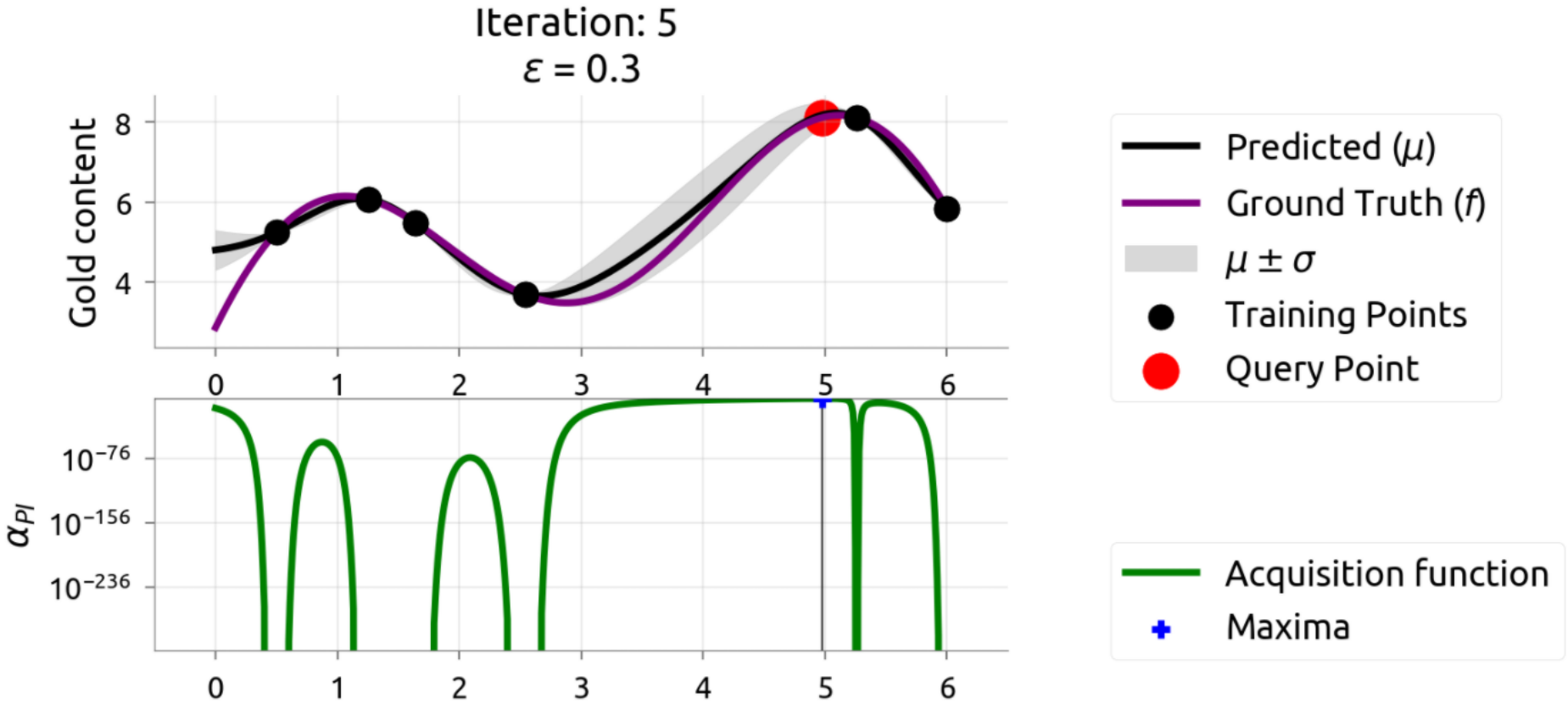
$\epsilon = 0.3$


Iteration: 3
ϵ = 0.3
Gold content
8
6
4
0
1
2
3
4
5
6
α_{PI}
10^{-76}
10^{-156}
10^{-236}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

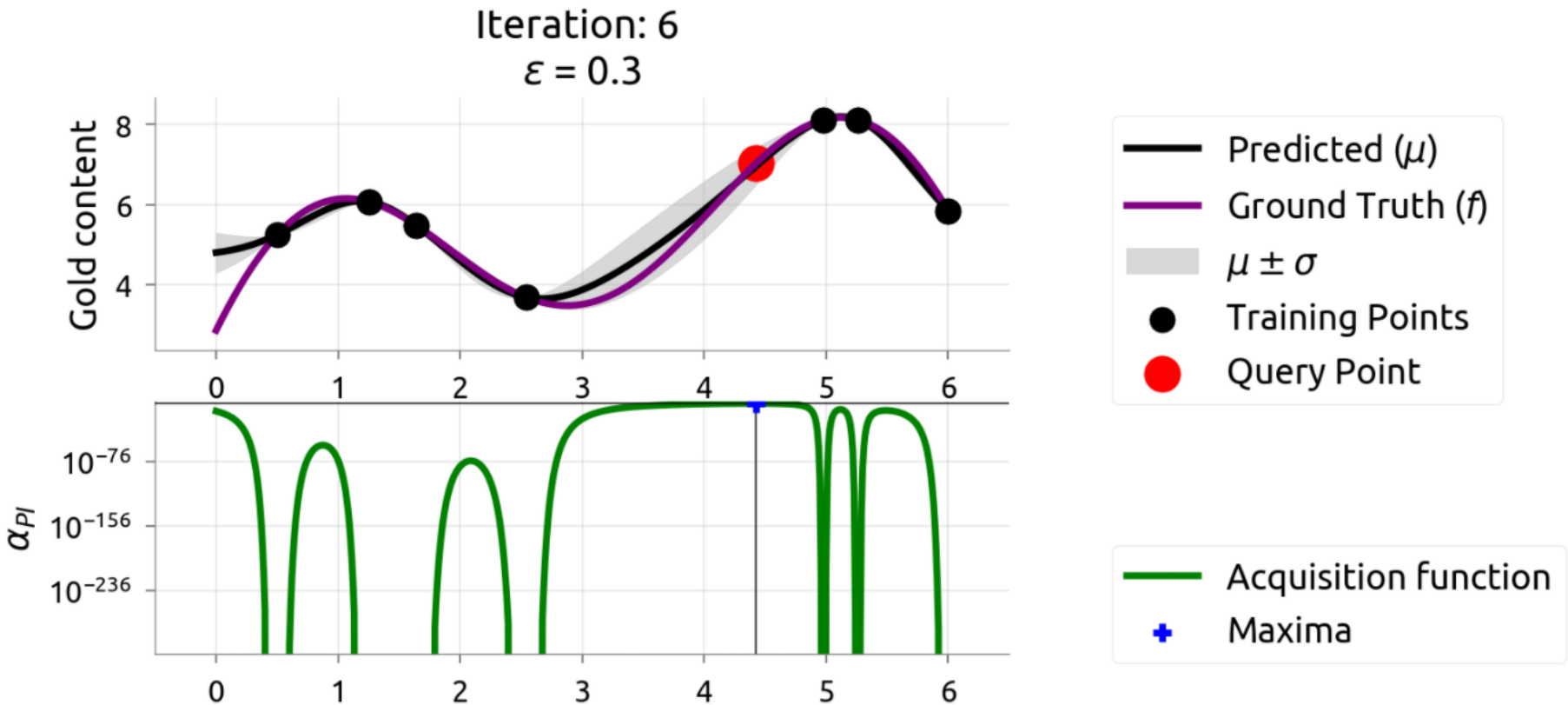
$\epsilon = 0.3$


Iteration: 4
ϵ = 0.3
Gold content
8
6
4
0
1
2
3
4
5
6
α_{PI}
10^{-76}
10^{-156}
10^{-236}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

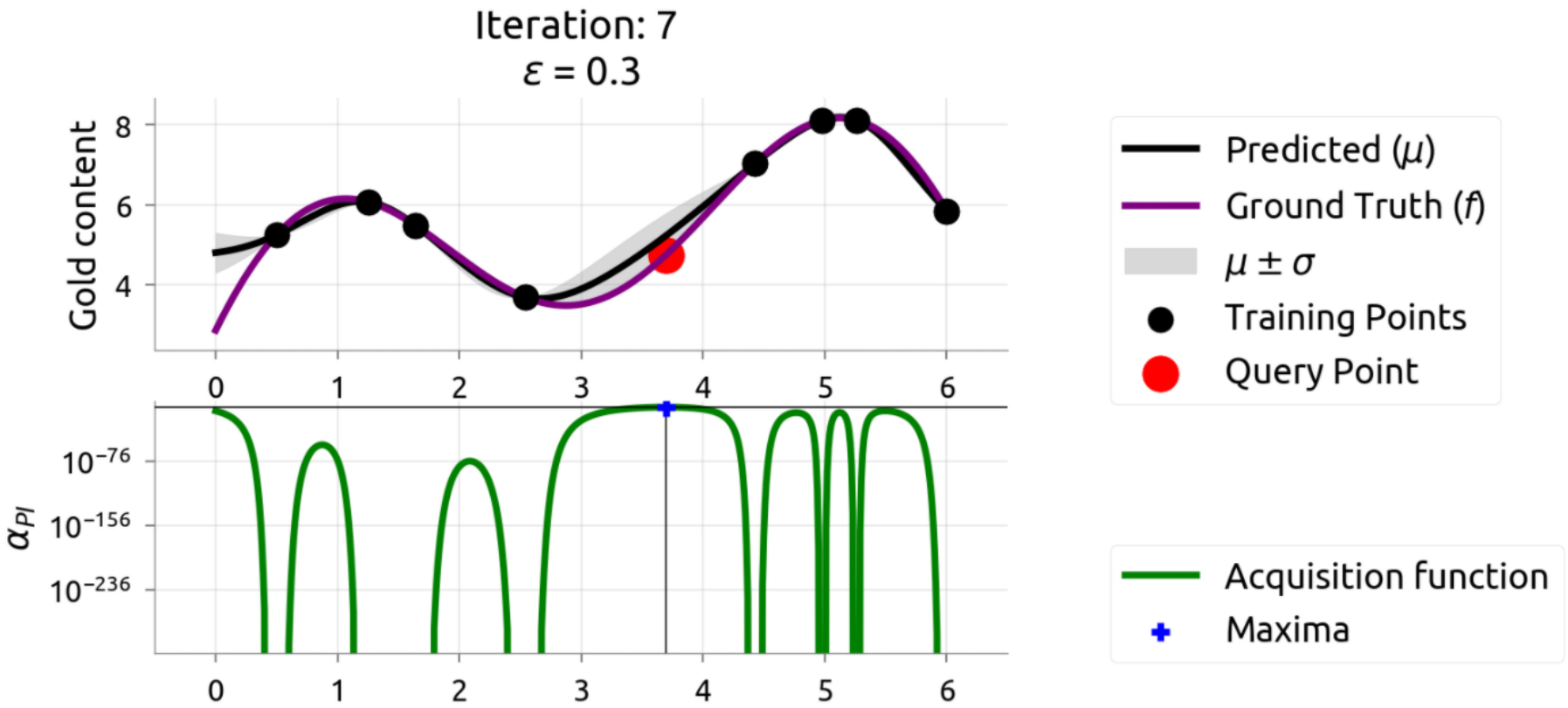
$\epsilon = 0.3$


Iteration: 5
ϵ = 0.3
Gold content
8
6
4
0
1
2
3
4
5
6
α_{PI}
10^{-76}
10^{-156}
10^{-236}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

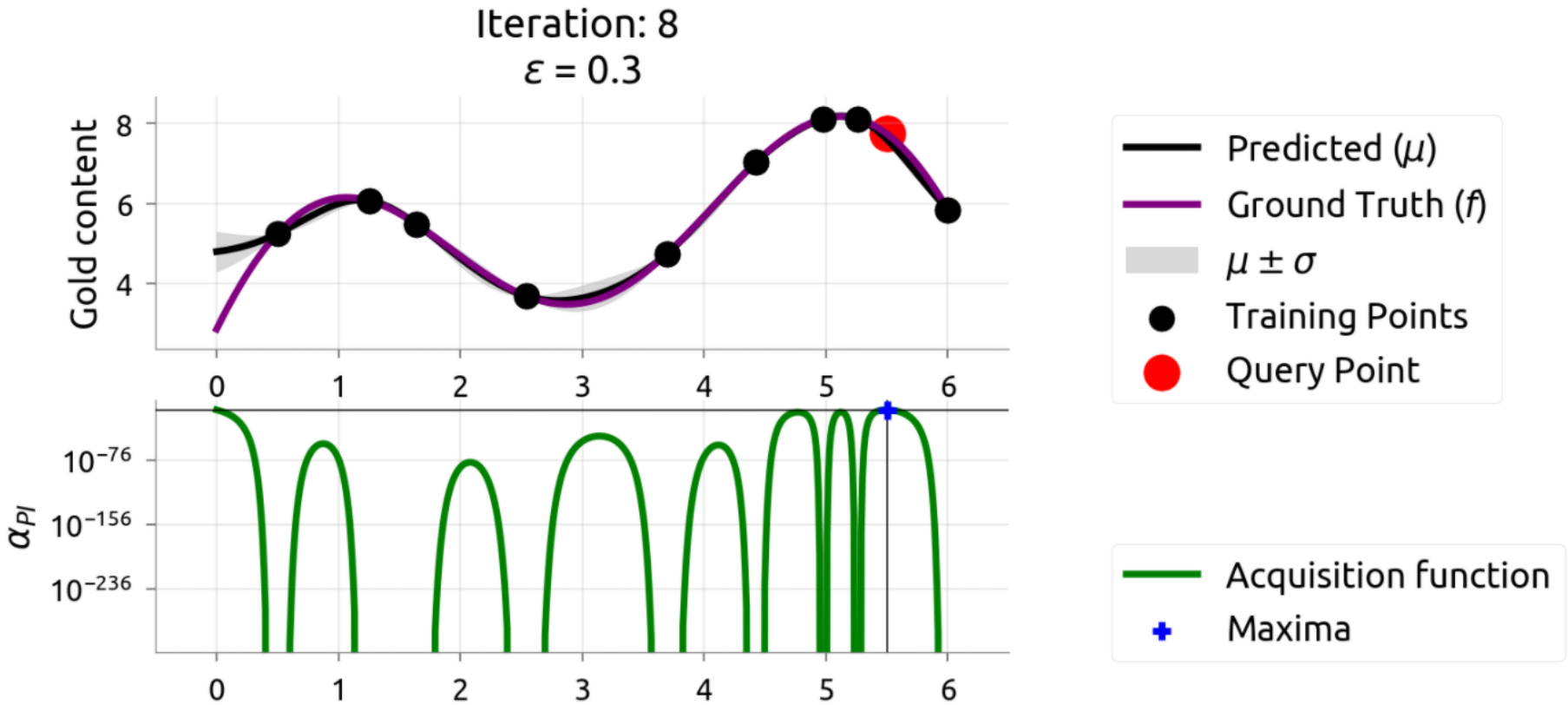
$\epsilon = 0.3$


Iteration: 6
ϵ = 0.3
Gold content
8
6
4
0
1
2
3
4
5
6
α_{PI}
10^{-76}
10^{-156}
10^{-236}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

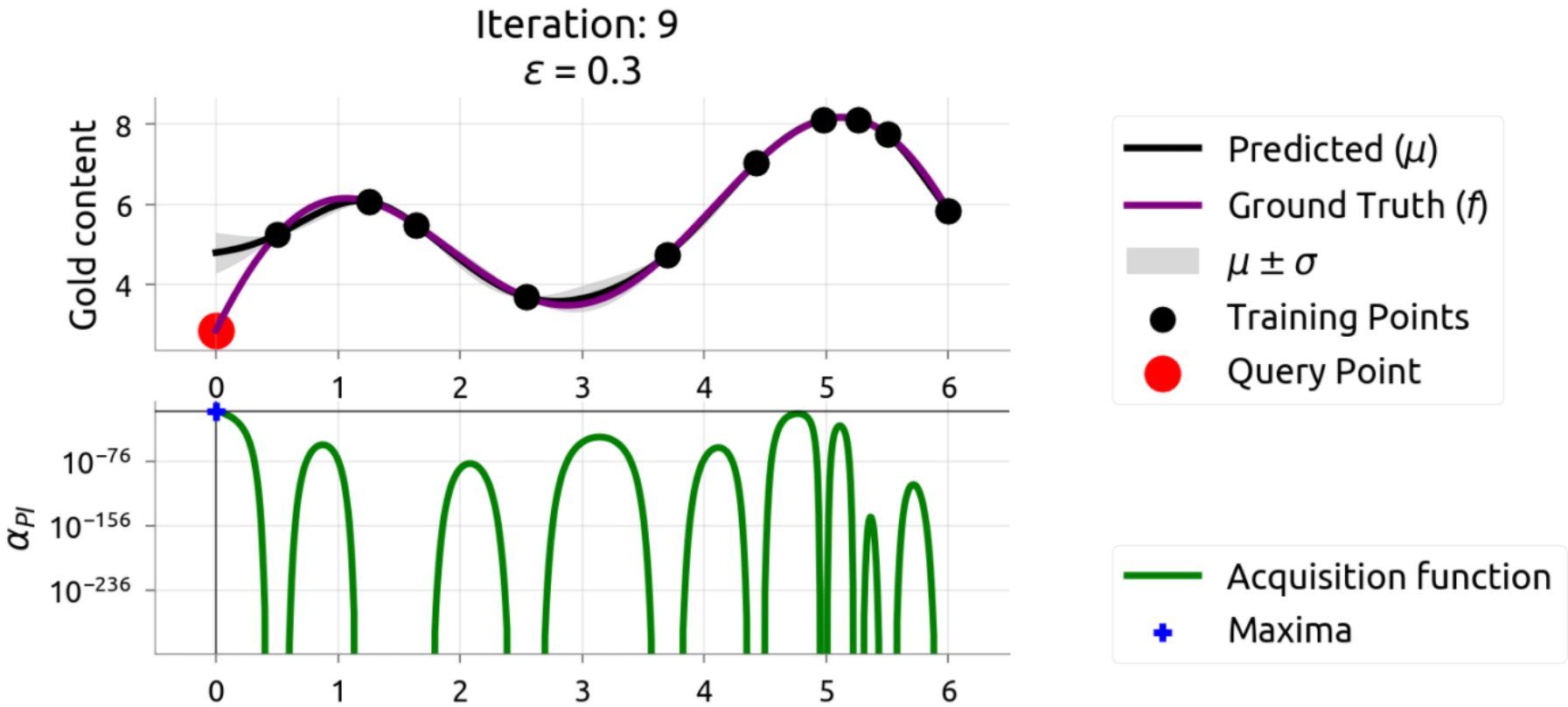
$\epsilon = 0.3$


Iteration: 7
ϵ = 0.3
Gold content
8
6
4
0
1
2
3
4
5
6
α_{PI}
10^{-76}
10^{-156}
10^{-236}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

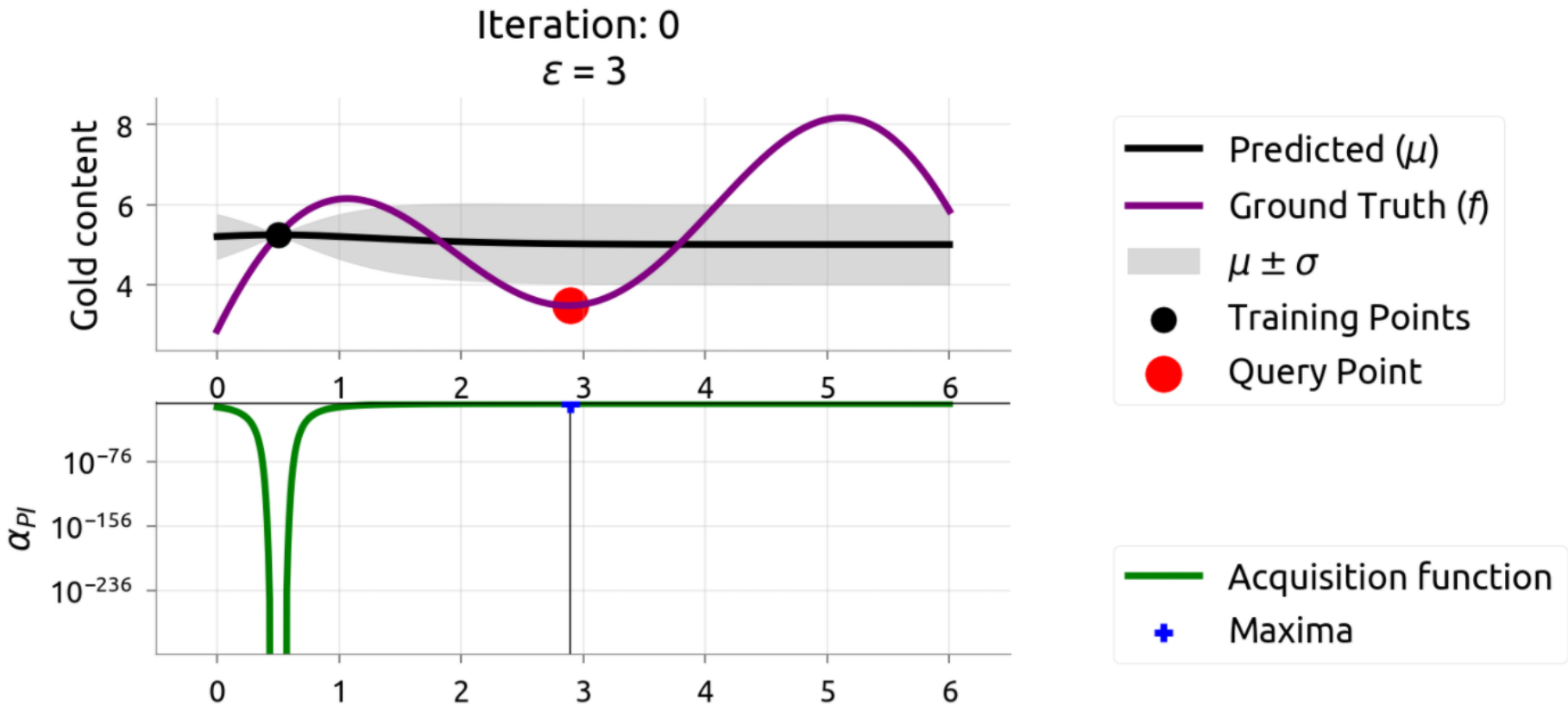
$\epsilon = 0.3$


Iteration: 8
ε = 0.3
Gold content
8
6
4
0
1
2
3
4
5
6
α_{PI}
10^{-76}
10^{-156}
10^{-236}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

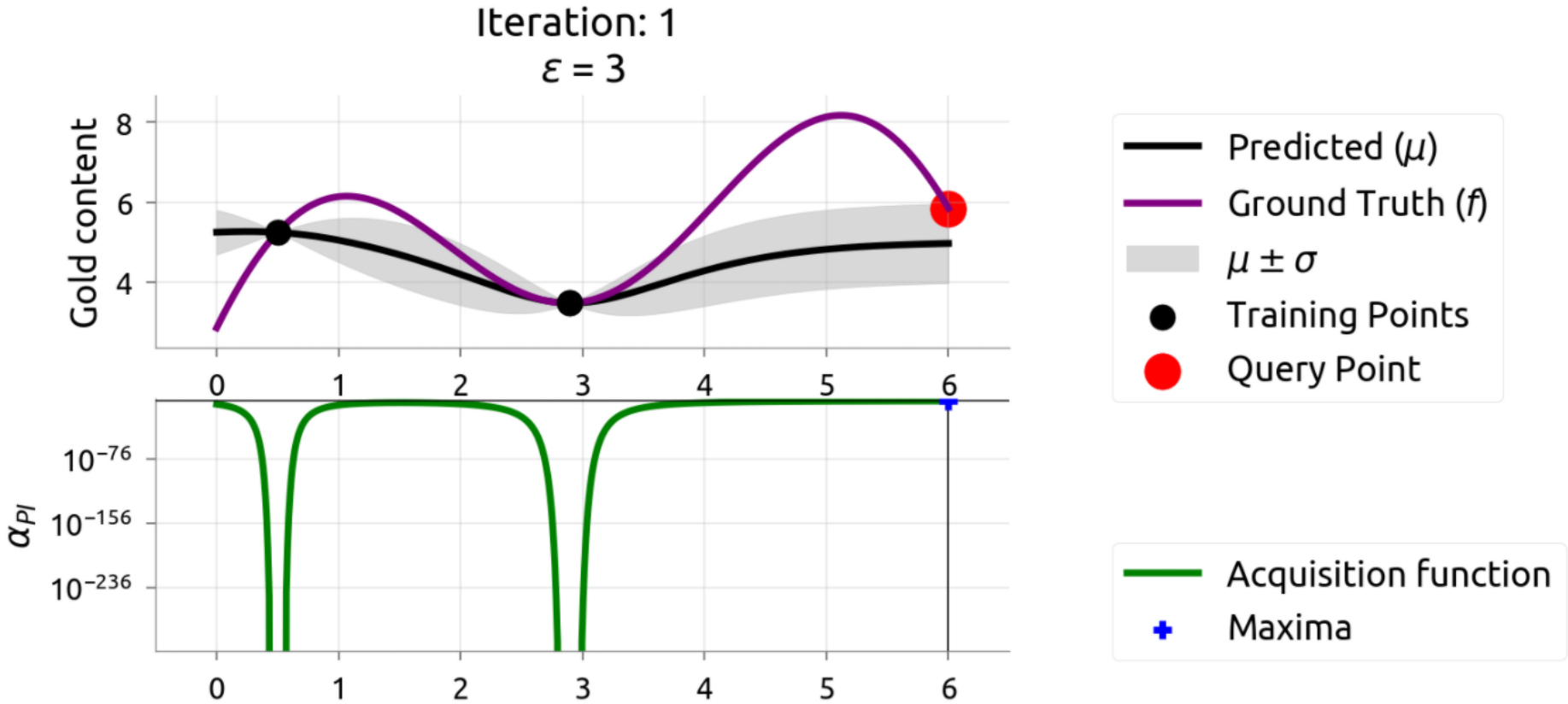
$\epsilon = 0.3$


Iteration: 9
ϵ = 0.3
Gold content
8
6
4
0
1
2
3
4
5
6
α_{PI}
10^{-76}
10^{-156}
10^{-236}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

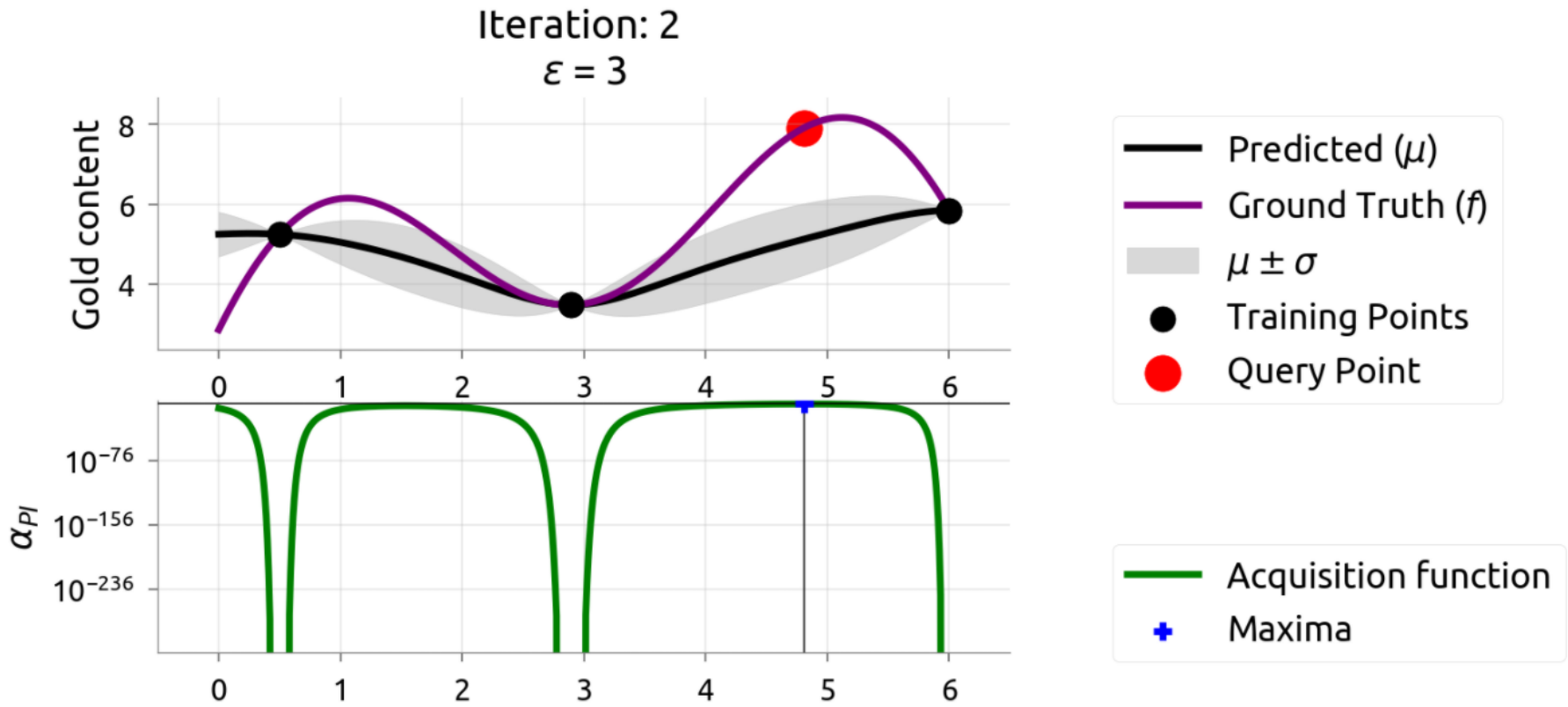
$\epsilon = 3$


Iteration: 0
ε = 3
Gold content
8
6
4
0
1
2
3
4
5
6
α_{PI}
10^{-76}
10^{-156}
10^{-236}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

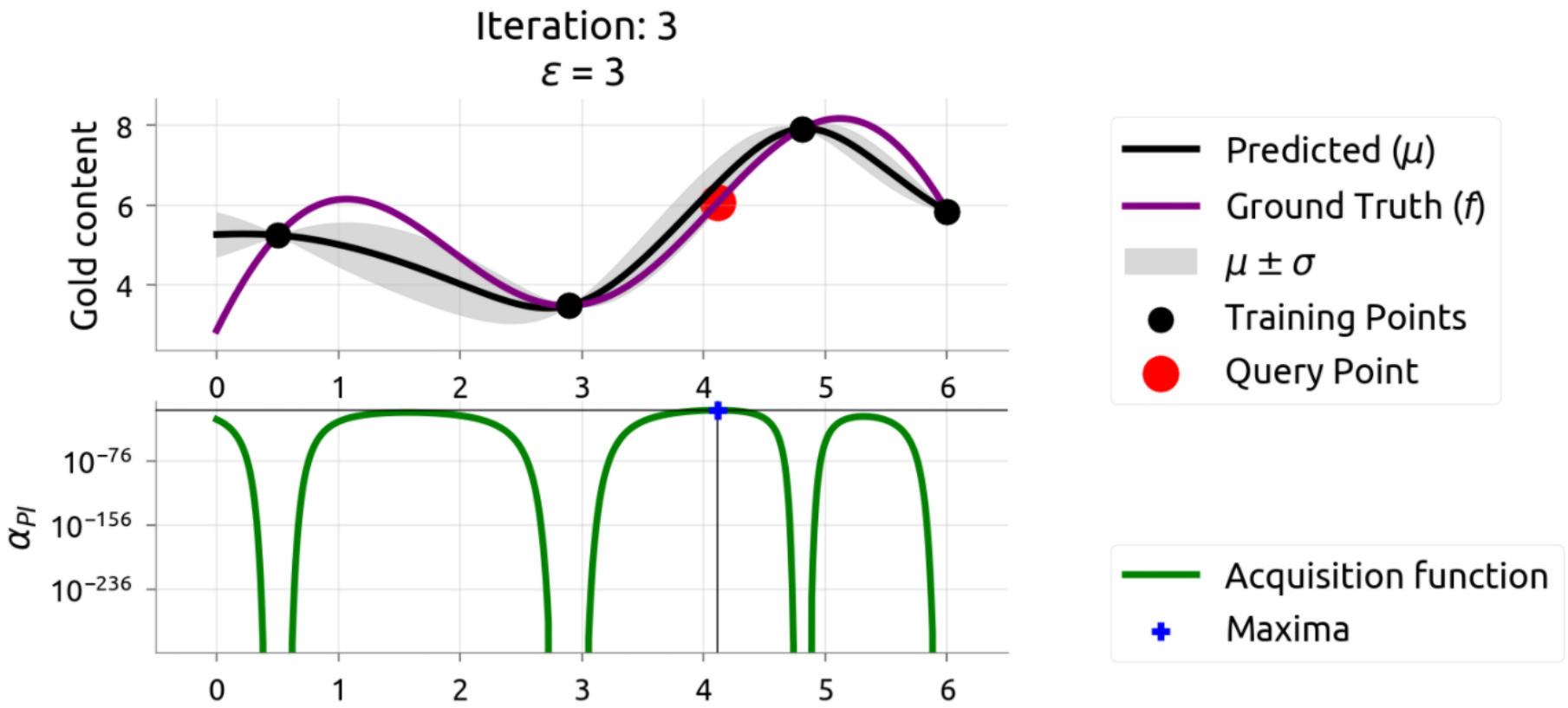
$\epsilon = 3$


Iteration: 1
ϵ = 3
Gold content
8
6
4
0
1
2
3
4
5
6
α_{PI}
10^{-76}
10^{-156}
10^{-236}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

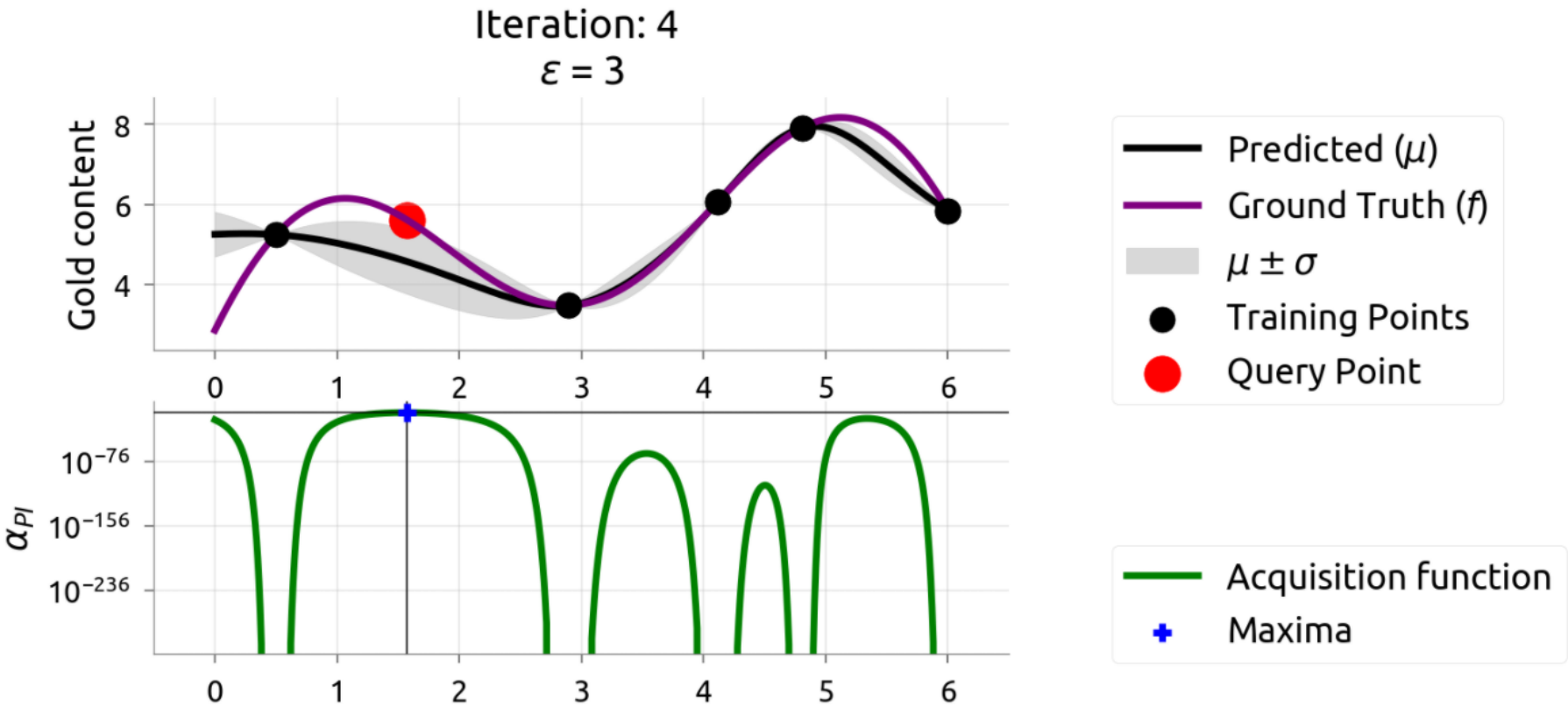
$\epsilon = 3$


Iteration: 2
ϵ = 3
Gold content
8
6
4
0
1
2
3
4
5
6
α_{PI}
10^{-76}
10^{-156}
10^{-236}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

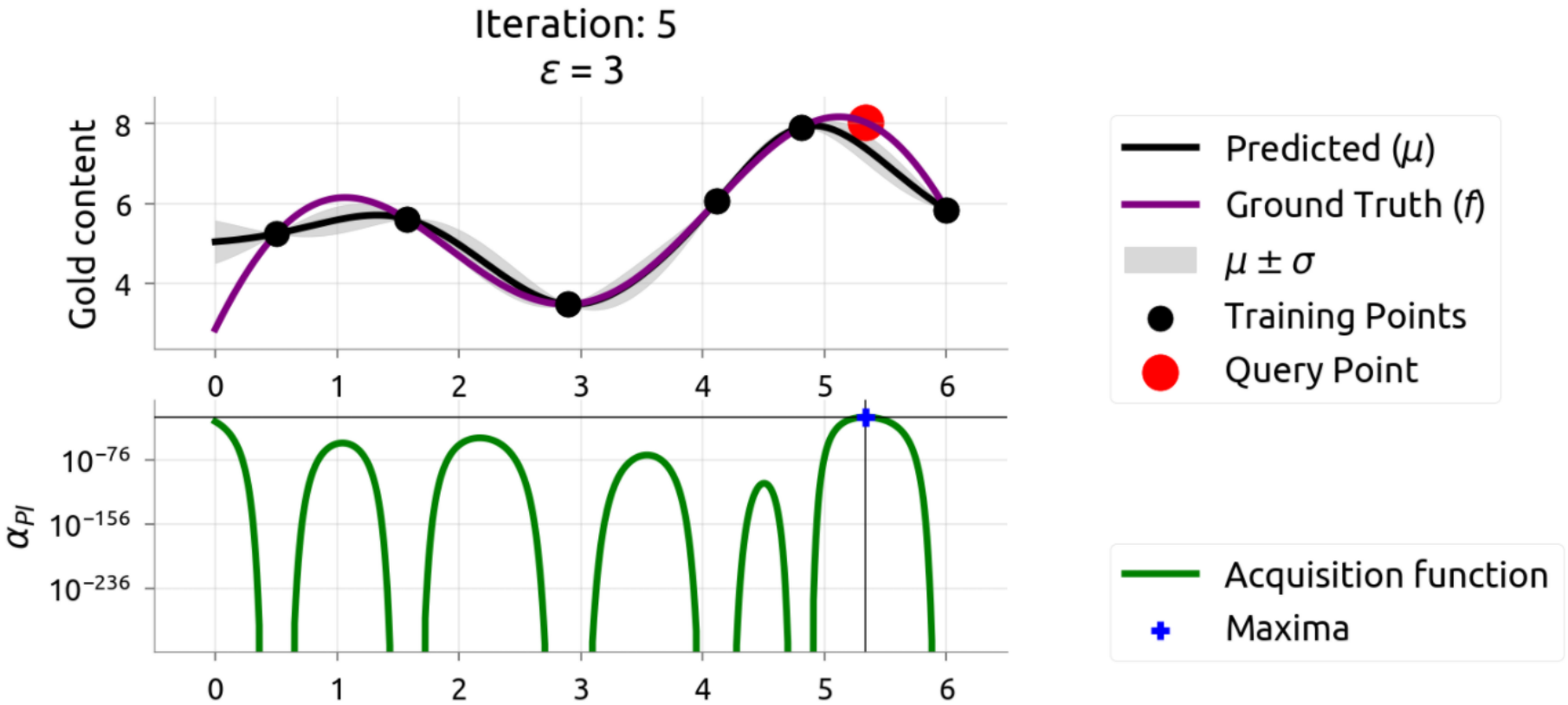
$\epsilon = 3$


Iteration: 3
ϵ = 3
Gold content
8
6
4
0
1
2
3
4
5
6
α_{PI}
10^{-76}
10^{-156}
10^{-236}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

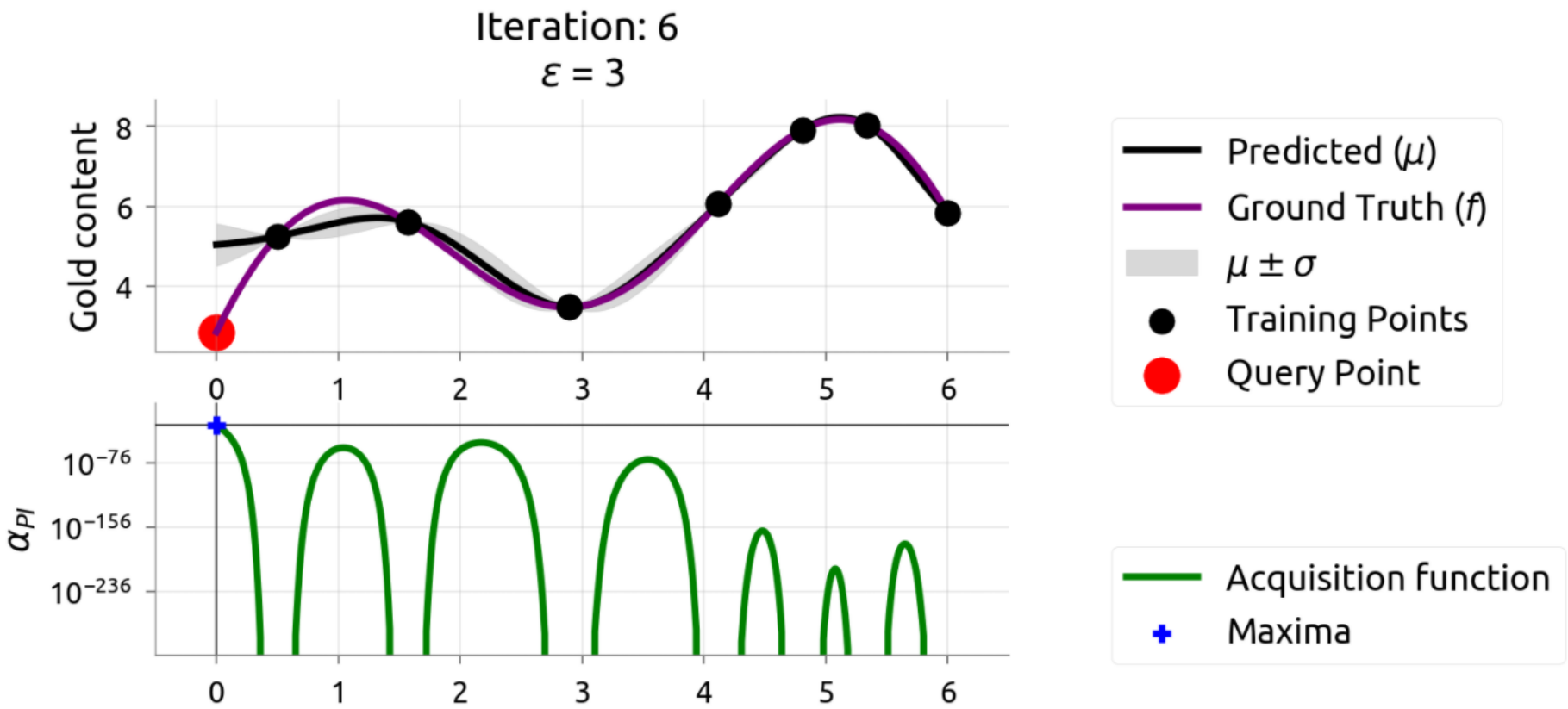
$\epsilon = 3$


Iteration: 4
$\varepsilon = 3$
Gold content
8
6
4
0
1
2
3
4
5
6
α_{PI}
10^{-76}
10^{-156}
10^{-236}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

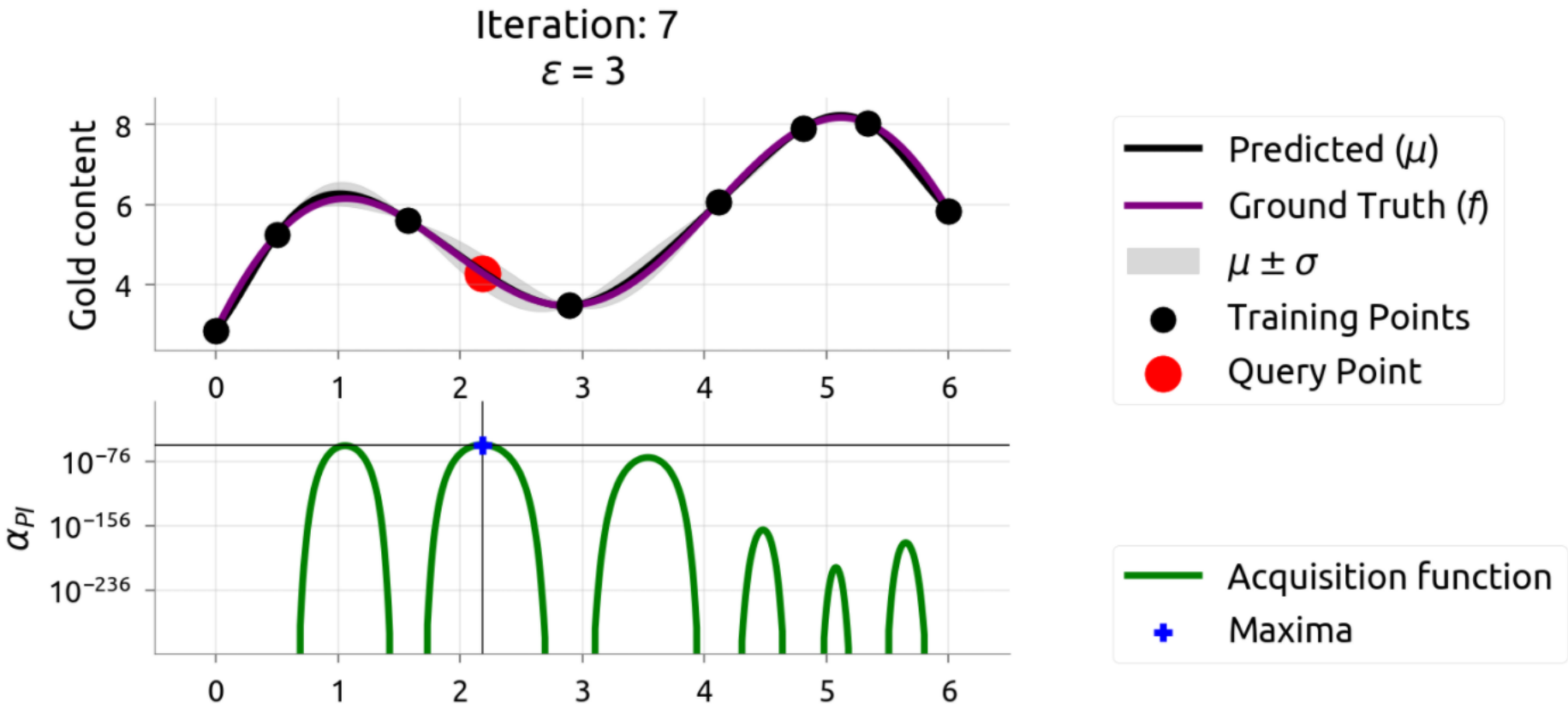
$\epsilon = 3$


Iteration: 5
ϵ = 3
Gold content
8
6
4
0
1
2
3
4
5
6
α_{PI}
10^{-76}
10^{-156}
10^{-236}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

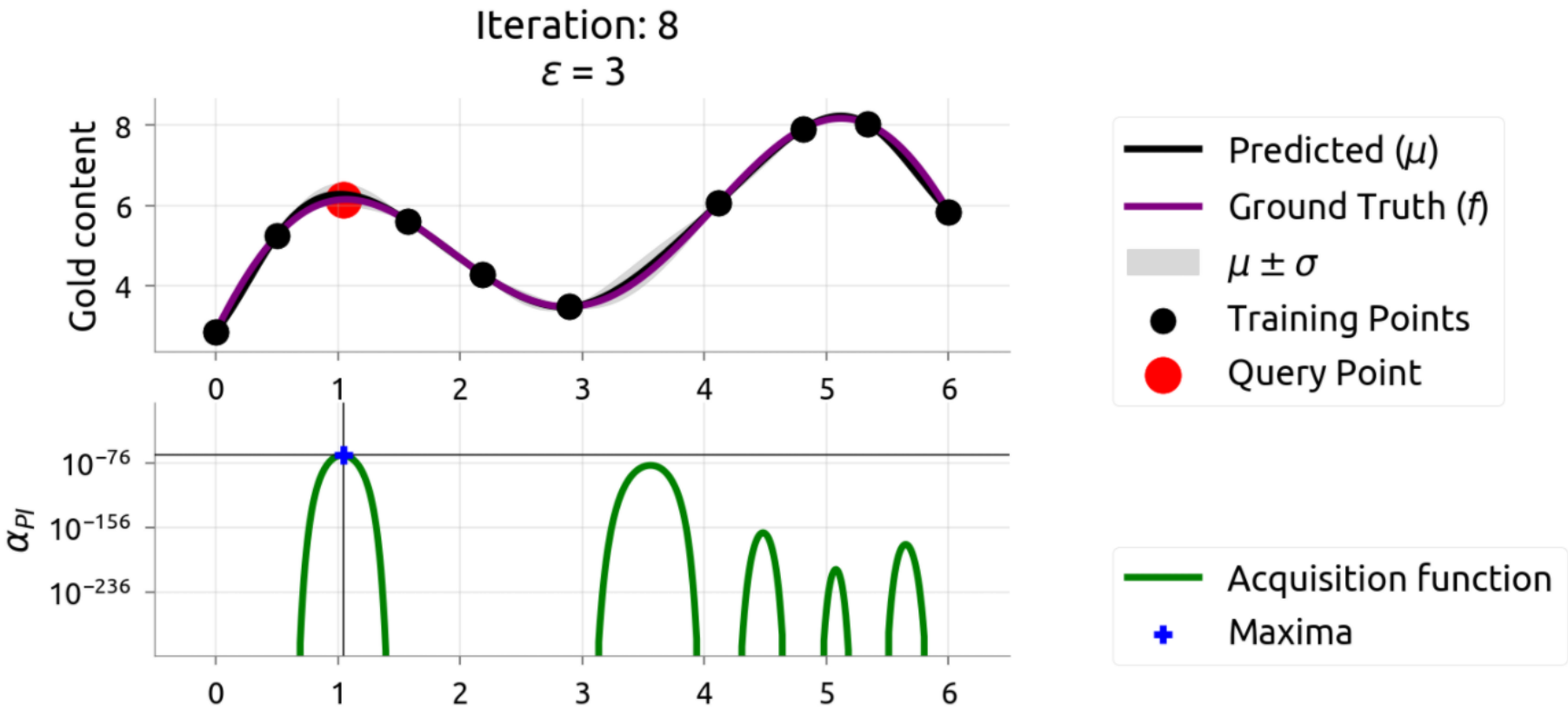
$\epsilon = 3$


Iteration: 6
ϵ = 3
Gold content
8
6
4
0
1
2
3
4
5
6
α_{PI}
10^{-76}
10^{-156}
10^{-236}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

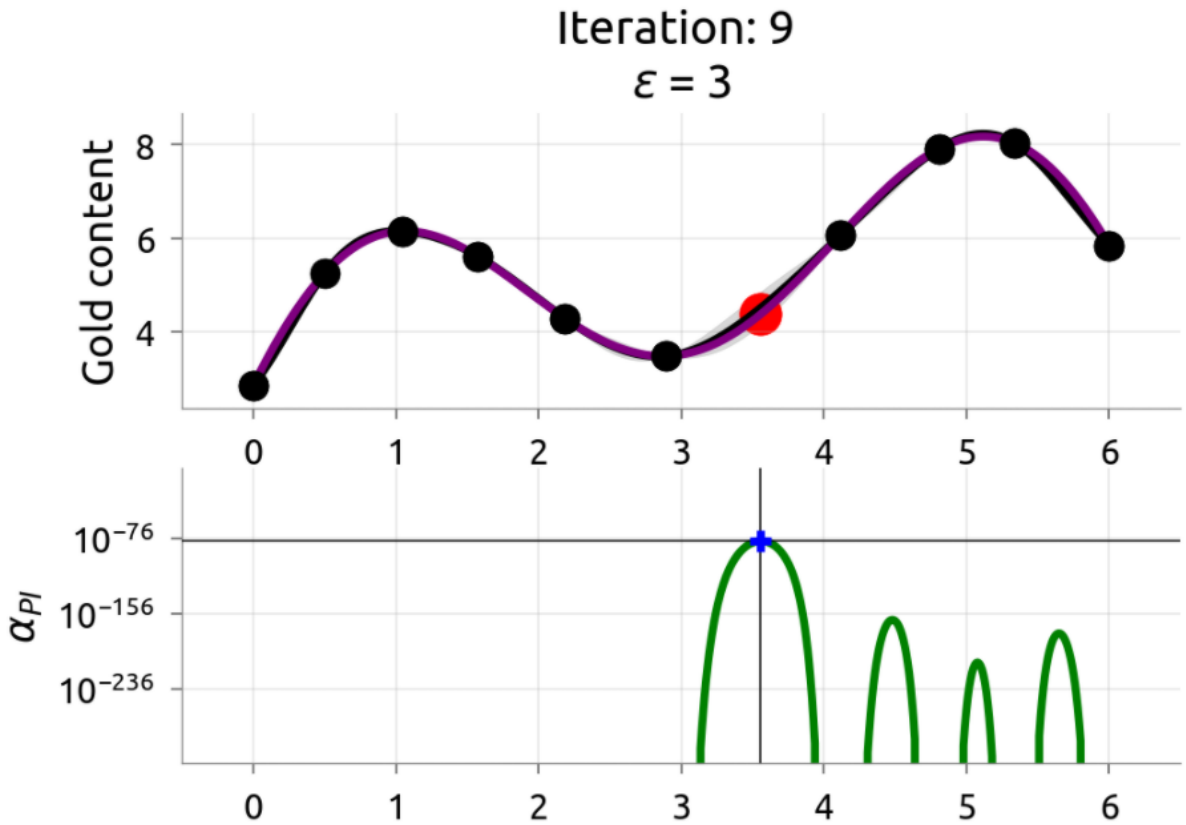
$\epsilon = 3$

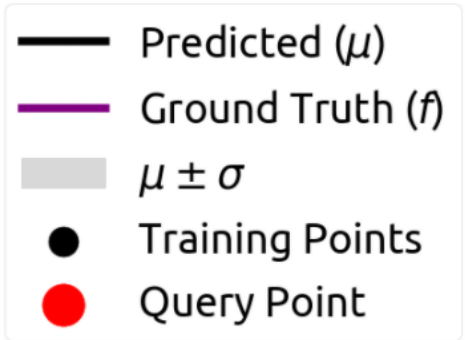

Iteration: 7
ϵ = 3
Gold content
8
6
4
0
1
2
3
4
5
6
α_{PI}
10^{-76}
10^{-156}
10^{-236}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

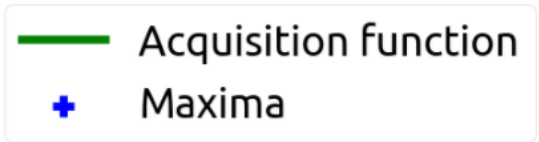
$\epsilon = 3$


Iteration: 8
ϵ = 3
Gold content
8
6
4
0
1
2
3
4
5
6
α_{PI}
10^{-76}
10^{-156}
10^{-236}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

$\epsilon = 3$


Iteration: 9
ε = 3
Gold content
8
6
4
0
1
2
3
4
5
6
α_{PI}
10^{-76}
10^{-156}
10^{-236}
0
1
2
3
4
5
6


Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point


Acquisition function
Maxima

Expected improvement (EI)

The main disadvantage of PI is that it does not take into account how much the improvement will be. Expected improvement (EI) computes the expectation of the improvement:

$$\mathbf{x}_{t+1} = \text{argmax}[\text{EI}(\mathbf{x})], \tag{10}$$

where

$$\text{EI}(\mathbf{x}) = \mathbb{E}[(\mu(\mathbf{x}) - f(x^+))\mathbb{I}(\mathbf{x} > x^+)] \tag{11}$$

$$\mathbb{I}(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{12}$$

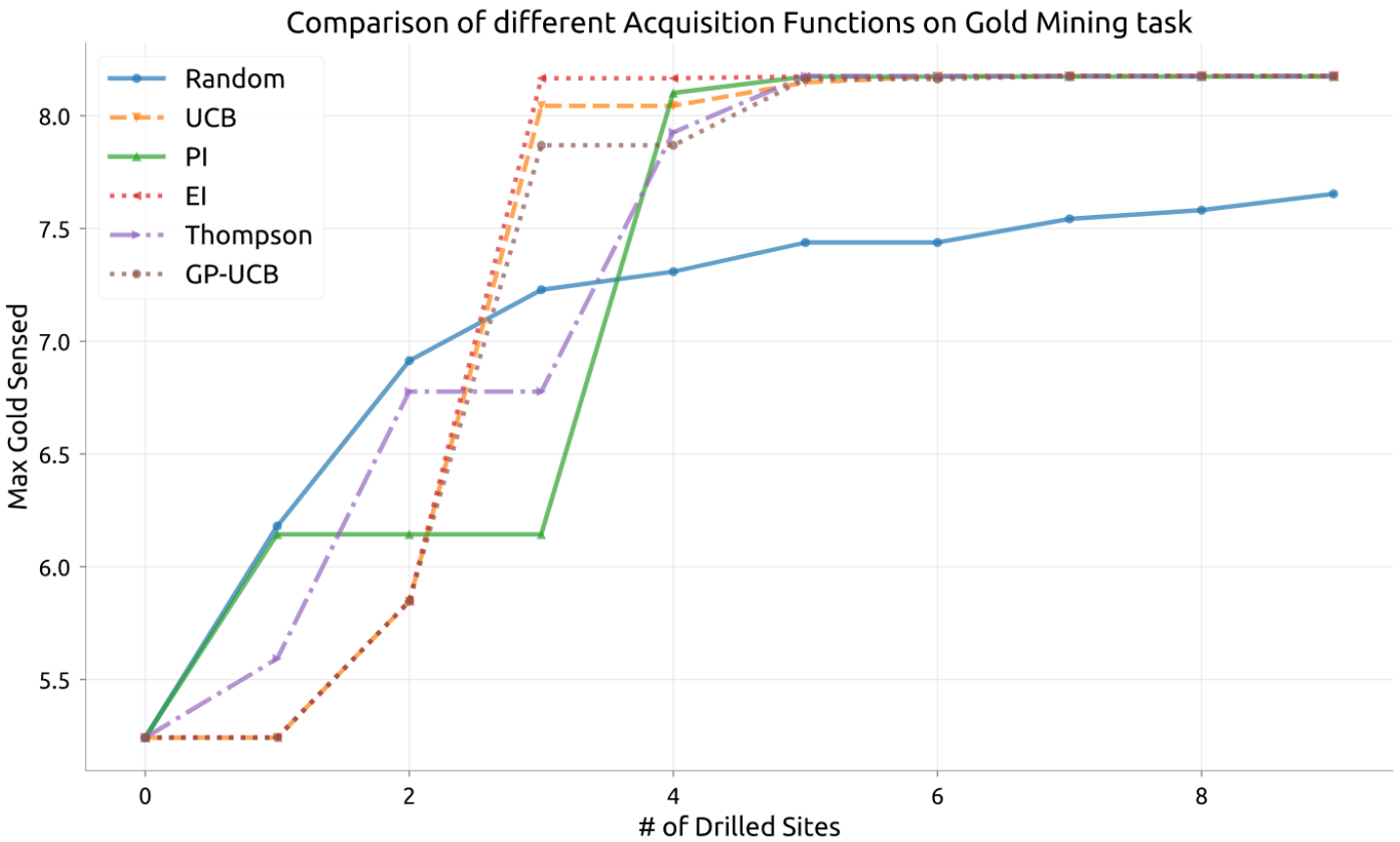
Upper Confidence Bound(UCB)

This acquisition function is defined as:

$$\text{UCB}(\mathbf{x}) = \mu(\mathbf{x}) + \beta^{1/2}\sigma(\mathbf{x}) \tag{13}$$

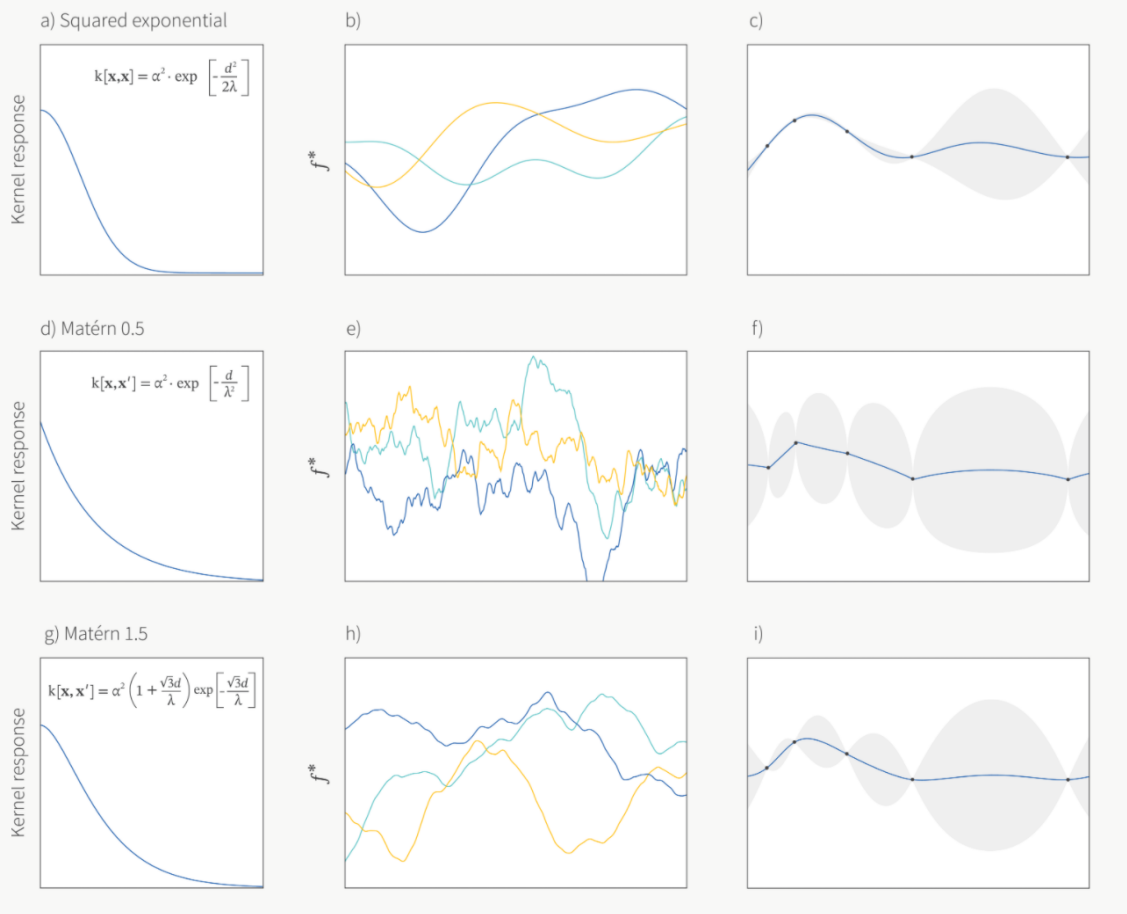
This favors either (i) regions where $\mu(\mathbf{x})$ is large (for exploitation) or (ii) regions where $\sigma(\mathbf{x})$ is large (for exploration). The positive parameter β trades off these two tendencies.

Acquisition Function Comparison



Other Issue

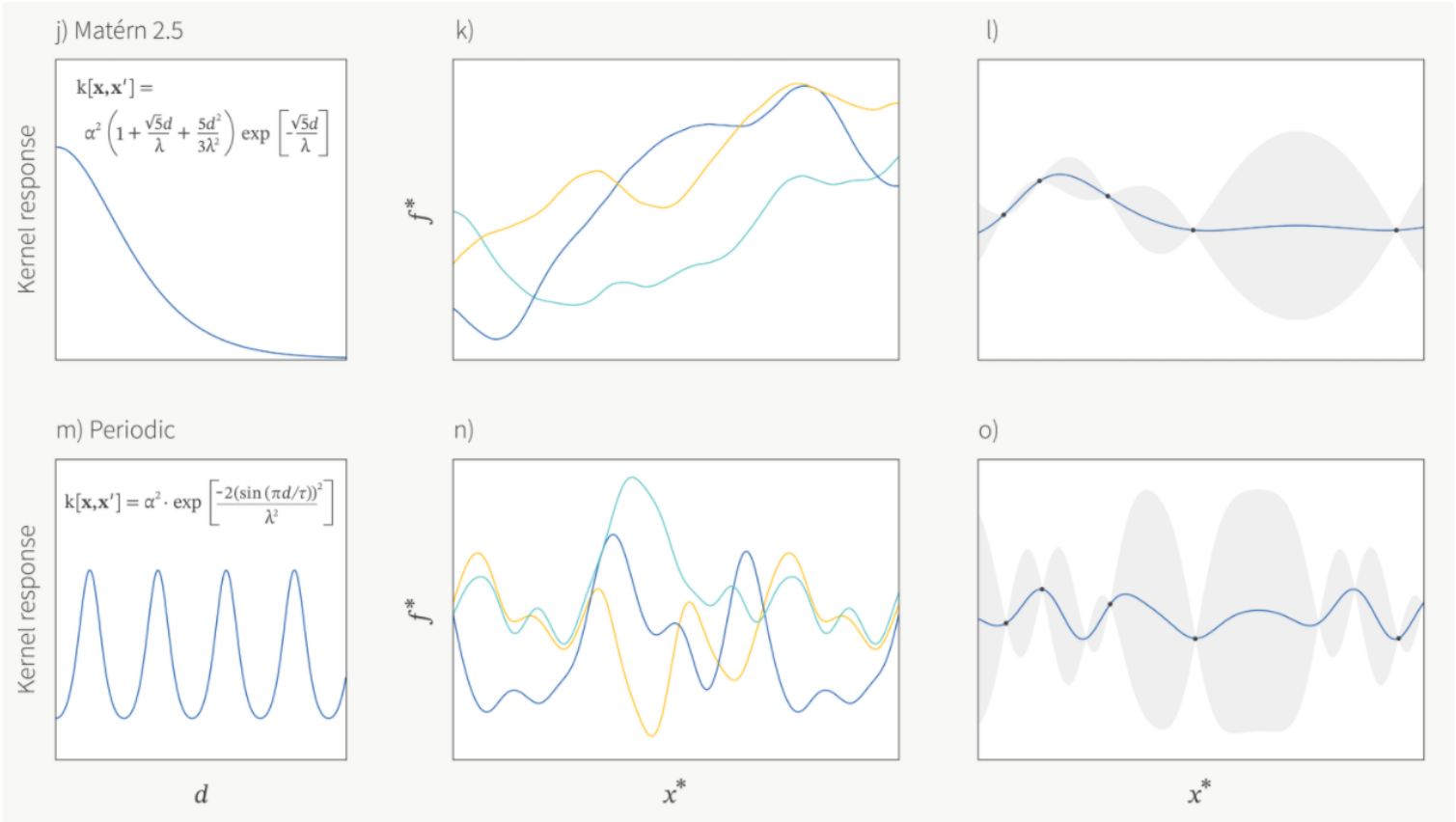
Choosing Kernel



where d is the Euclidean distance between the points.

Other Issue

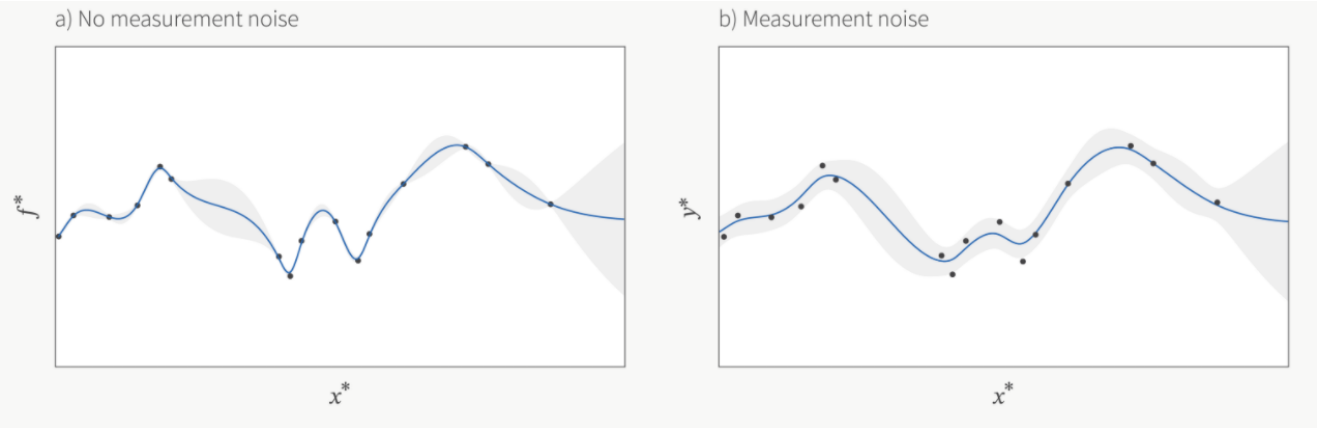
Choosing Kernel



where d is the Euclidean distance between the points.

Other Issue

Data with noise



we add an extra noise term to the expression for the Gaussian process covariance:

$$\mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))] = k(\mathbf{x}, \mathbf{x}') + \text{noise term} \tag{14}$$

Application

BO has emerged as a powerful solution for these varied problems:

- Interactive user interfaces
- Robotics
- Environmental monitoring
- Information extraction
- Automatic machine learning
- Sensor networks
- Adaptive Monte Carlo
- Reinforcement learning

Summary

The pseudo code of BO

Algorithm 1 Basic pseudo-code for Bayesian optimization

Place a Gaussian process prior on f
Observe f at n_0 points according to an initial space-filling experimental design. Set $n = n_0$.
while $n \leq N$ **do**
  Update the posterior probability distribution on f using all available data
  Let x_n be a maximizer of the acquisition function over x, where the acquisition function is computed using the current posterior distribution.
  Observe $y_n = f(x_n)$.
  Increment n
end while
Return a solution: either the point evaluated with the largest $f(x)$, or the point with the largest posterior mean.

Reference

For further references see

- www.borealisai.com/en/blog/tutorial-8-bayesian-optimization/
- https://distill.pub/2020/bayesian-optimization/
- Taking the Human Out of the Loop: A Review of Bayesian Optimization