Bayesian Optimization and Active Learning

Bayesian Optimization and Active Learning

Data Science in Electron Microscopy

Philipp Pelz

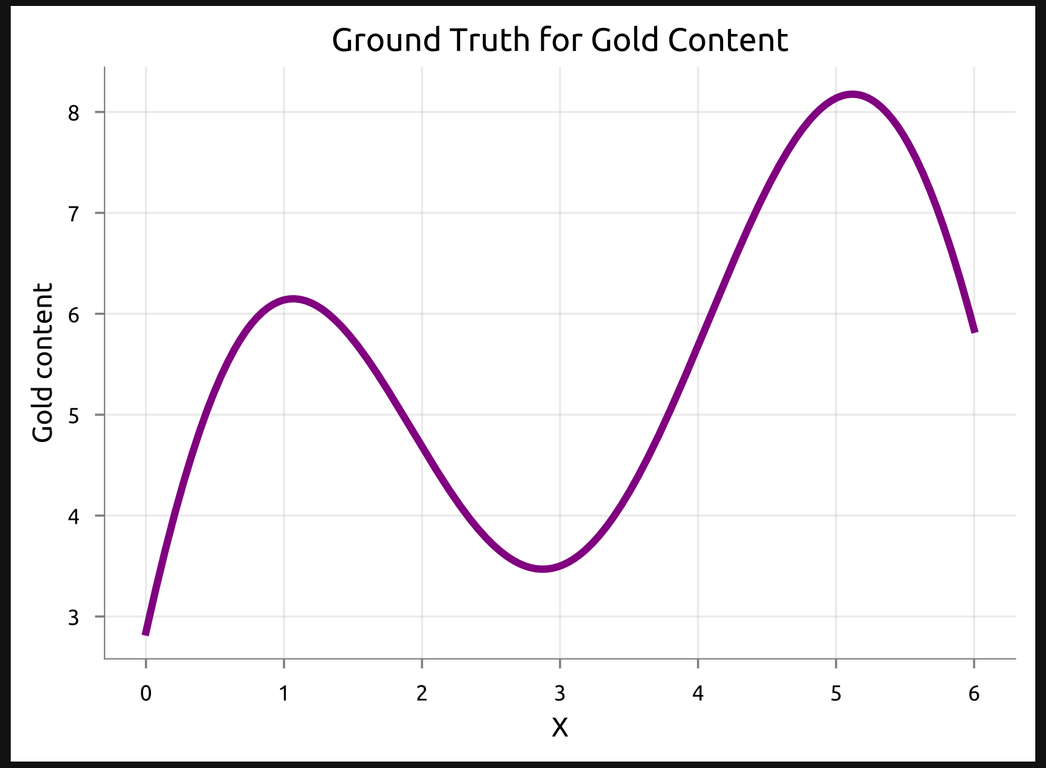
2025-01-30

https://github.com/ECLIPSE-Lab/WS24_DataScienceForEM ##
Introduction

- Bayesian Optimization is a method for optimizing black-box functions that are expensive to evaluate.
- Useful for tuning hyperparameters in machine learning models.

Gold Mining Analogy

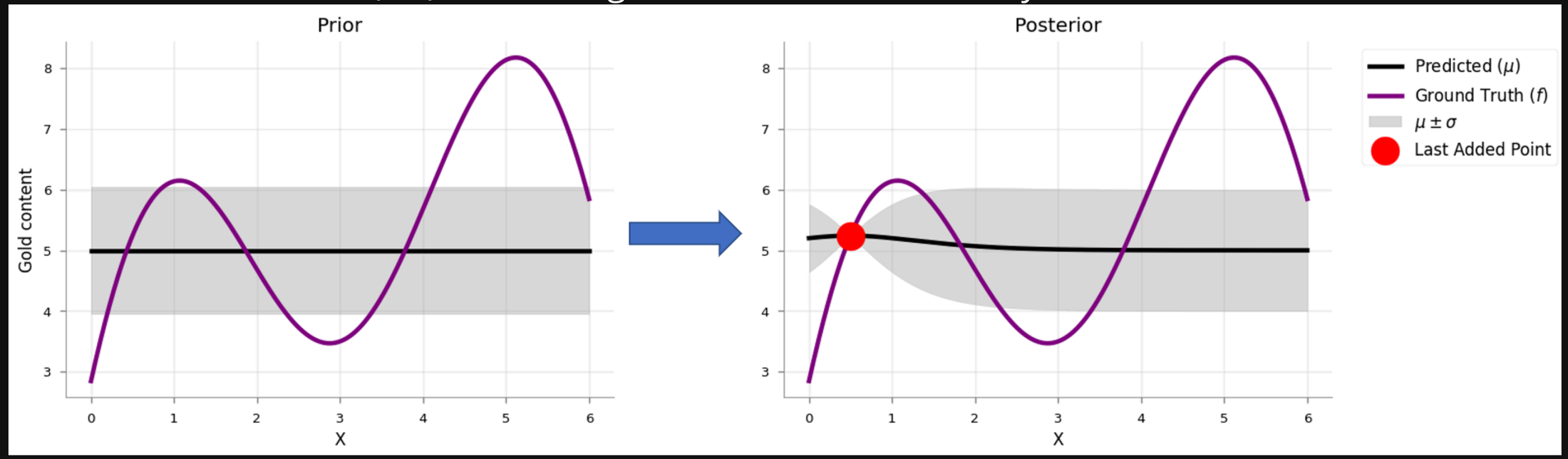
- Goal: Find the maximum gold content along a line with minimal drillings.
- Two objectives:
  1. **Active Learning**: Estimate gold distribution accurately.
  2. **Bayesian Optimization**: Find the location of maximum gold content.



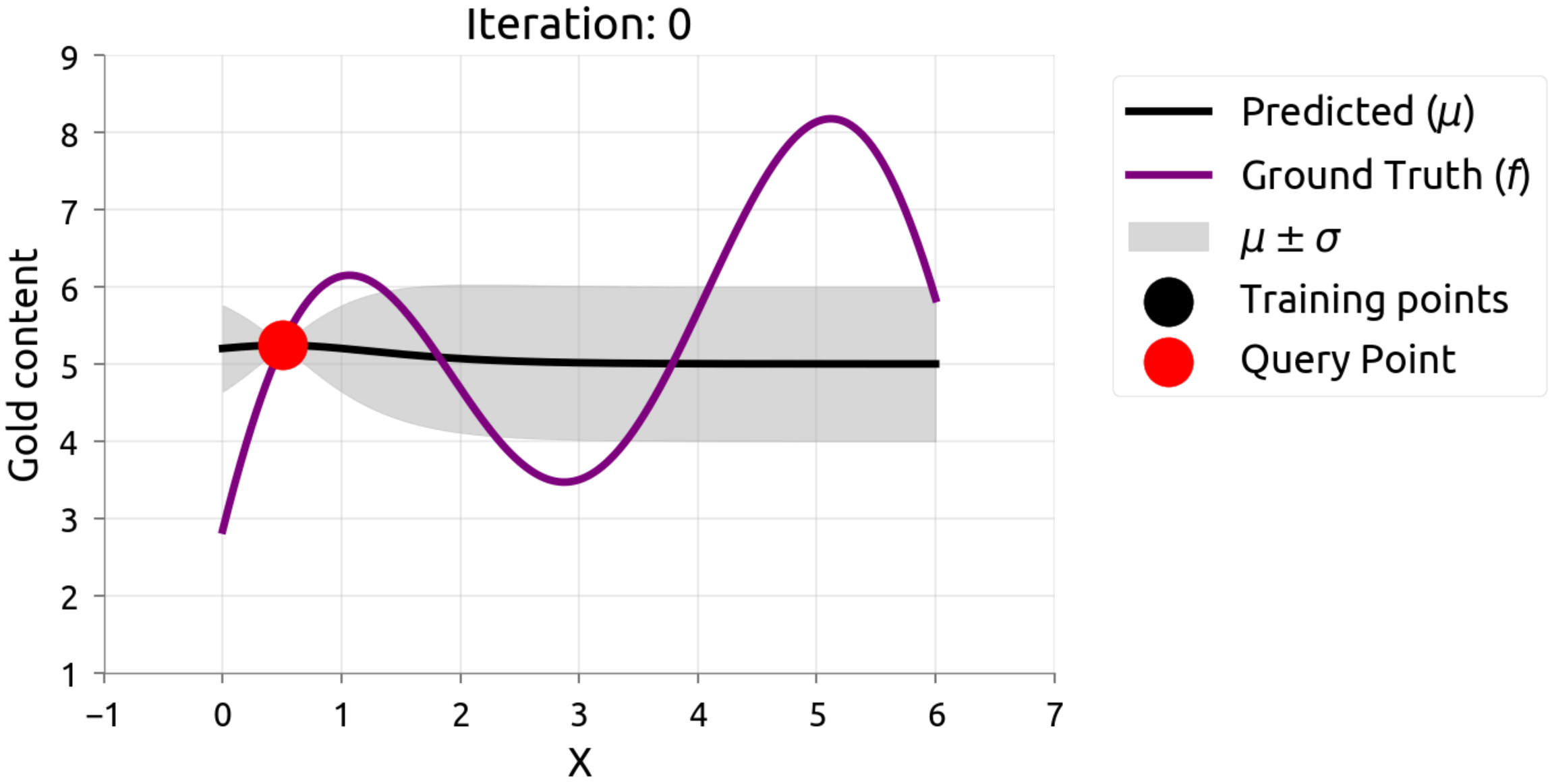
Gold Mining

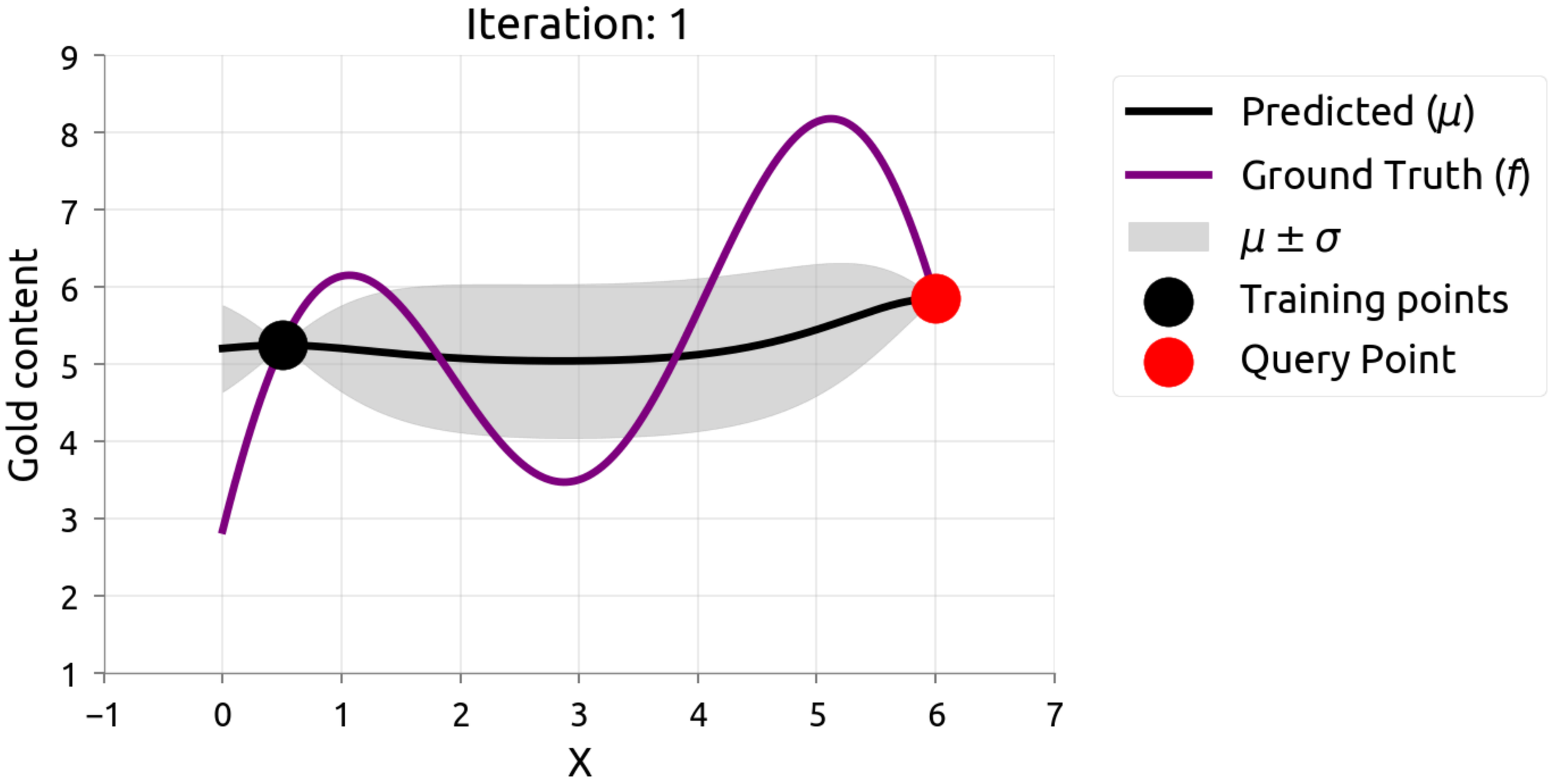
Active Learning

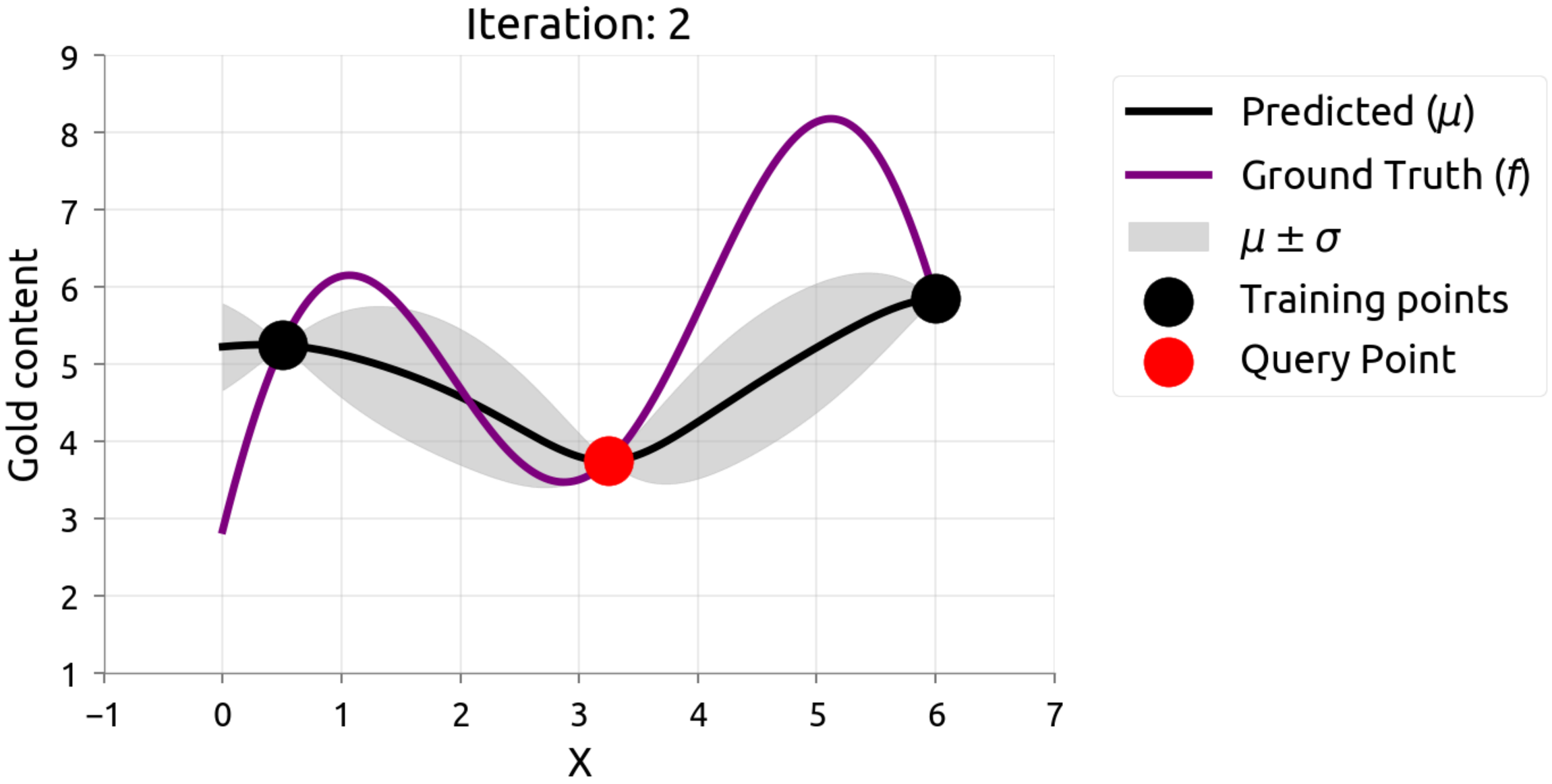
- Aim: Minimize labeling costs while maximizing modeling accuracy.
- Strategy: Label the point with the highest model uncertainty (variance).
- Use Gaussian Process (GP) as a surrogate model for uncertainty estimates.

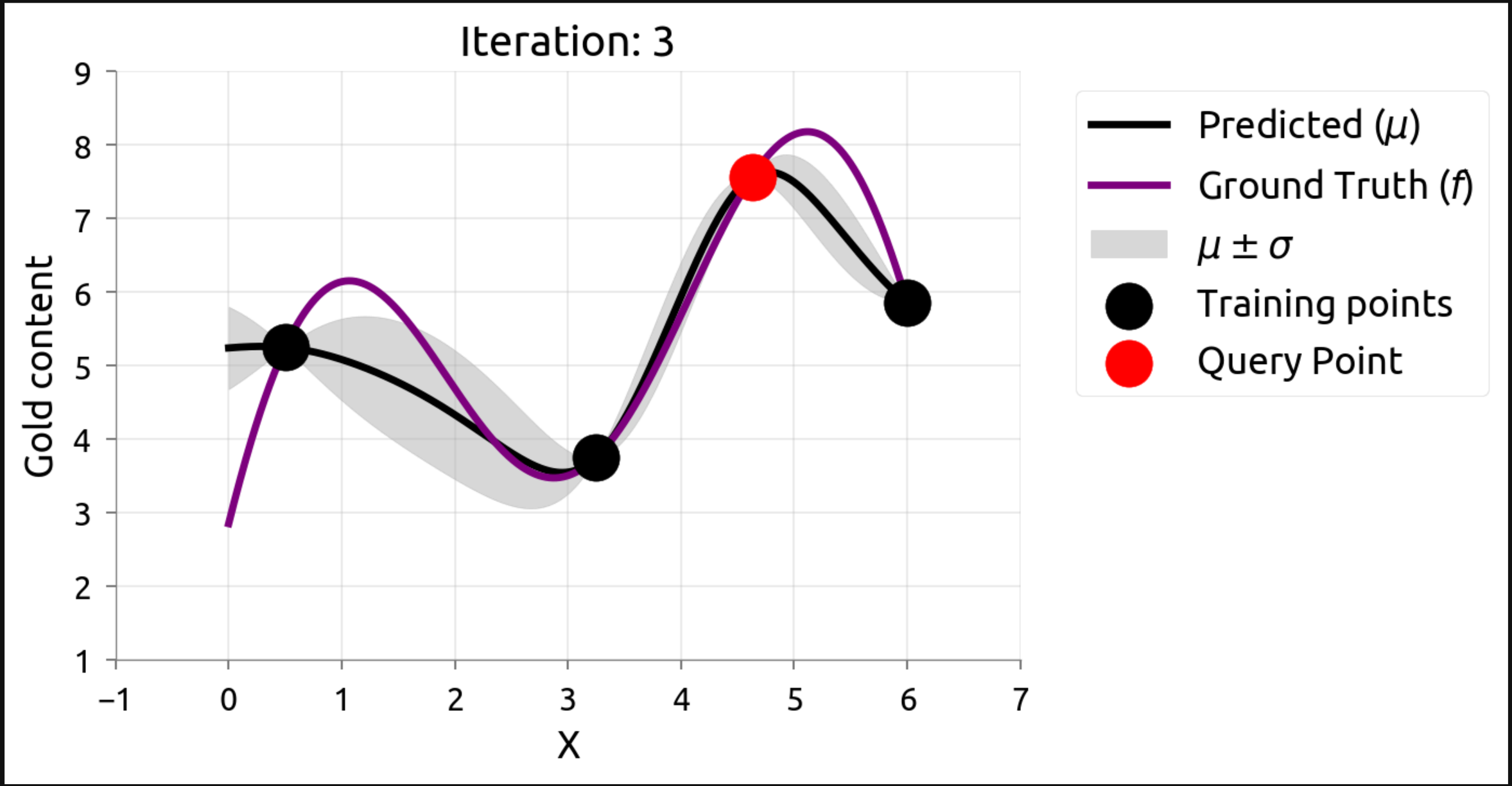


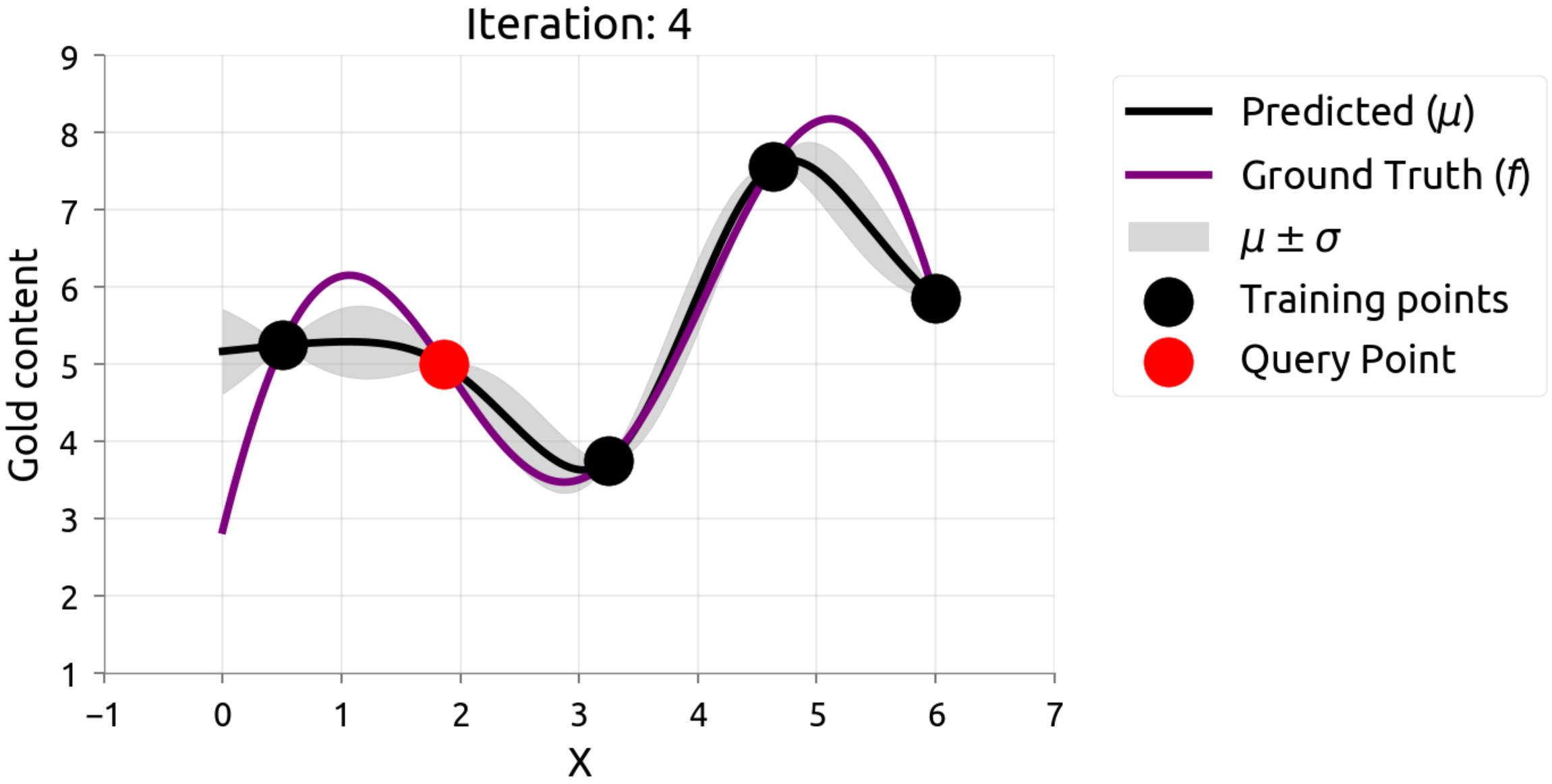
- Each new data point updates our surrogate model, moving it closer to the ground truth. The black line and the grey shaded region indicate the mean μ and uncertainty μ±σ in our gold distribution estimate before and after drilling.

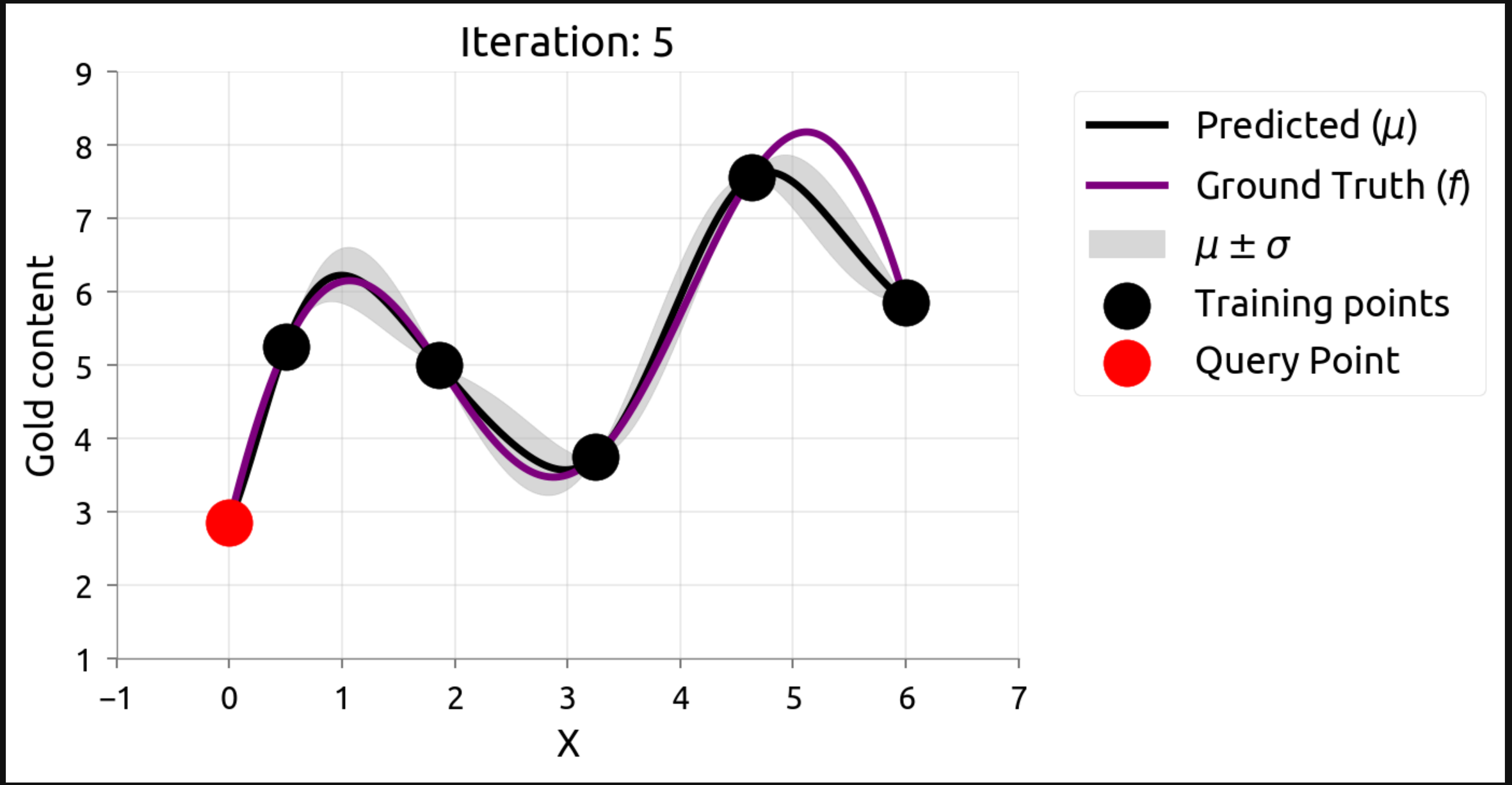

Iteration: 0
Gold content
X
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training points
Query Point

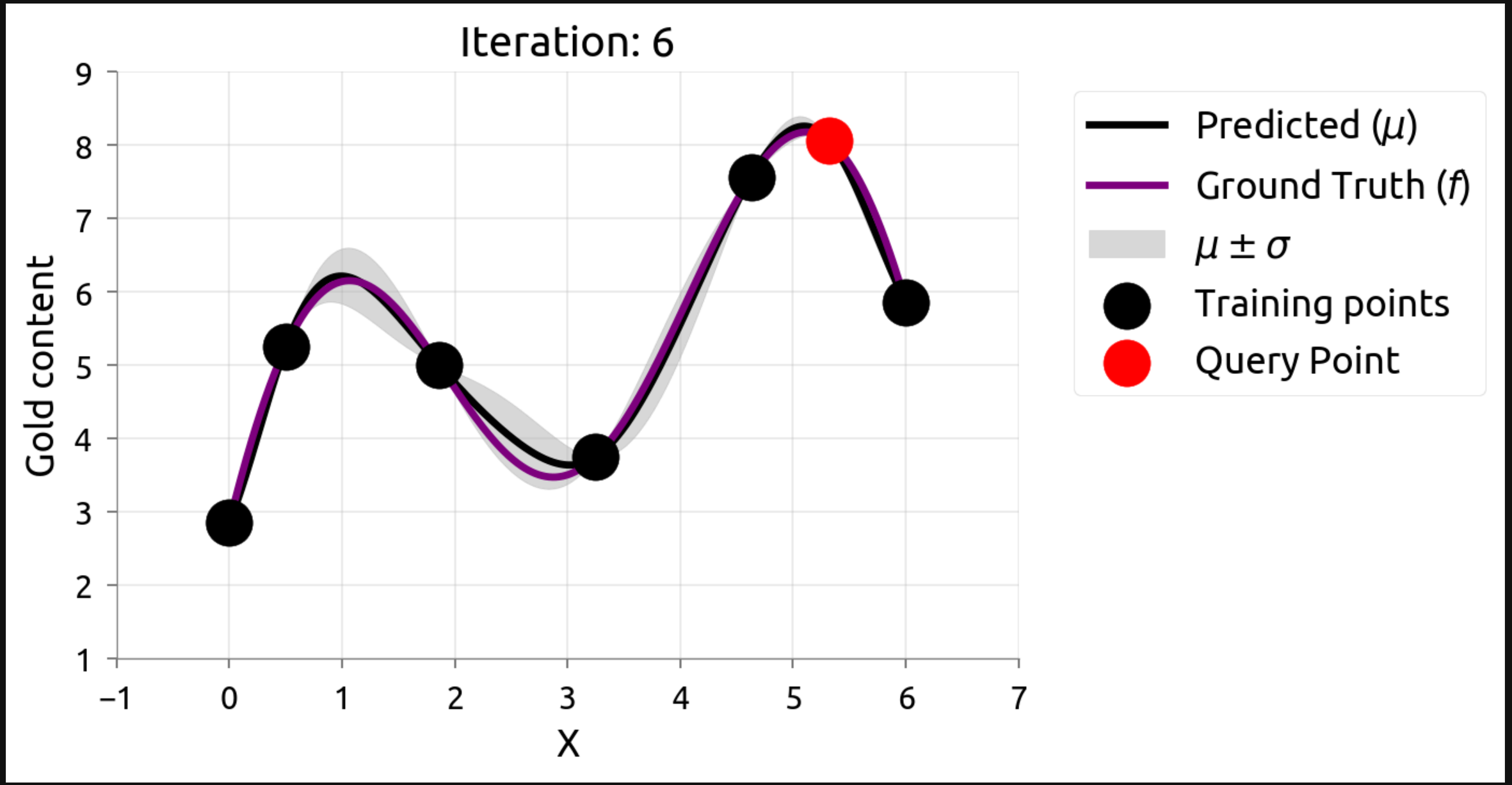

Iteration: 1
Gold content
X
9
8
7
6
5
4
3
2
1
−1
0
1
2
3
4
5
6
7
Predicted (μ)
Ground Truth (f)
μ ± σ
Training points
Query Point

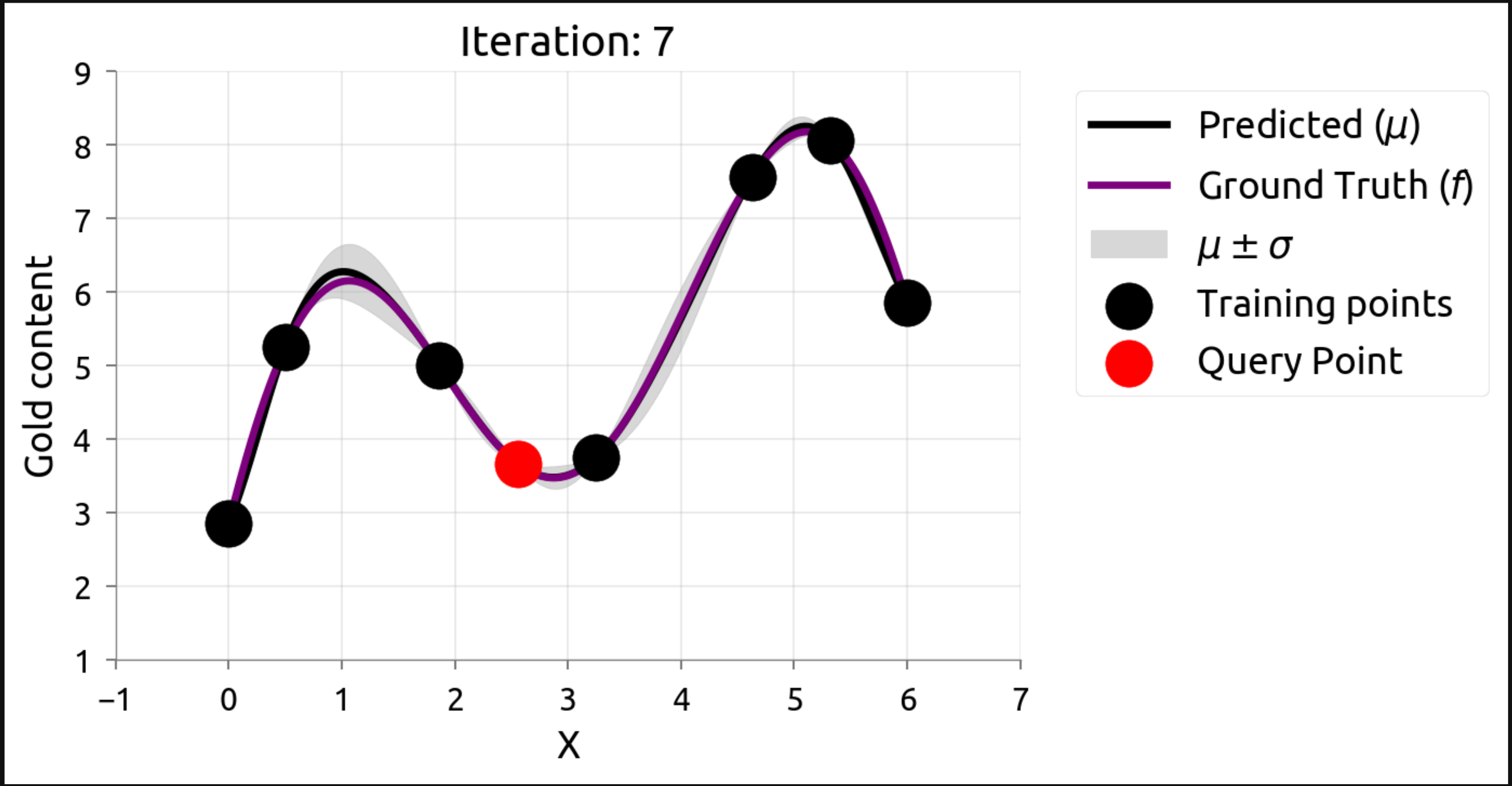

Iteration: 2
Gold content
X
Predicted (μ)
Ground Truth (f)
μ ± σ
Training points
Query Point

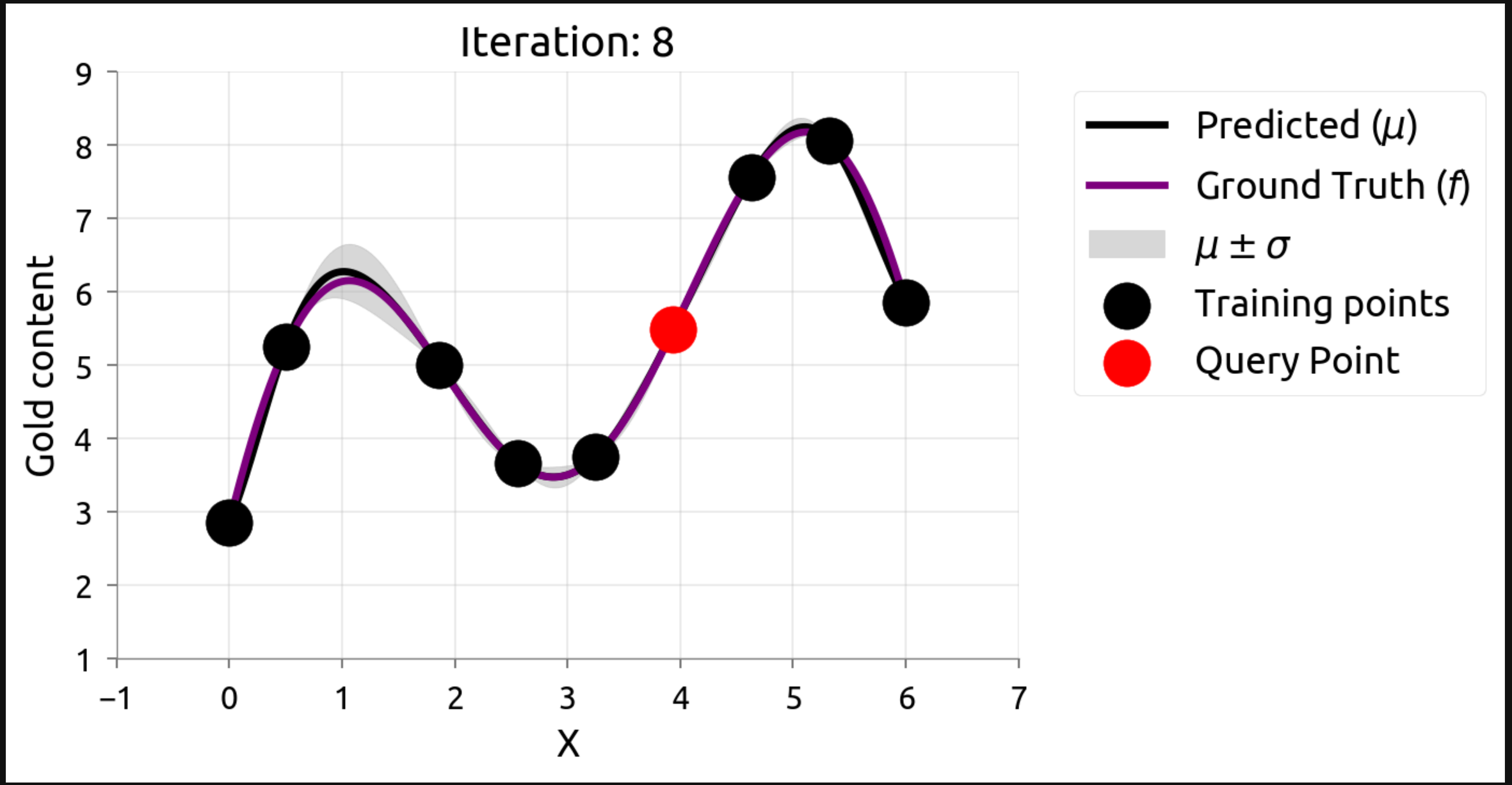

Iteration: 3
Gold content
X
9
8
7
6
5
4
3
2
1
−1
0
1
2
3
4
5
6
7
Predicted (μ)
Ground Truth (f)
μ ± σ
Training points
Query Point


Iteration: 4
Gold content
X
Predicted (μ)
Ground Truth (f)
μ ± σ
Training points
Query Point


Iteration: 5
Gold content
X
Predicted (μ)
Ground Truth (f)
μ ± σ
Training points
Query Point


Iteration: 6
Gold content
X
Predicted (μ)
Ground Truth (f)
μ ± σ
Training points
Query Point


Iteration: 7
Gold content
X
Predicted (μ)
Ground Truth (f)
μ ± σ
Training points
Query Point


Iteration: 8
Gold content
X
9
8
7
6
5
4
3
2
1
−1
0
1
2
3
4
5
6
7
Predicted (μ)
Ground Truth (f)
μ ± σ
Training points
Query Point

Bayesian Optimization

- Aim: Find the maximum of an unknown function efficiently.
- Balance exploration (unknown regions) and exploitation (known high-value regions).
- Key component: Acquisition function.

To solve this problem, we will follow the following algorithm:

1. We first choose a surrogate model for modeling the true function f and define its prior.
2. Given the set of observations (function evaluations), use Bayes rule to obtain the posterior.
3. Use an acquisition function α(x), which is a function of the posterior, to decide the next sample point
4. Add newly sampled data to the set of observations and goto step #2 till convergence or budget elapses.

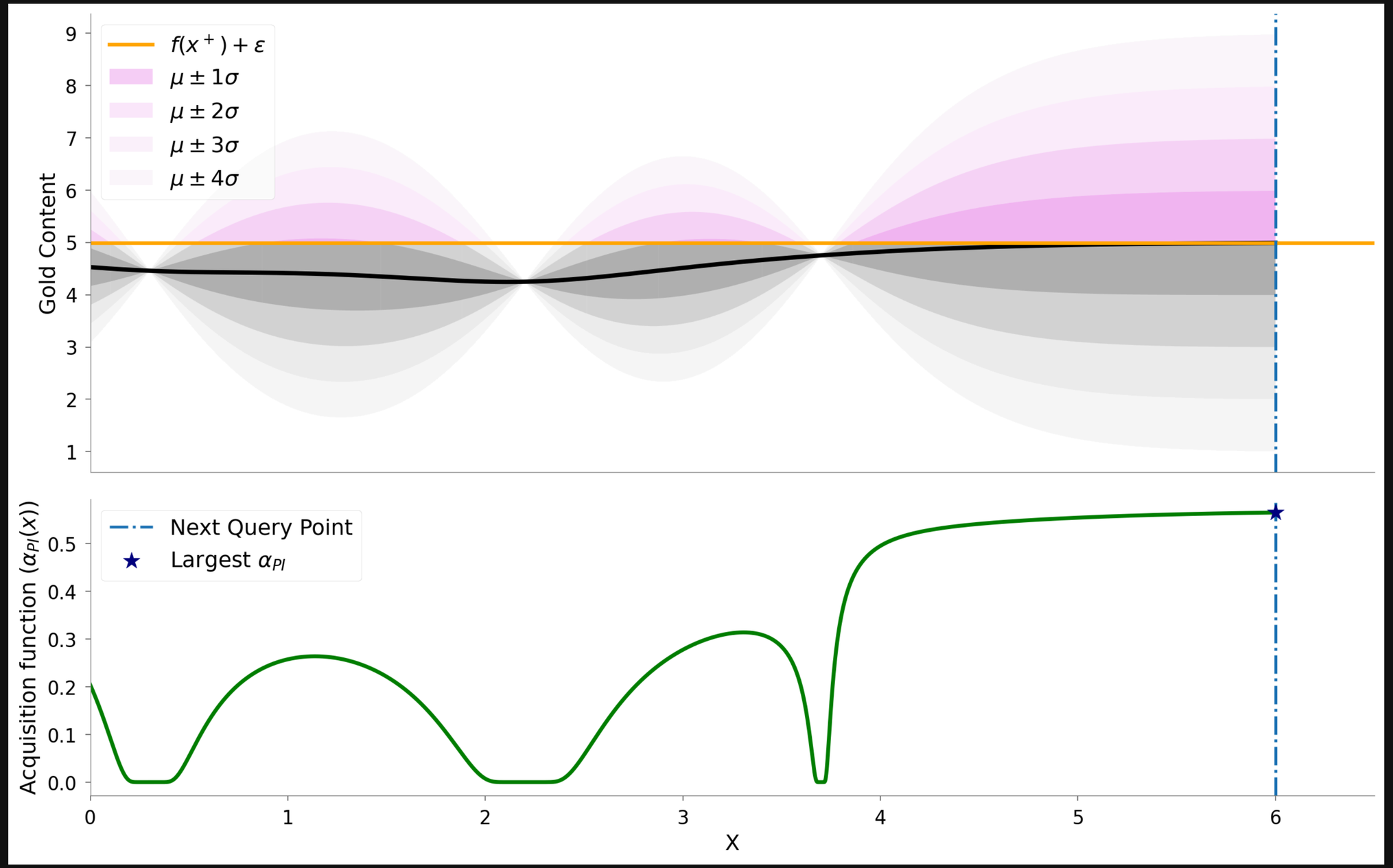
Surrogate Models and Gaussian Processes

- Surrogate models estimate the unknown function.
- GPs are flexible and provide uncertainty estimates.
- Update the surrogate model using Bayes’ rule after each evaluation.

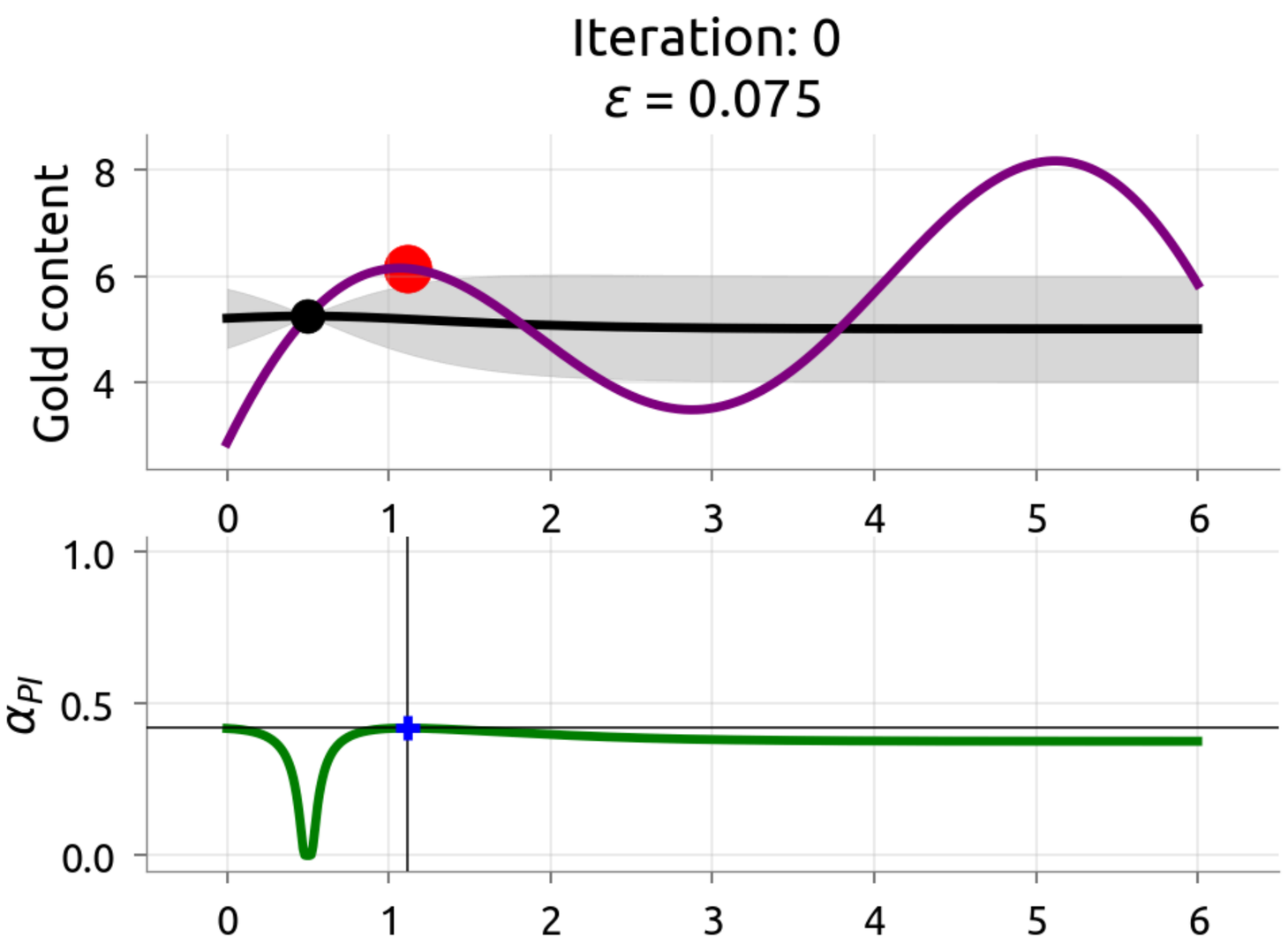
Acquisition Functions

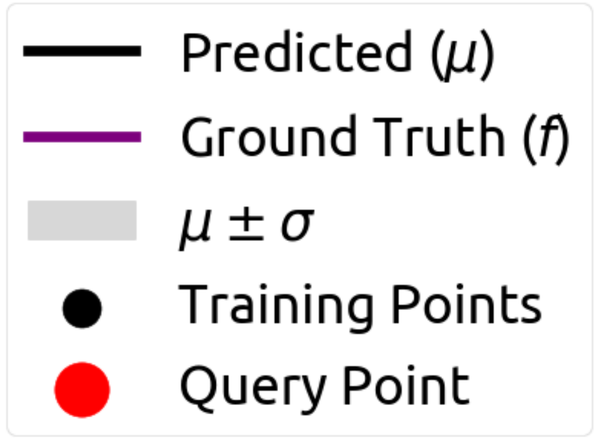
Probability of Improvement (PI)

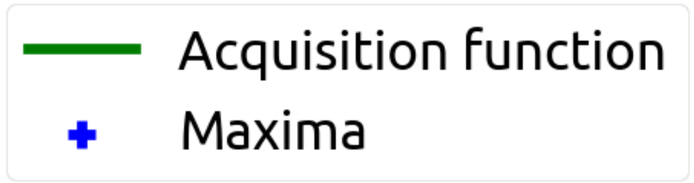
- Chooses the point with the highest probability of improvement over the current best.
- Mathematically, we write the selection of next point as follows:
  $x_{t+1} = argmax(\alpha_{PI}(x)) = argmax(P(f(x) \geq (f(x^+) + \epsilon)))$
- we are just finding the upper-tail probability (or the CDF) of the surrogate posterior. Moreover, if we are using a GP as a surrogate the expression above converts to
- $x^+ = \text{argmax}_{x_i \in x_{1:t}} f(x_i)$

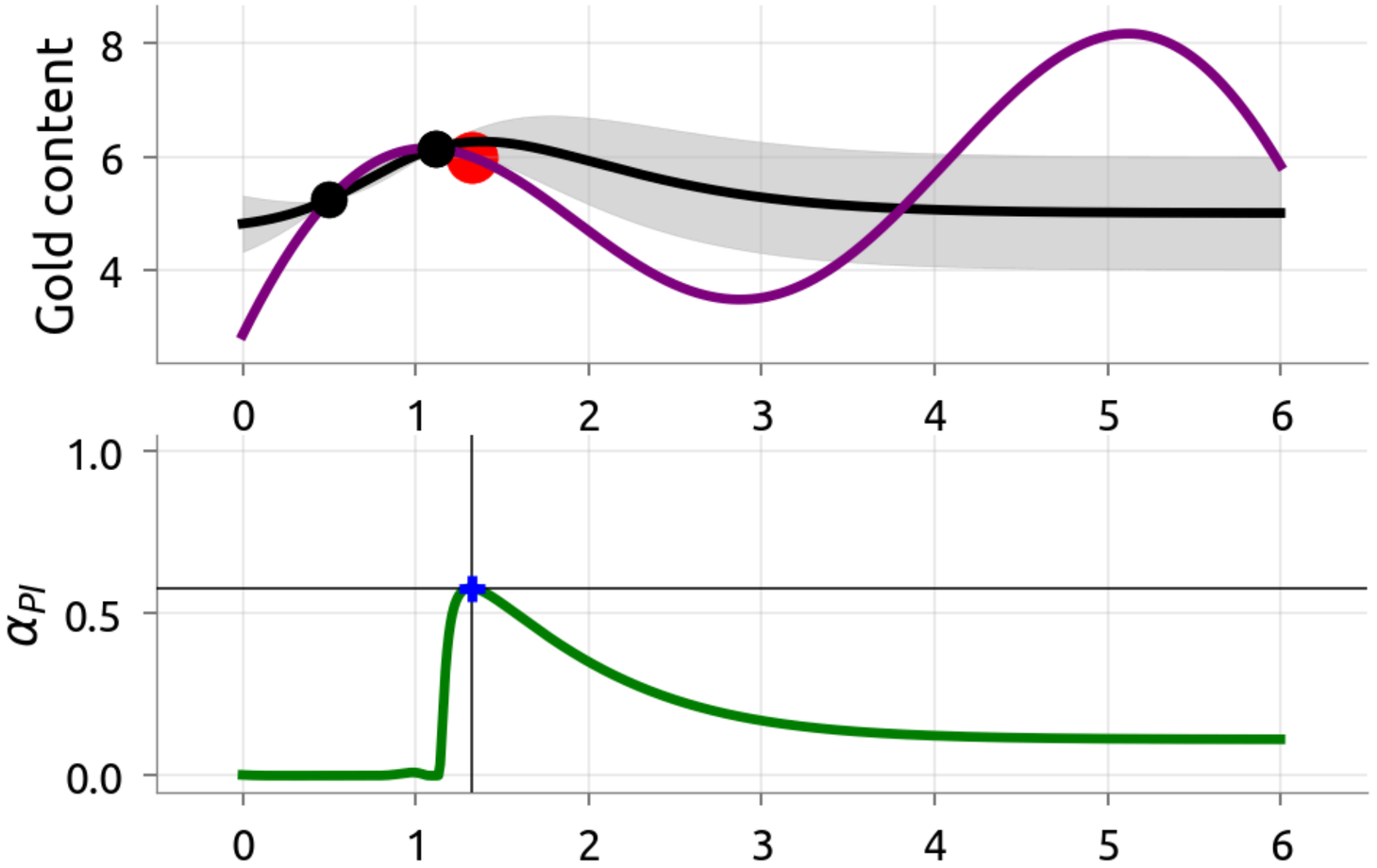

$f(x^+) + \varepsilon$
$\mu \pm 1\sigma$
$\mu \pm 2\sigma$
$\mu \pm 3\sigma$
$\mu \pm 4\sigma$
Gold Content
Next Query Point
Largest α_{PI}
Acquisition function ($\alpha_{PI}(x)$)
X

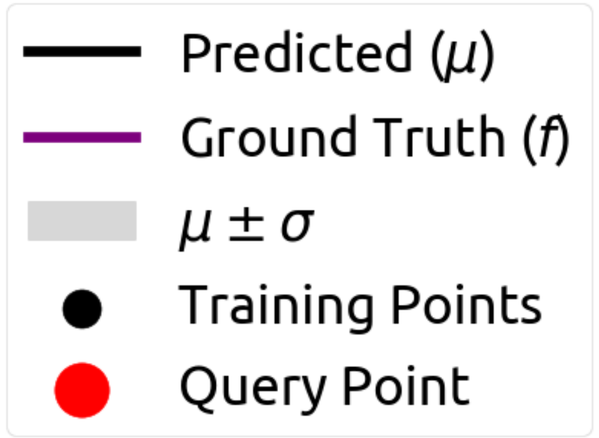
Intuition behind ϵ in PI: ϵ = 0.075

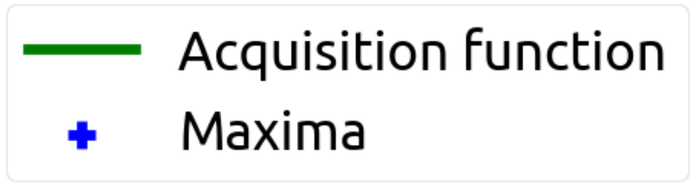


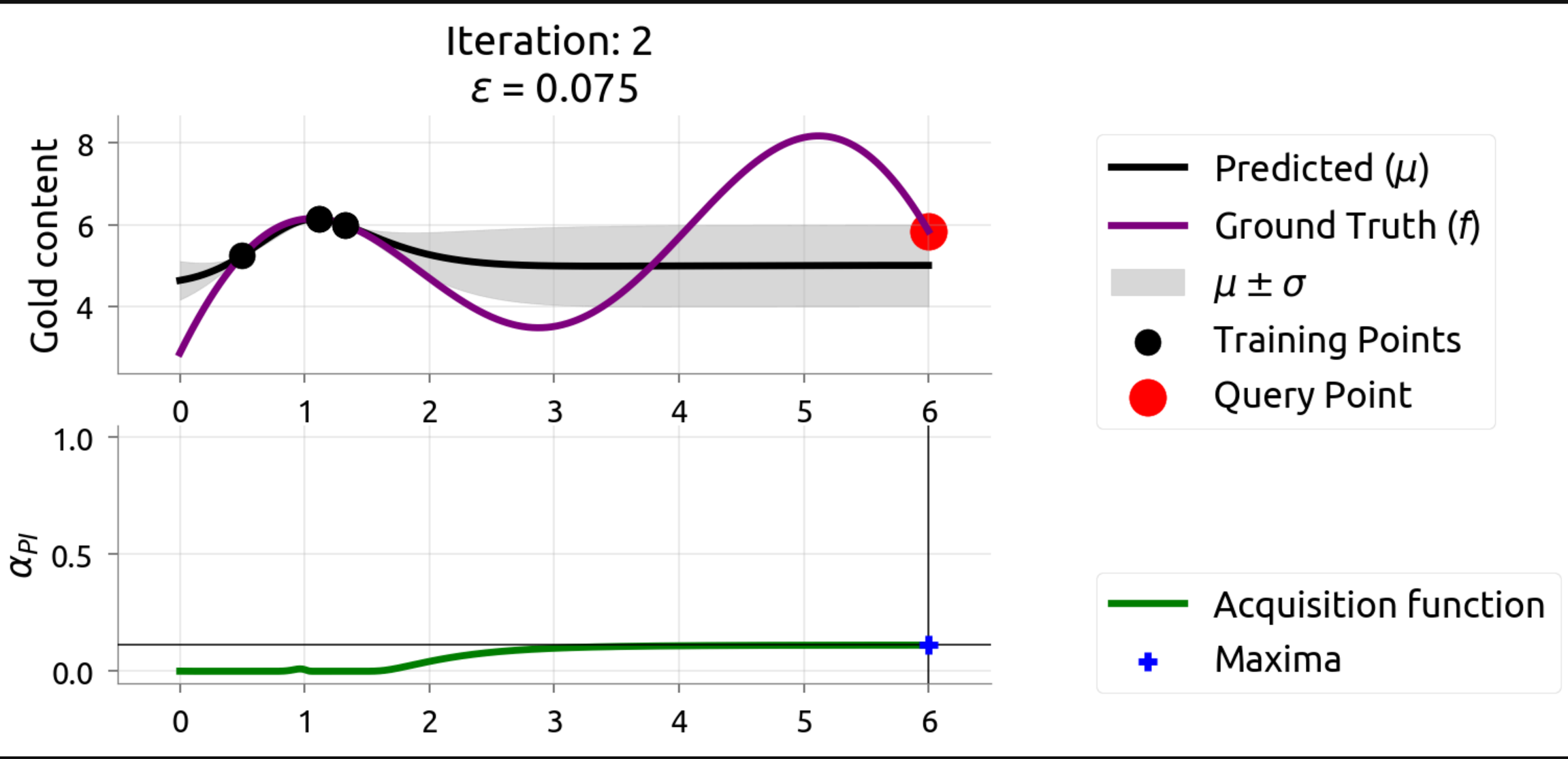


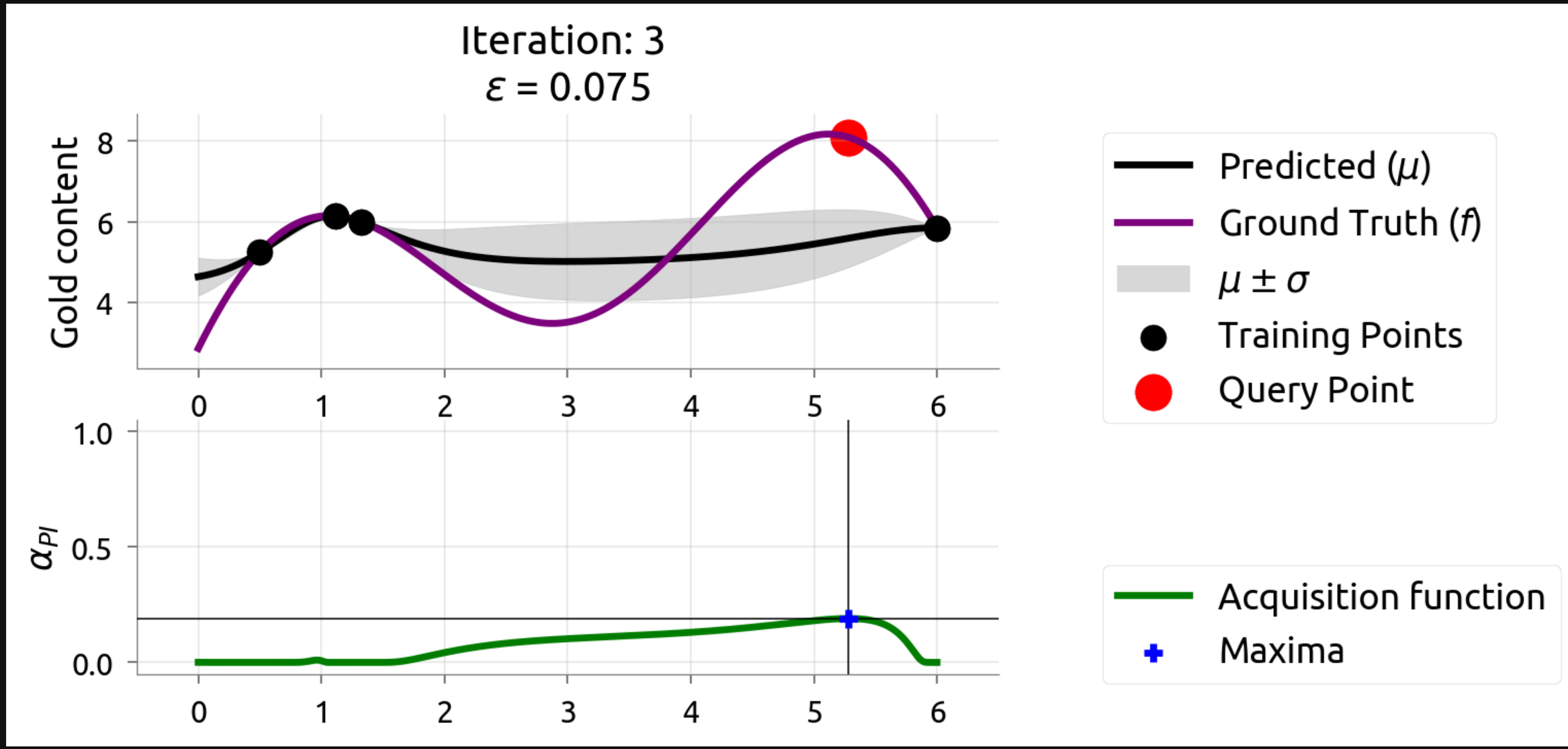


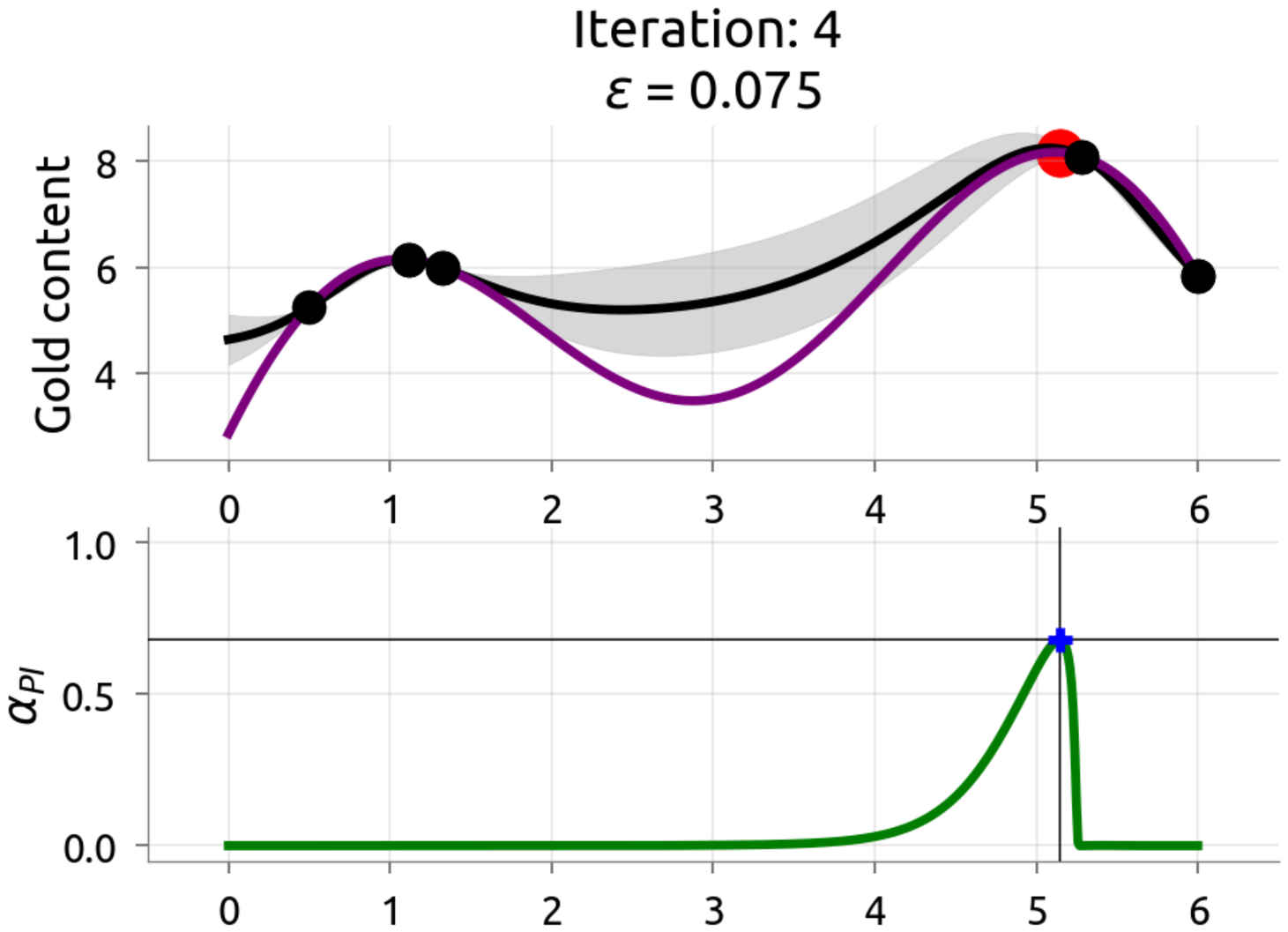

Iteration: 1
ε = 0.075
Gold content
8
6
4
0
1
2
3
4
5
6
1.0
0.5
0.0
α_{PI}
0
1
2
3
4
5
6

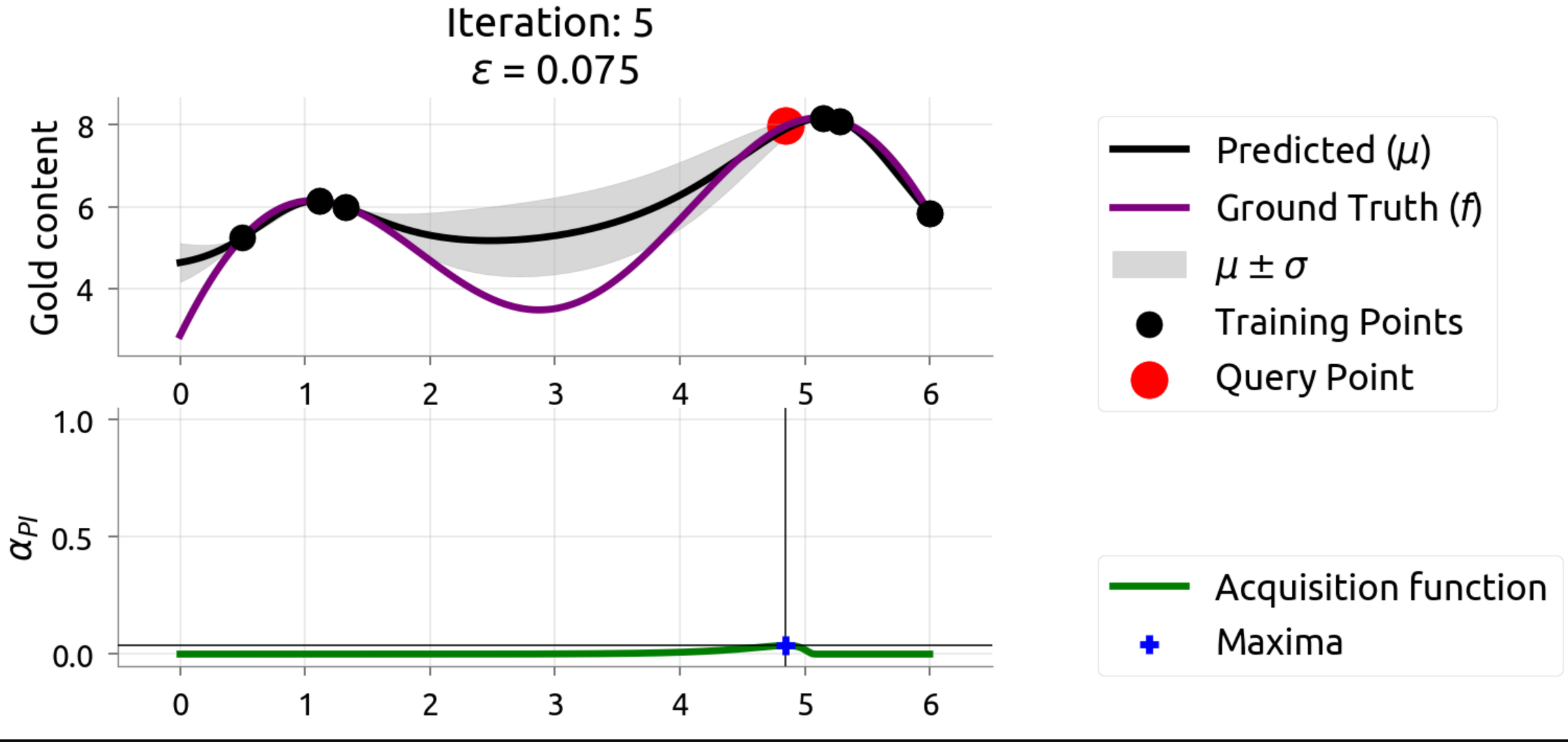

Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point

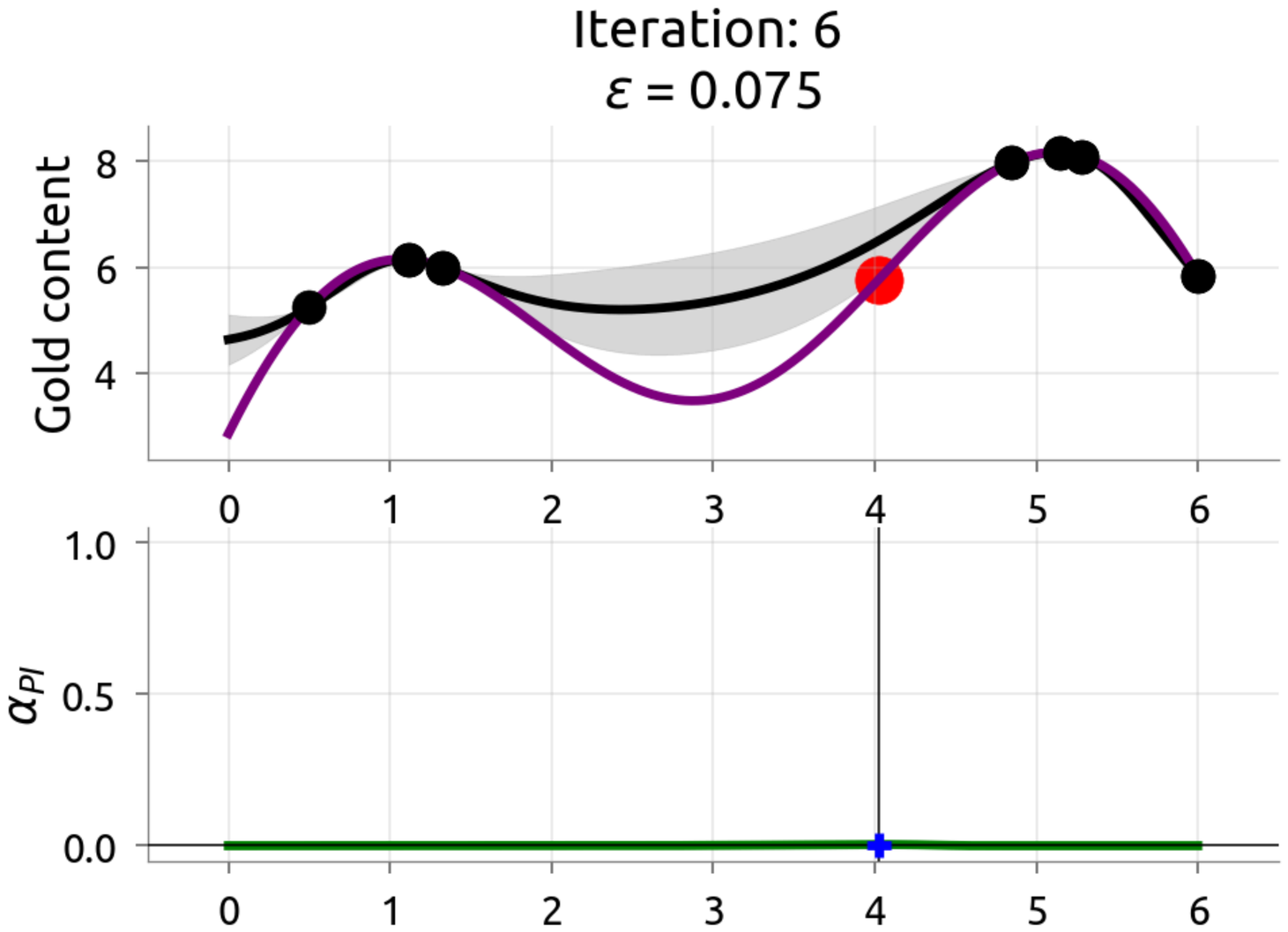

Acquisition function
Maxima

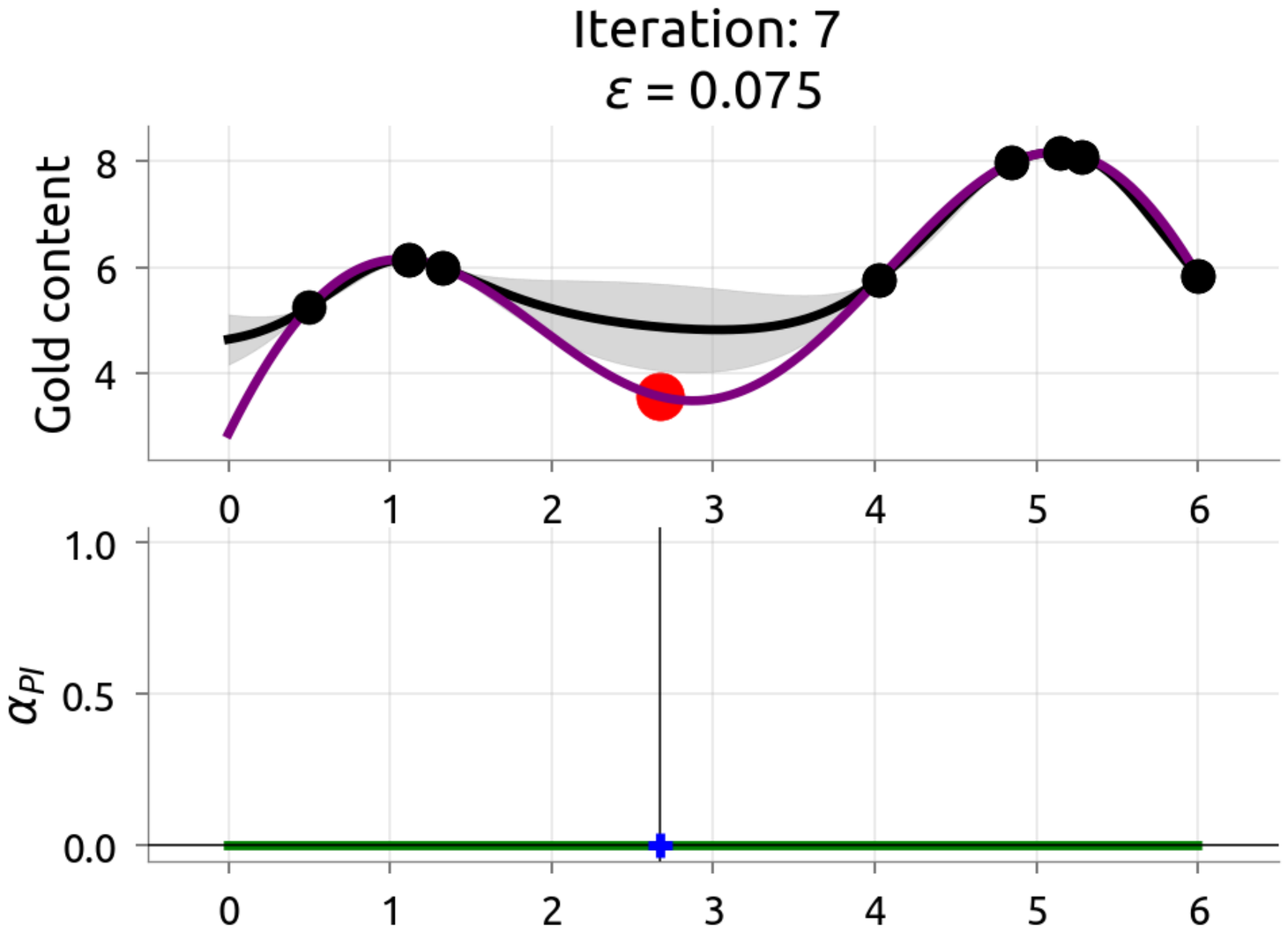

Iteration: 2
ε = 0.075
Gold content
8
6
4
0
1
2
3
4
5
6
1.0
0.5
0.0
α_{PI}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

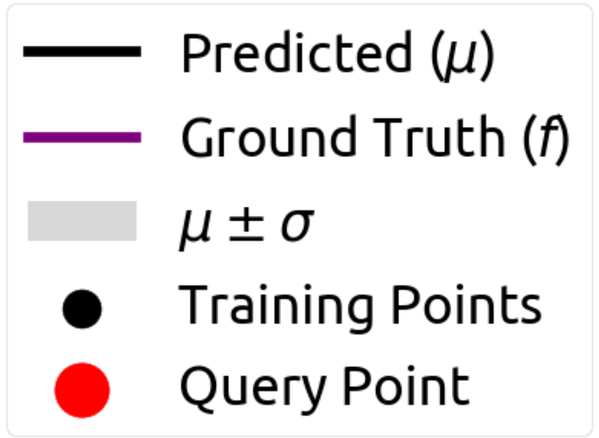

Iteration: 3
ε = 0.075
Gold content
8
6
4
0
1
2
3
4
5
6
1.0
0.5
0.0
α_{PI}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

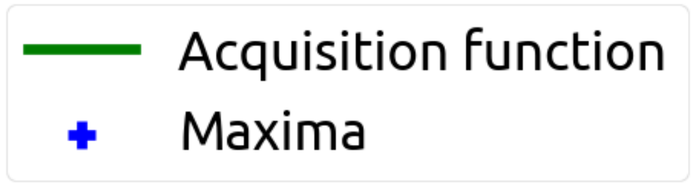

Iteration: 4
ε = 0.075
Gold content
8
6
4
0
1
2
3
4
5
6
1.0
0.5
0.0
α_{PI}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

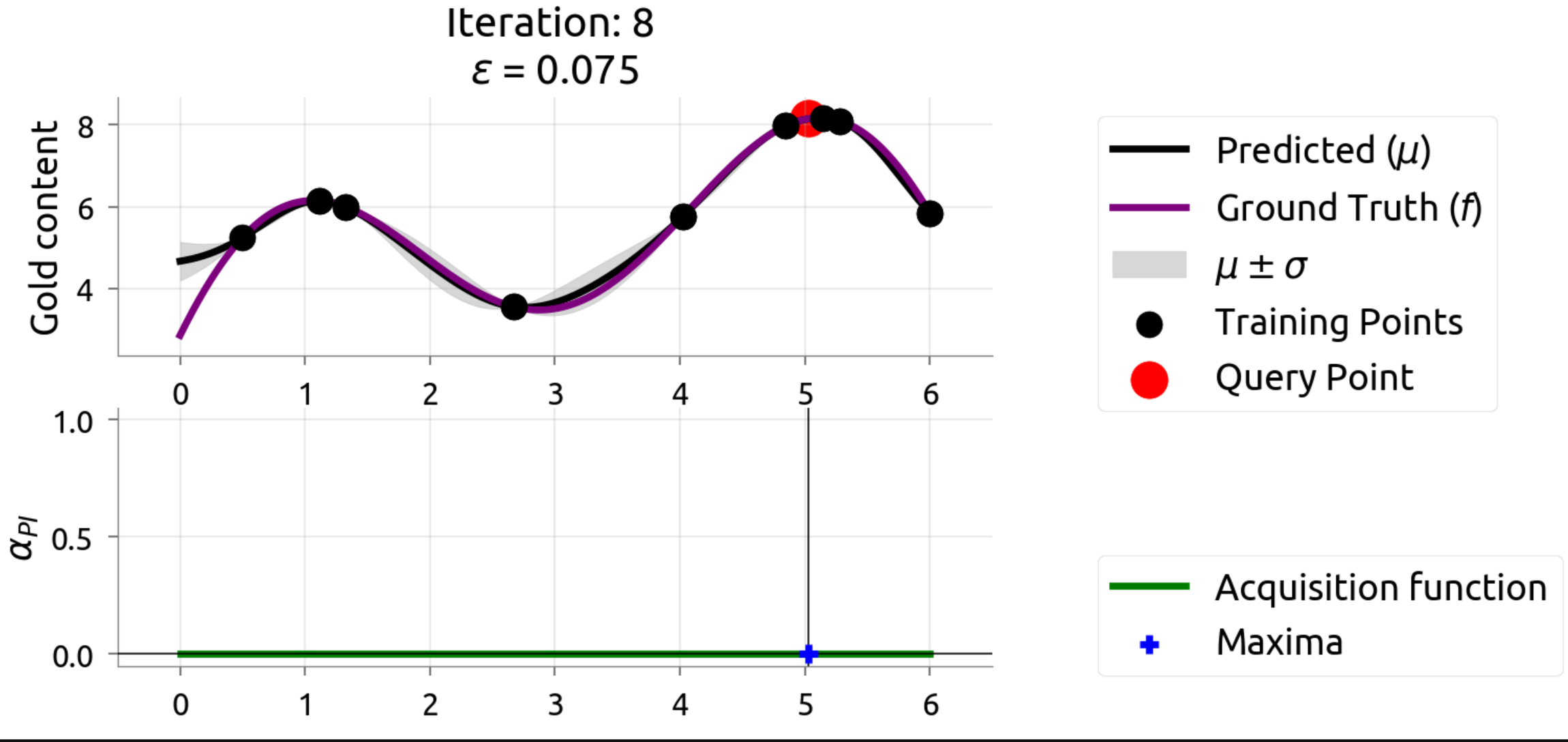

Iteration: 5
ε = 0.075
Gold content
8
6
4
0
1
2
3
4
5
6
1.0
0.5
0.0
α_{PI}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

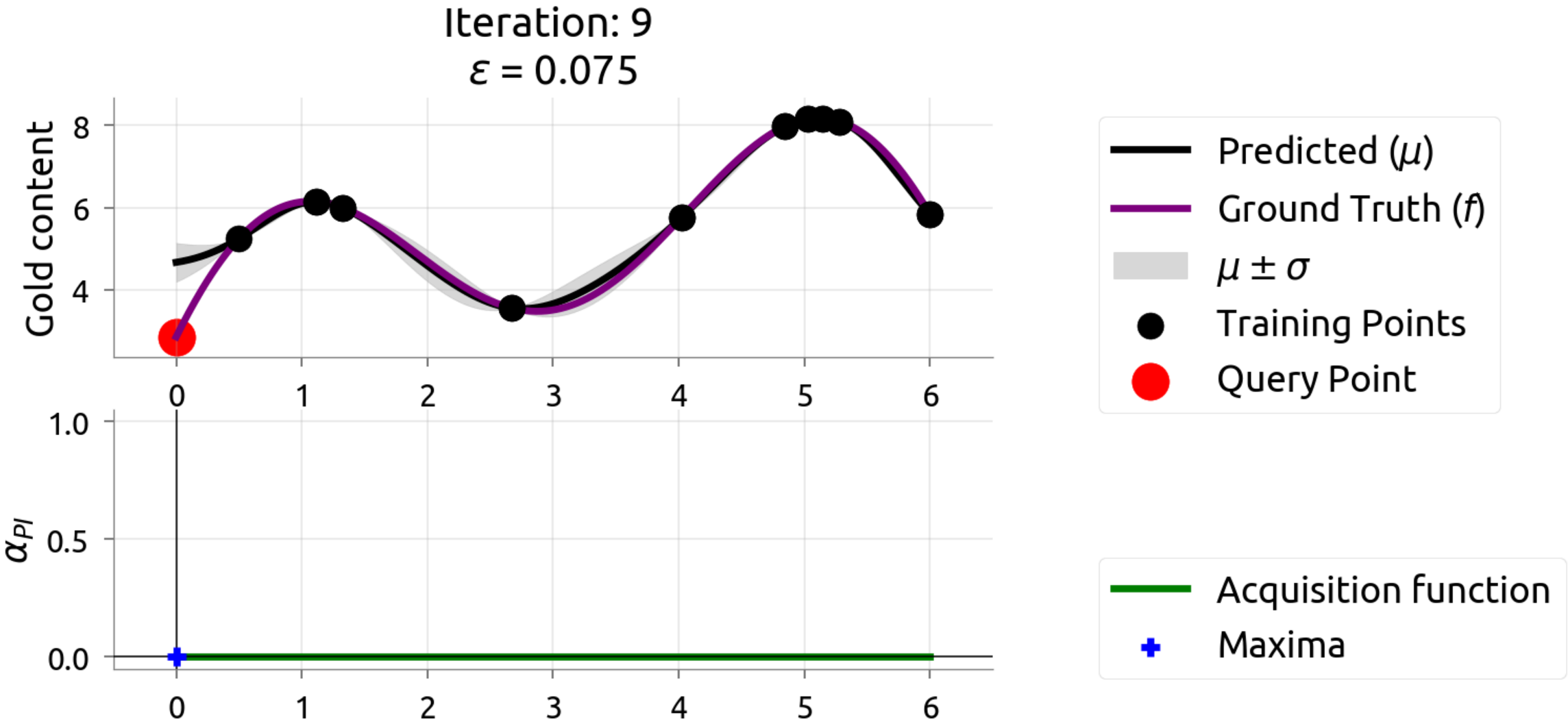

Iteration: 6
ε = 0.075
Gold content
8
6
4
0
1
2
3
4
5
6
1.0
0.5
0.0
α_{PI}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima


Iteration: 7
ε = 0.075
Gold content
8
6
4
0
1
2
3
4
5
6
1.0
0.5
0.0
α_{PI}
0
1
2
3
4
5
6


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

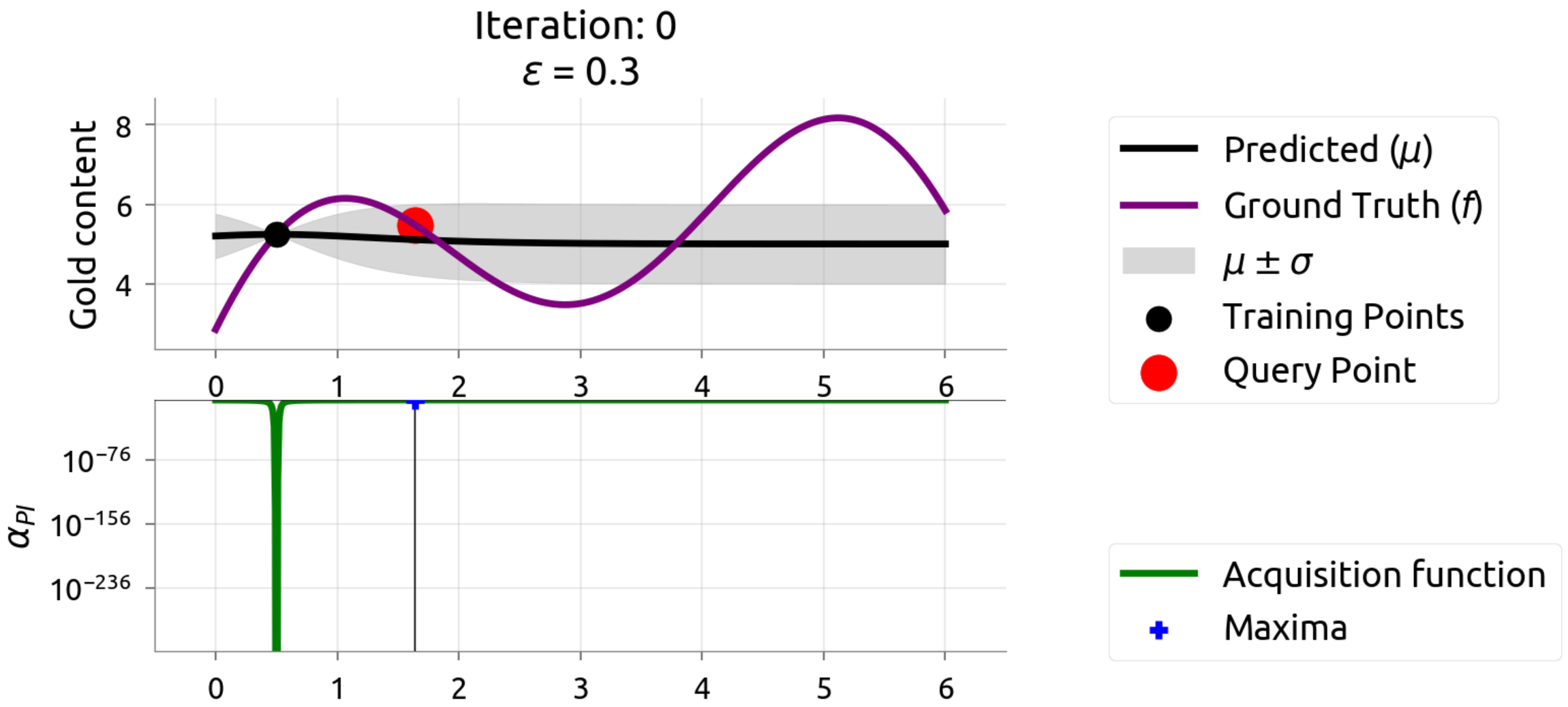

Acquisition function
Maxima

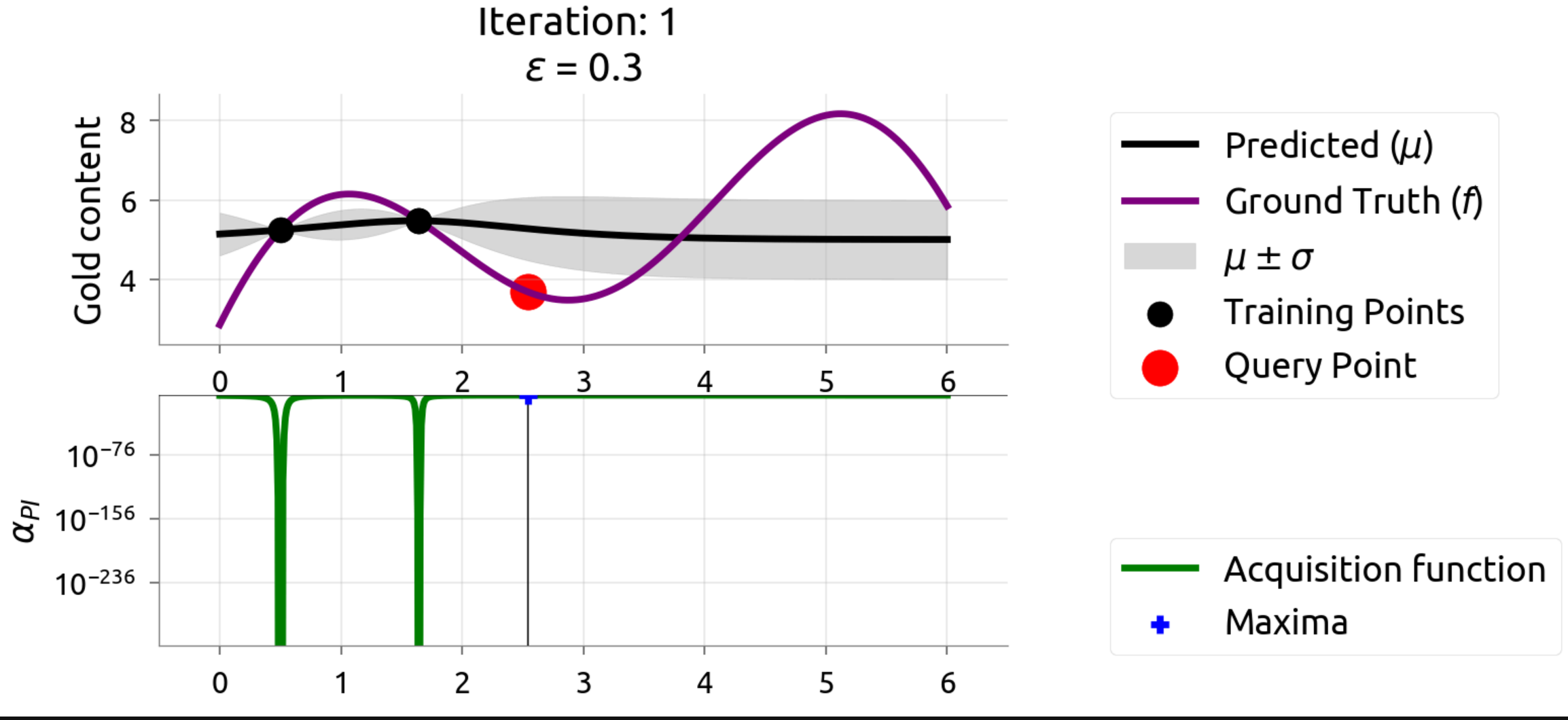

Iteration: 8
ε = 0.075
Gold content
8
6
4
0
1
2
3
4
5
6
1.0
0.5
0.0
α_{PI}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

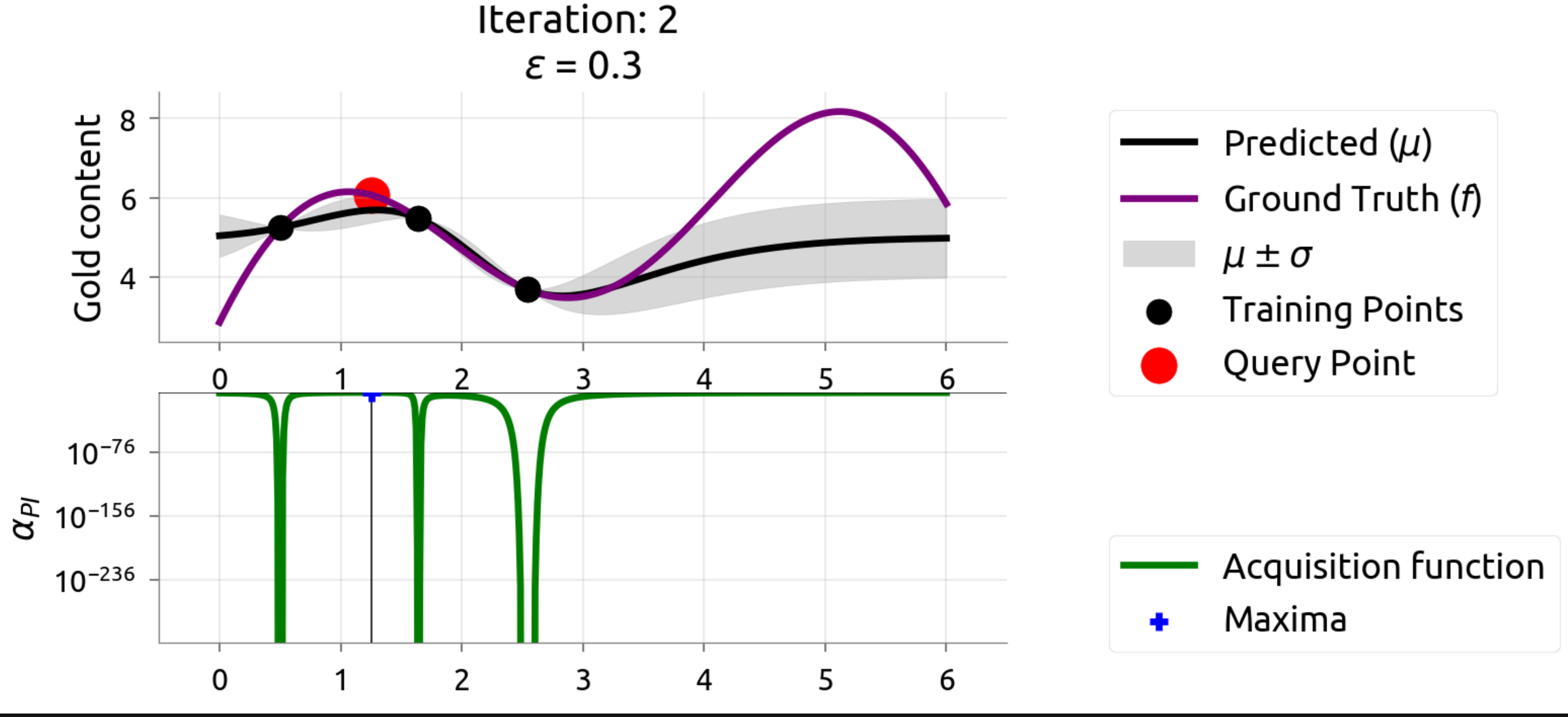

Iteration: 9
ε = 0.075
Gold content
8
6
4
0
1
2
3
4
5
6
α_{PI}
1.0
0.5
0.0
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

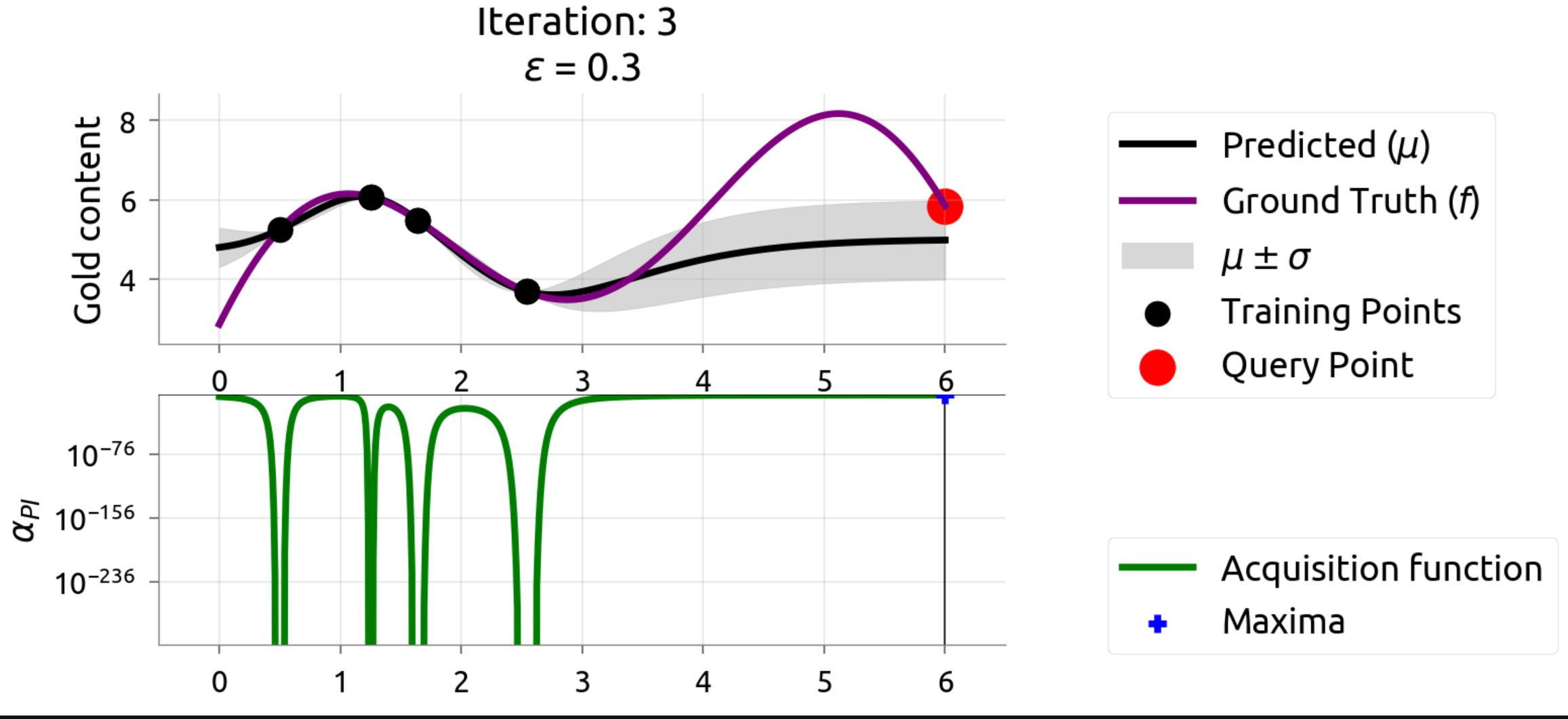
- Looking at the graph above, we see that we reach the global maxima in a few iterations .
- Our surrogate possesses a large uncertainty in $x \in [2,4]$ in the first few iterations
- acquisition function initially exploits regions with a high promise , which leads to high uncertainty in the region $x \in [2,4]$.
- observation also shows that we do not need to construct an accurate estimate of the black-box function to find its maximum.

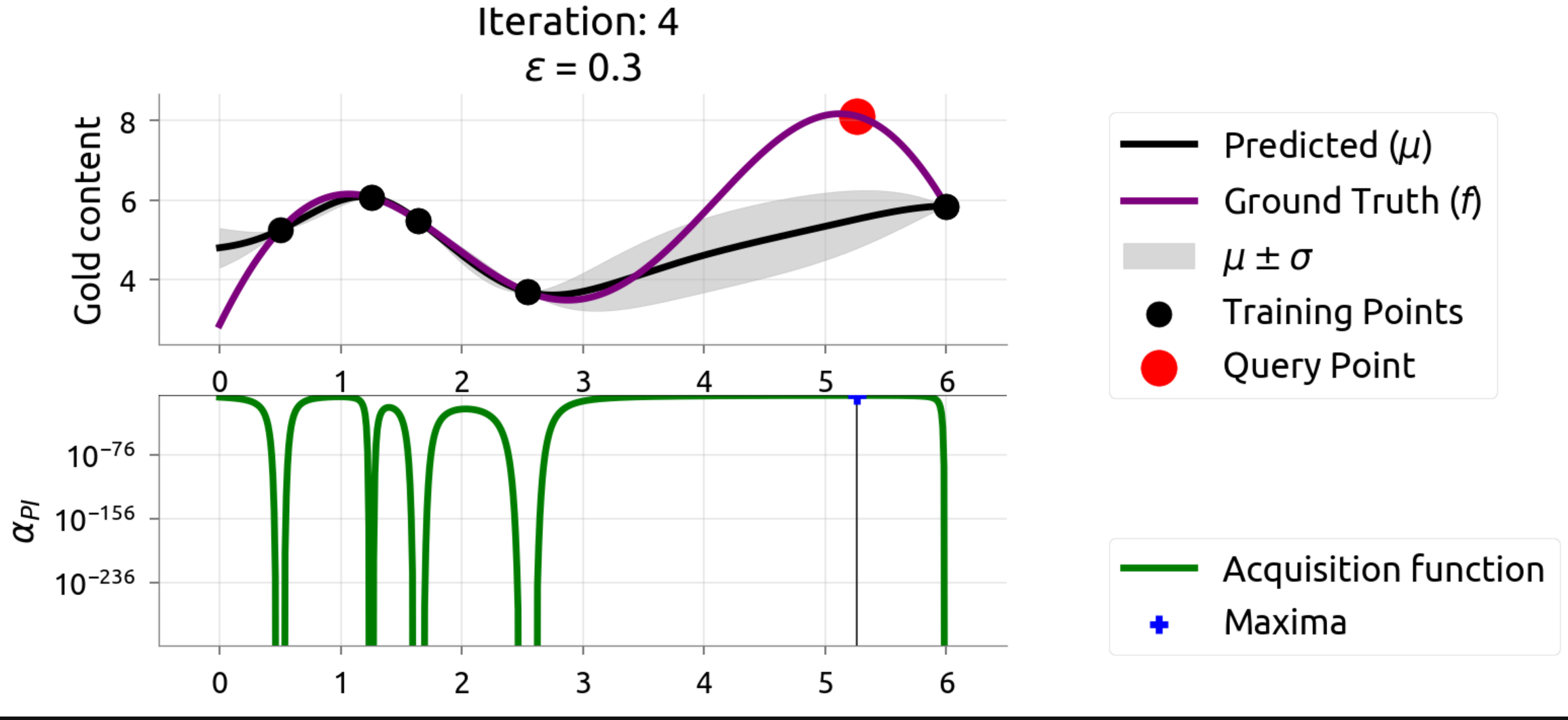
Intuition behind ϵ in PI: ϵ = 0.3

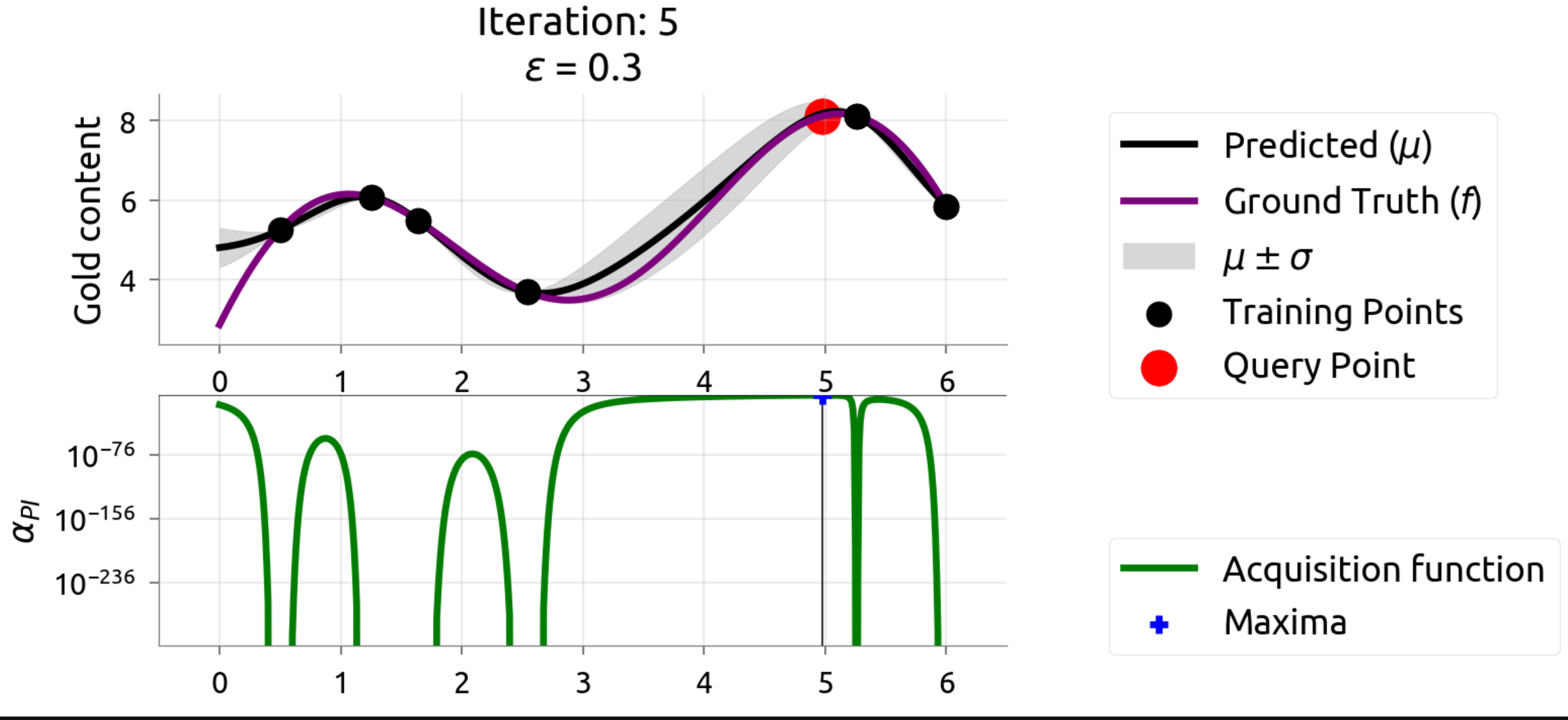


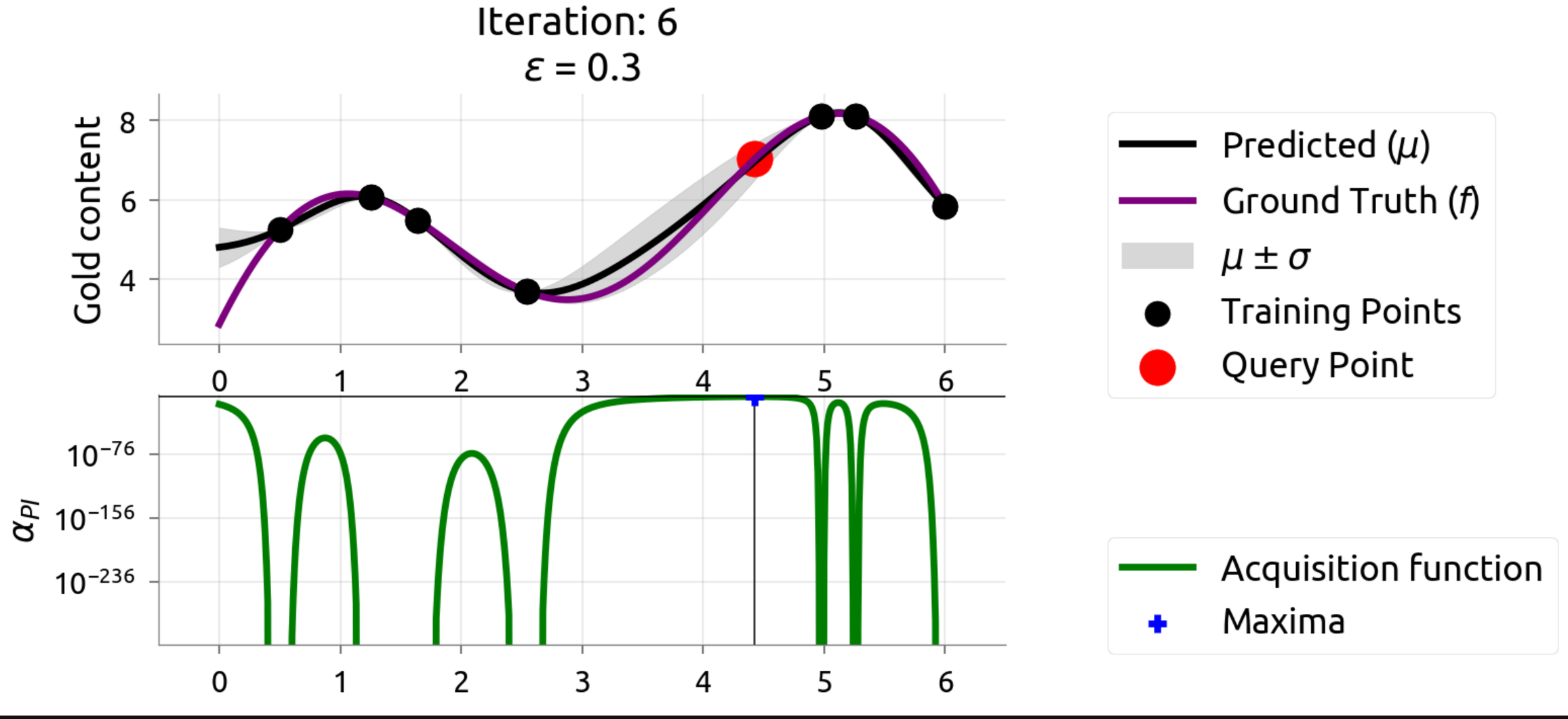

Iteration: 1
$\varepsilon = 0.3$
Gold content
8
6
4
0
1
2
3
4
5
6
α_{PI}
10^{-76}
10^{-156}
10^{-236}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

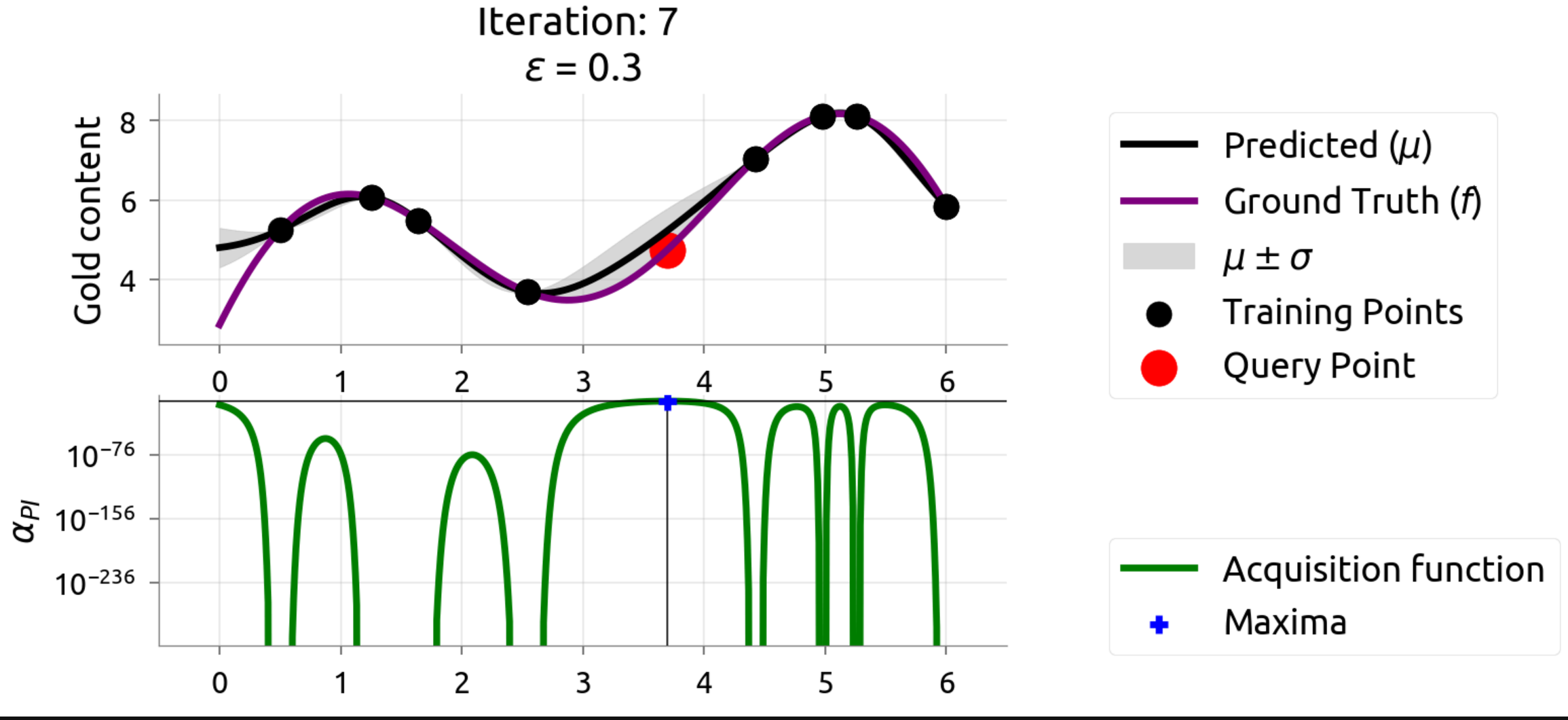

Iteration: 2
$\varepsilon = 0.3$
Gold content
8
6
4
0
1
2
3
4
5
6
α_{PI}
10^{-76}
10^{-156}
10^{-236}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

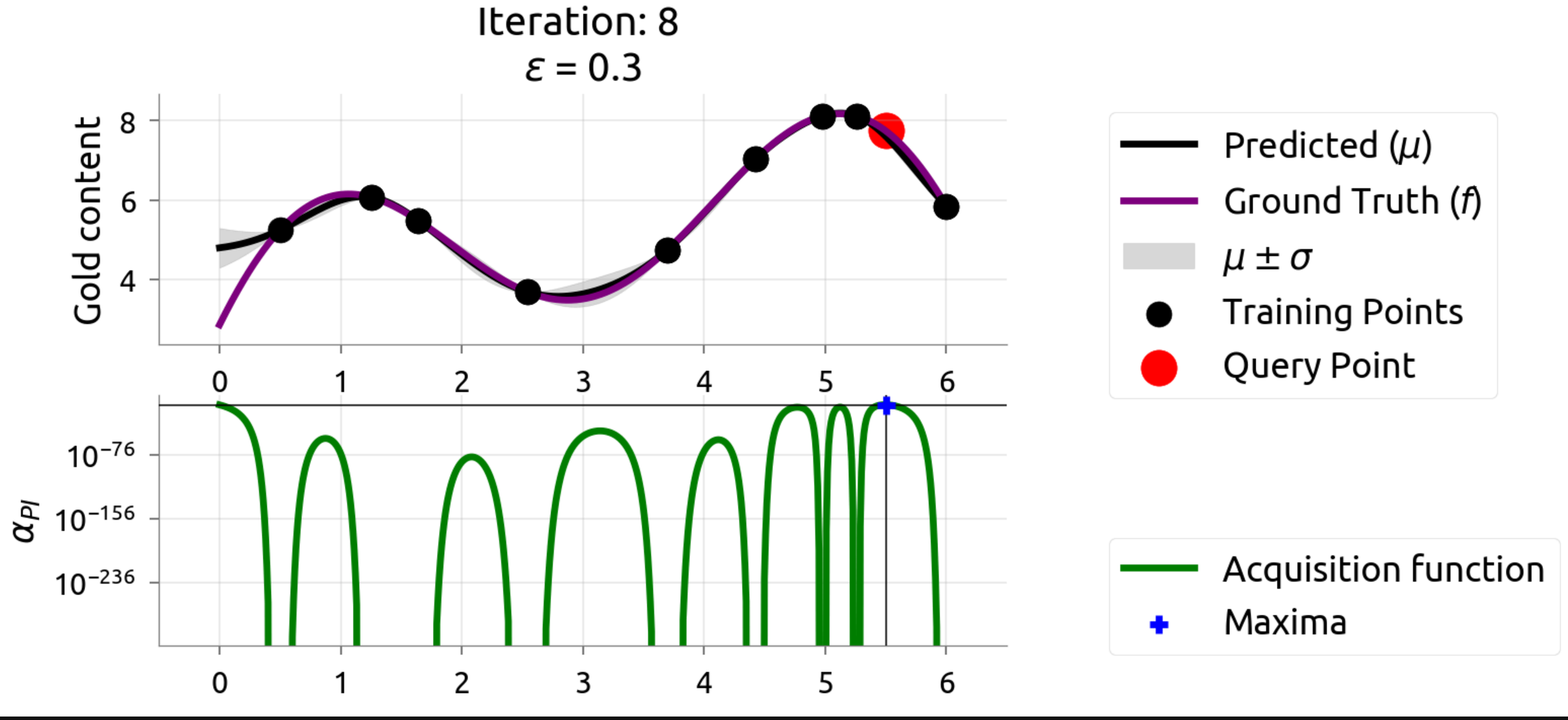

Iteration: 3
$\varepsilon = 0.3$
Gold content
8
6
4
0
1
2
3
4
5
6
α_{PI}
10^{-76}
10^{-156}
10^{-236}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

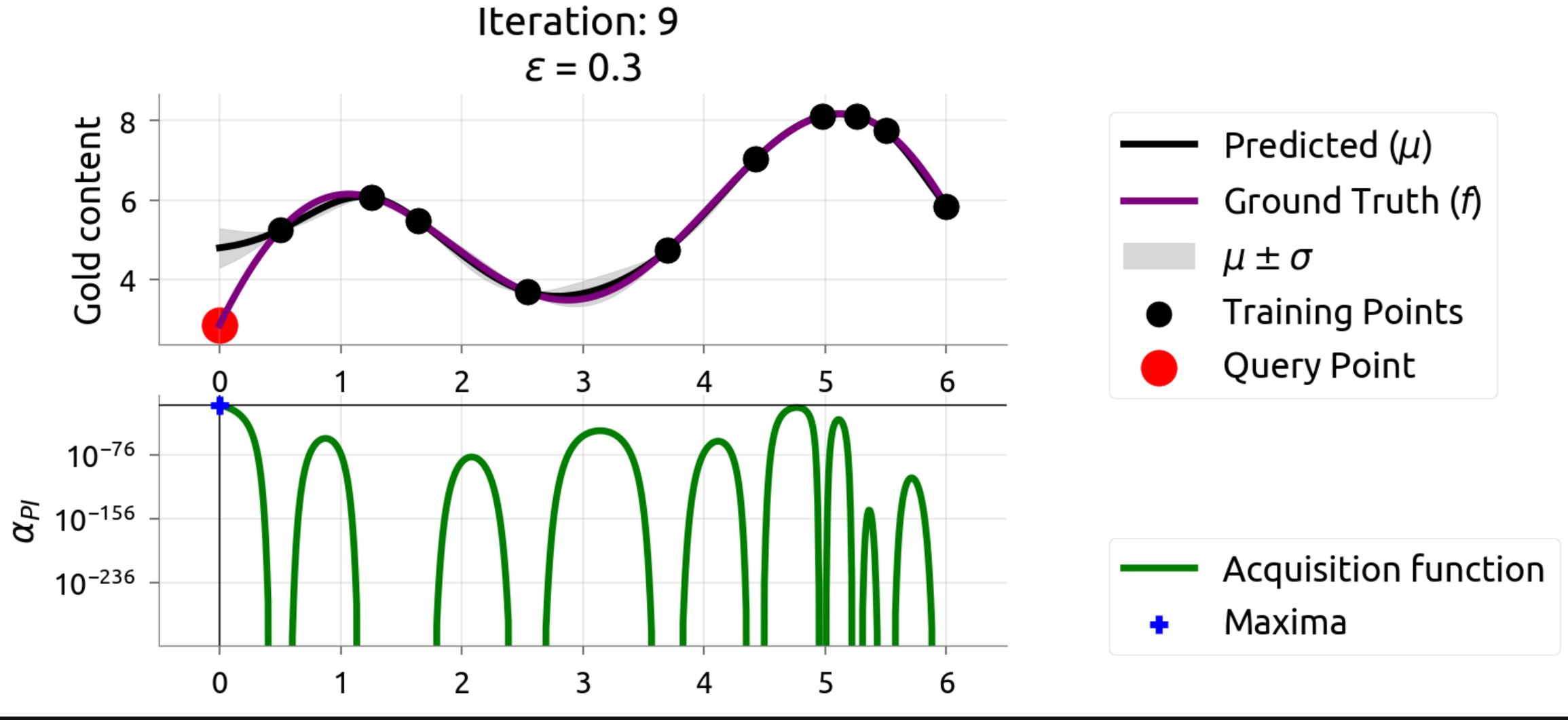

Iteration: 4
$\varepsilon = 0.3$
Gold content
8
6
4
0
1
2
3
4
5
6
α_{PI}
10^{-76}
10^{-156}
10^{-236}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima


Iteration: 5
$\varepsilon = 0.3$
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


Iteration: 6
$\varepsilon = 0.3$
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


Iteration: 7
$\varepsilon = 0.3$
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


Iteration: 8
$\varepsilon = 0.3$
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


Iteration: 9
$\varepsilon = 0.3$
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

- We see that we made things worse!
- Our model now uses ϵ=3, and we are unable to exploit when we land near the global maximum. Moreover, with high exploration, the setting becomes similar to active learning.
- Our quick experiments above help us conclude that ϵ controls the degree of exploration in the PI acquisition function.

Expected Improvement: Introduction

- Probability of improvement considers *how likely* an improvement is.
- Expected Improvement (EI) considers *how much* we can improve.
- **Key Idea**: Choose the next query point with the highest expected improvement over the current max $f(x^+)$.

EI: Mathematical Formulation

- **Equation**: $x_{t+1} = \arg\min_x \mathbb{E}\left(||h_{t+1}(x) - f(x^\star)|| \mid \mathcal{D}_t\right)$
- **Components**:
  - f: Actual ground truth function.
  - h_{t+1}: Posterior mean of the surrogate at $t + 1^{th}$ timestep.
  - $\mathcal{D}_t$: Training data.
  - $x^\star$: Actual position where f takes the maximum value.

EI: Mockus' Acquisition Function

- we are trying to select the point that minimizes the distance to the objective evaluated at the maximum
- Unfortunately, we do not know the ground truth function, f, so we need to estimate it.
- Mockus[8] proposed the following acquisition function to overcome the issue
- **Equation**: $x_{t+1} = \text{argmax}_x \mathbb{E}\left(max\{0,\ h_{t+1}(x) - f(x^+)\} \mid \mathcal{D}_t\right)$
- $f(x^+)$: Maximum value encountered so far.

EI: Analytical Expression for GP Surrogate

- $EI(x) = \begin{cases} (\mu_t(x) - f(x^+) - \epsilon)\Phi(Z) + \sigma_t(x)\phi(Z), & \text{if } \sigma_t(x) > 0 \\ 0, & \text{if } \sigma_t(x) = 0 \end{cases}$

$$Z = \frac{\mu_t(x) - f(x^+) - \epsilon}{\sigma_t(x)}$$

where Φ and ϕ are the cumulative distribution function and probability density function of the standard normal distribution, respectively.

EI: When is EI High?

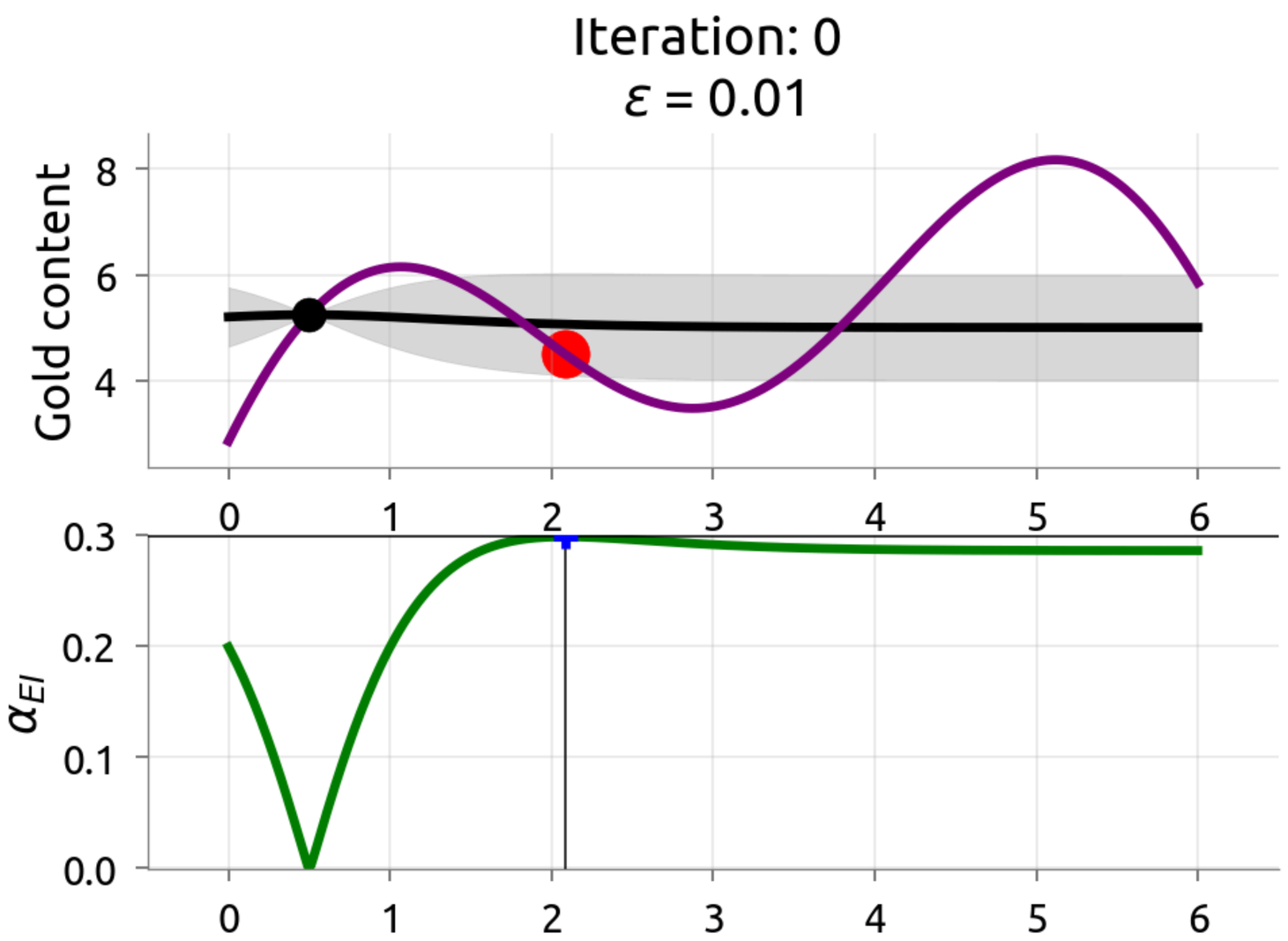
- EI is high when:
  - The expected value of $\mu_t(x) - f(x^+)$ is high.
  - The uncertainty $\sigma_t(x)$ around a point is high.

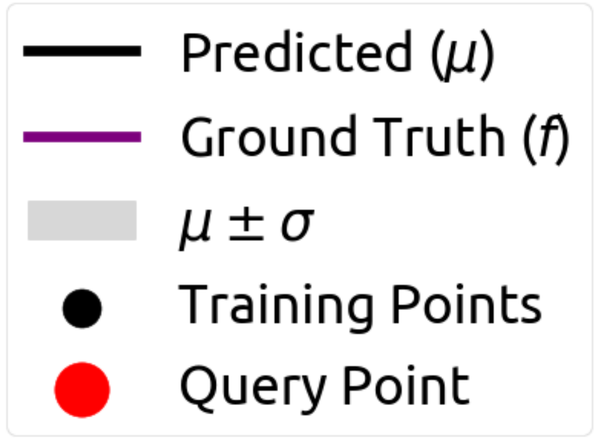
EI: Moderating Exploration with ϵ

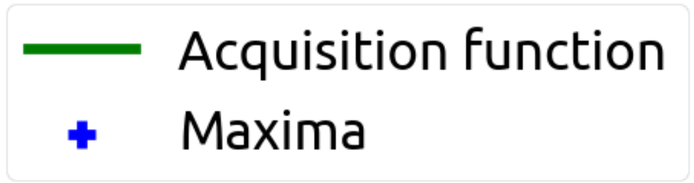
- Adjusting ϵ moderates exploration.
- **Examples**:
  - $\epsilon = 0.01$: Close to the global maxima in few iterations.
  - $\epsilon = 0.3$: More exploration, less exploitation near the global maxima.
  - $\epsilon = 3$: Too much exploration, quick reach near the global maxima, less exploitation.

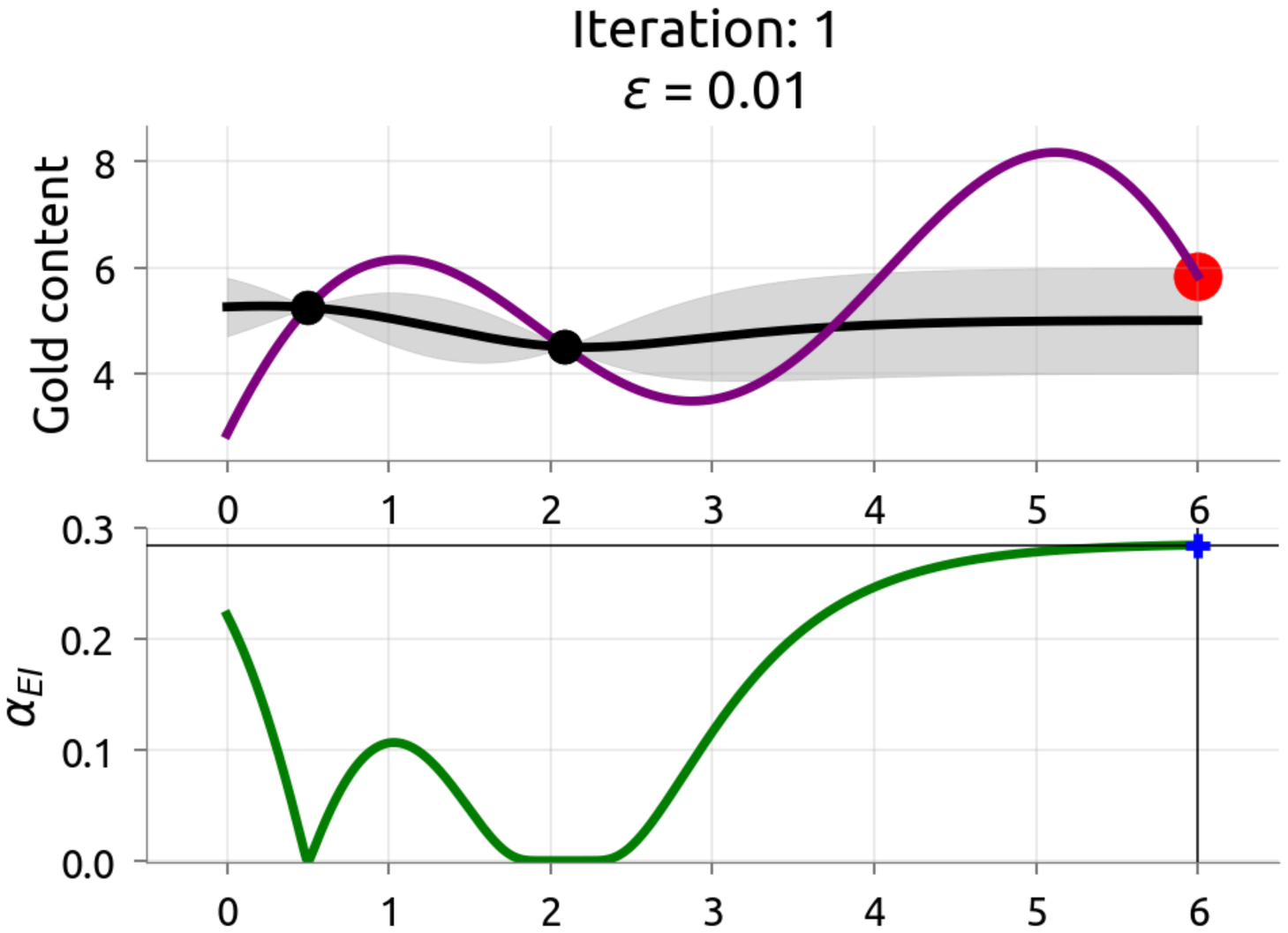
EI: Visualizations

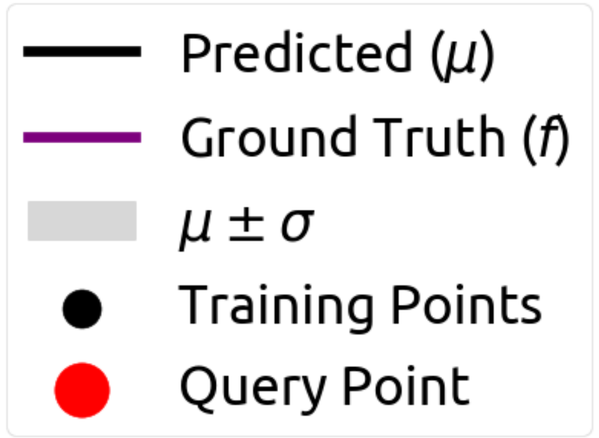
Intuition behind ϵ in EI: ϵ = 0.3

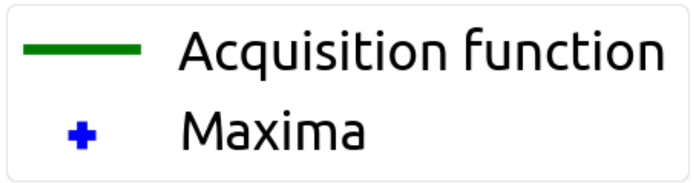


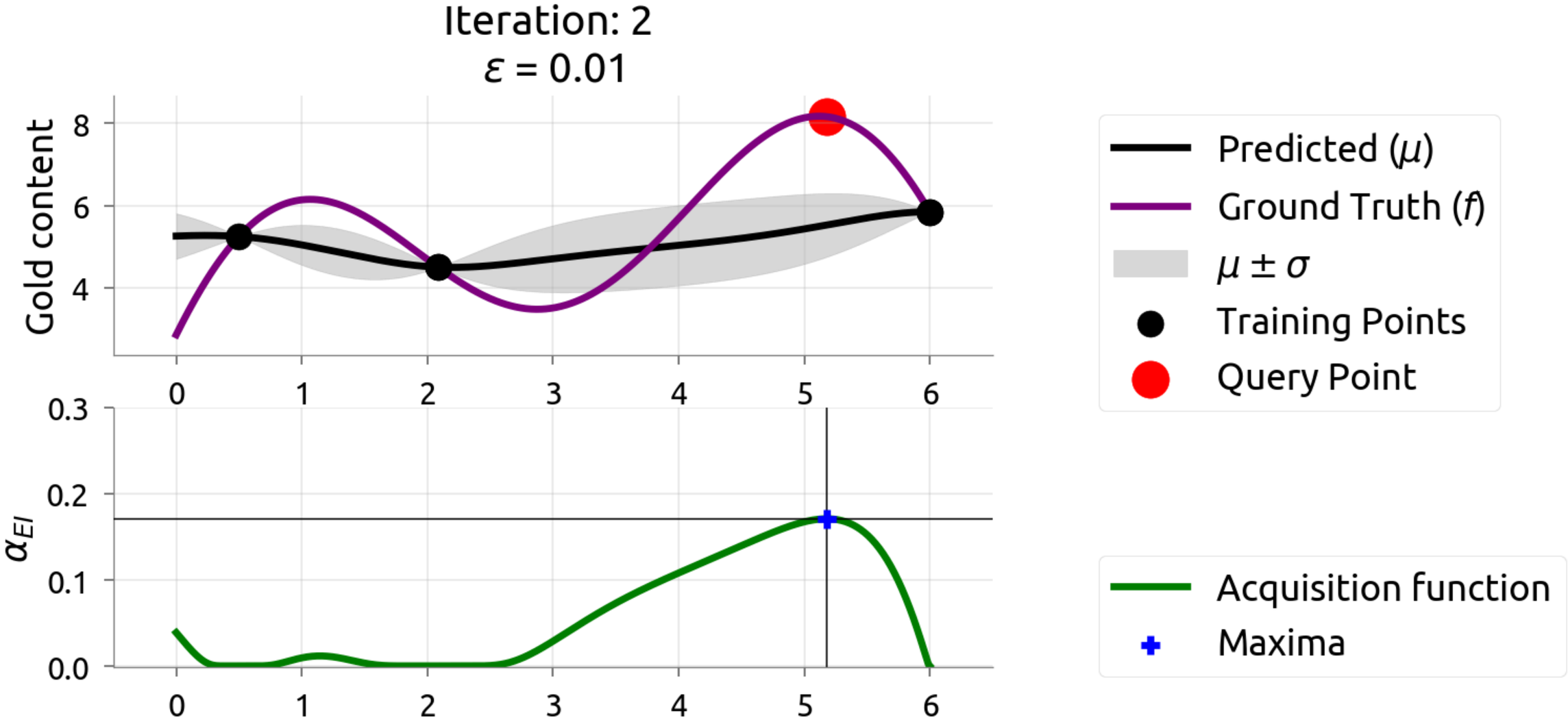


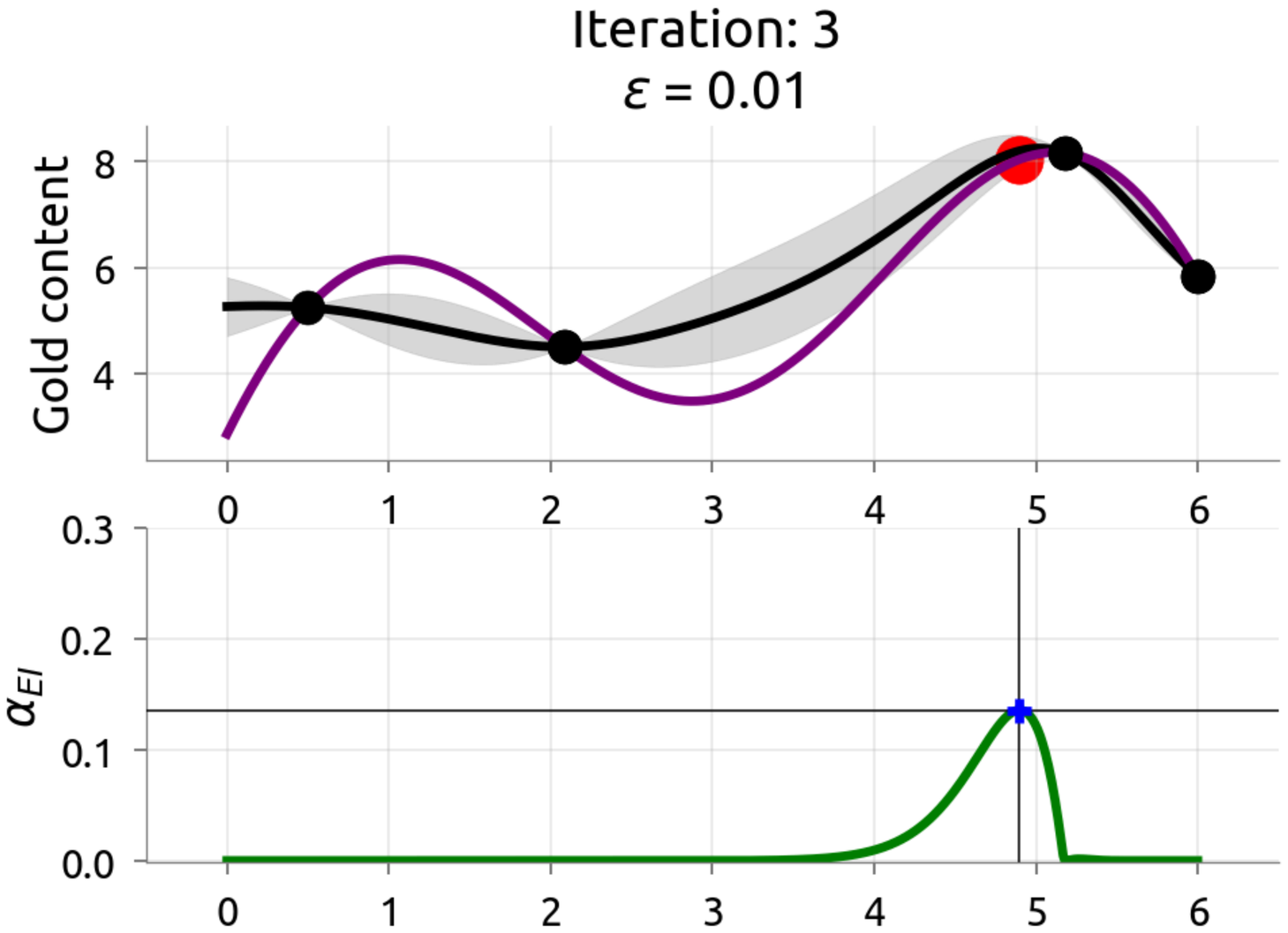


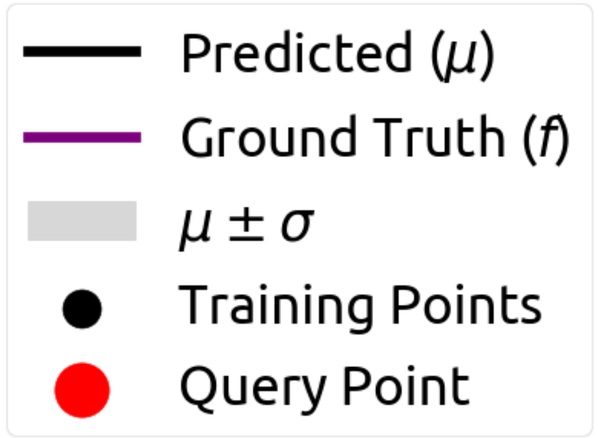

Iteration: 1
ε = 0.01
Gold content
8
6
4
0
1
2
3
4
5
6
0.3
0.2
0.1
0.0
α_{EI}
0
1
2
3
4
5
6

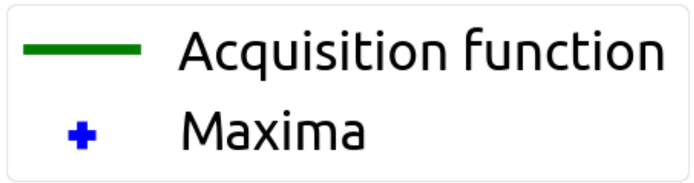

Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point

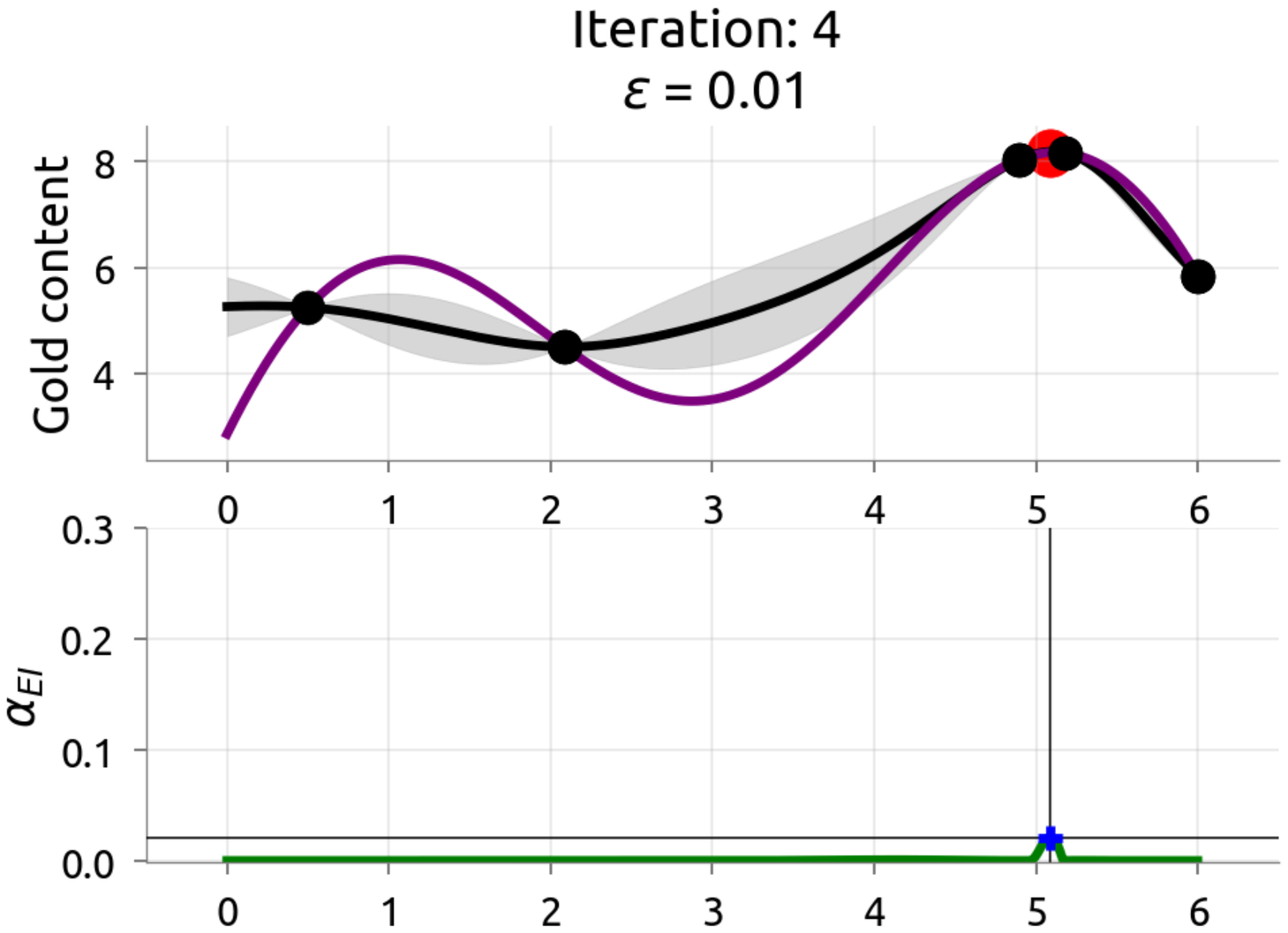

Acquisition function
Maxima

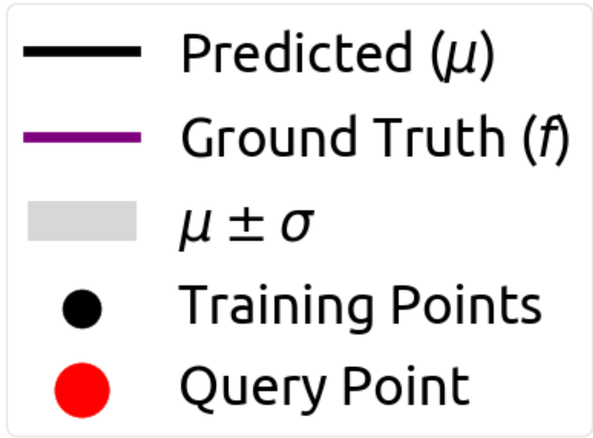

Iteration: 2
ε = 0.01
Gold content
8
6
4
0
1
2
3
4
5
6
α_{EI}
0.3
0.2
0.1
0.0
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

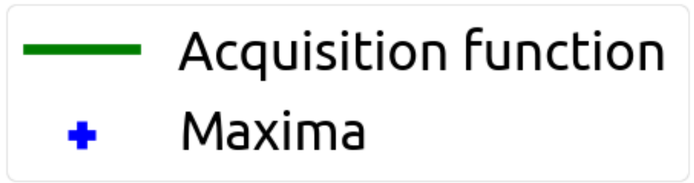

Iteration: 3
ε = 0.01
Gold content
8
6
4
0
1
2
3
4
5
6
0.3
0.2
0.1
0.0
α_{EI}
0
1
2
3
4
5
6

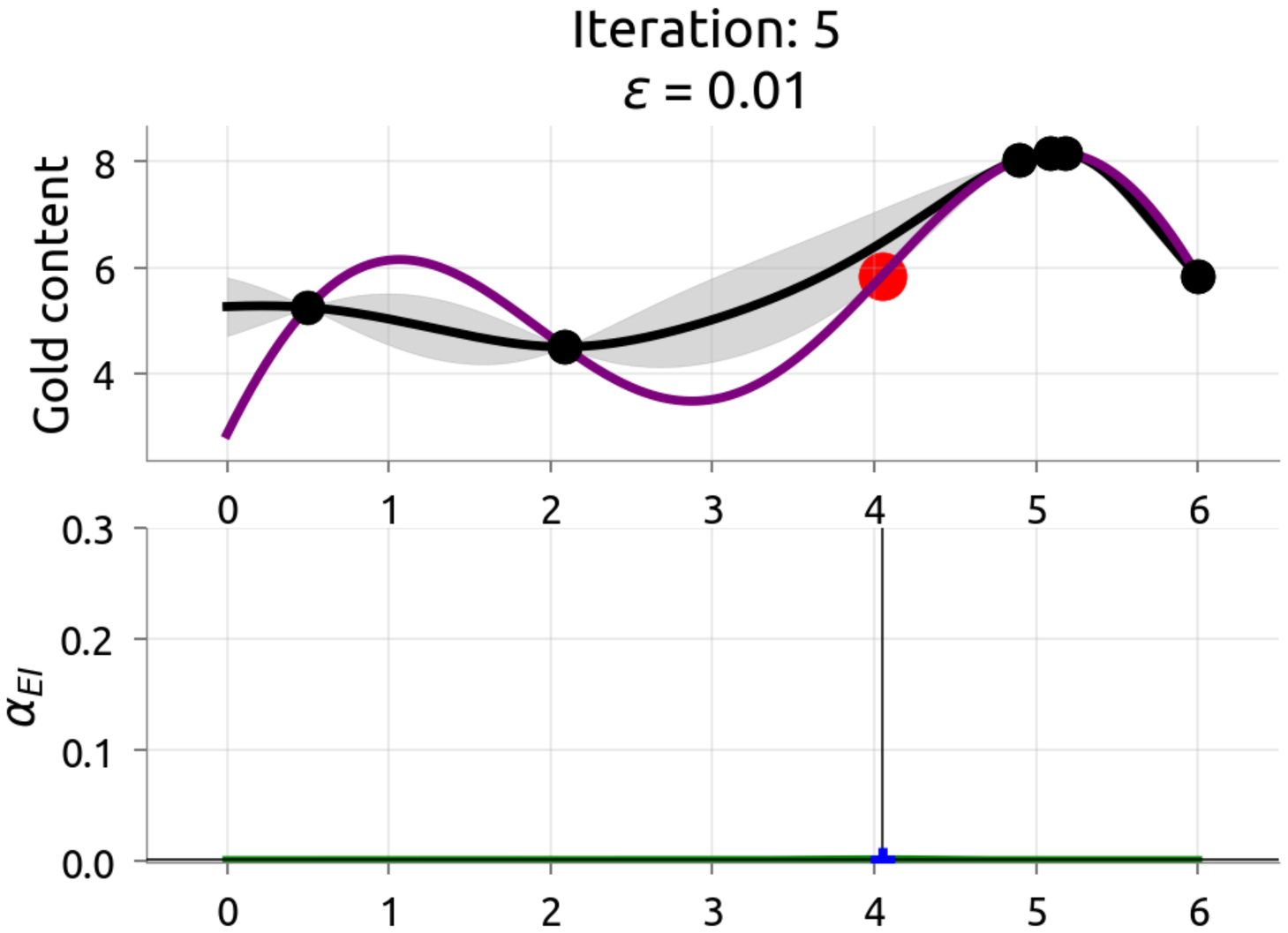

Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point

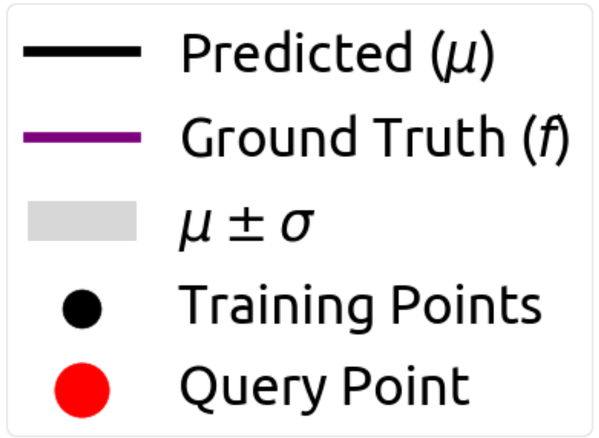

Acquisition function
Maxima

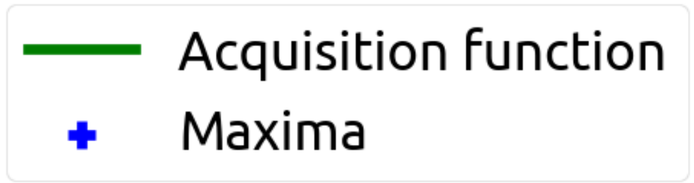

Iteration: 4
ε = 0.01
Gold content
8
6
4
0
1
2
3
4
5
6
0.3
0.2
0.1
0.0
α_{EI}
0
1
2
3
4
5
6

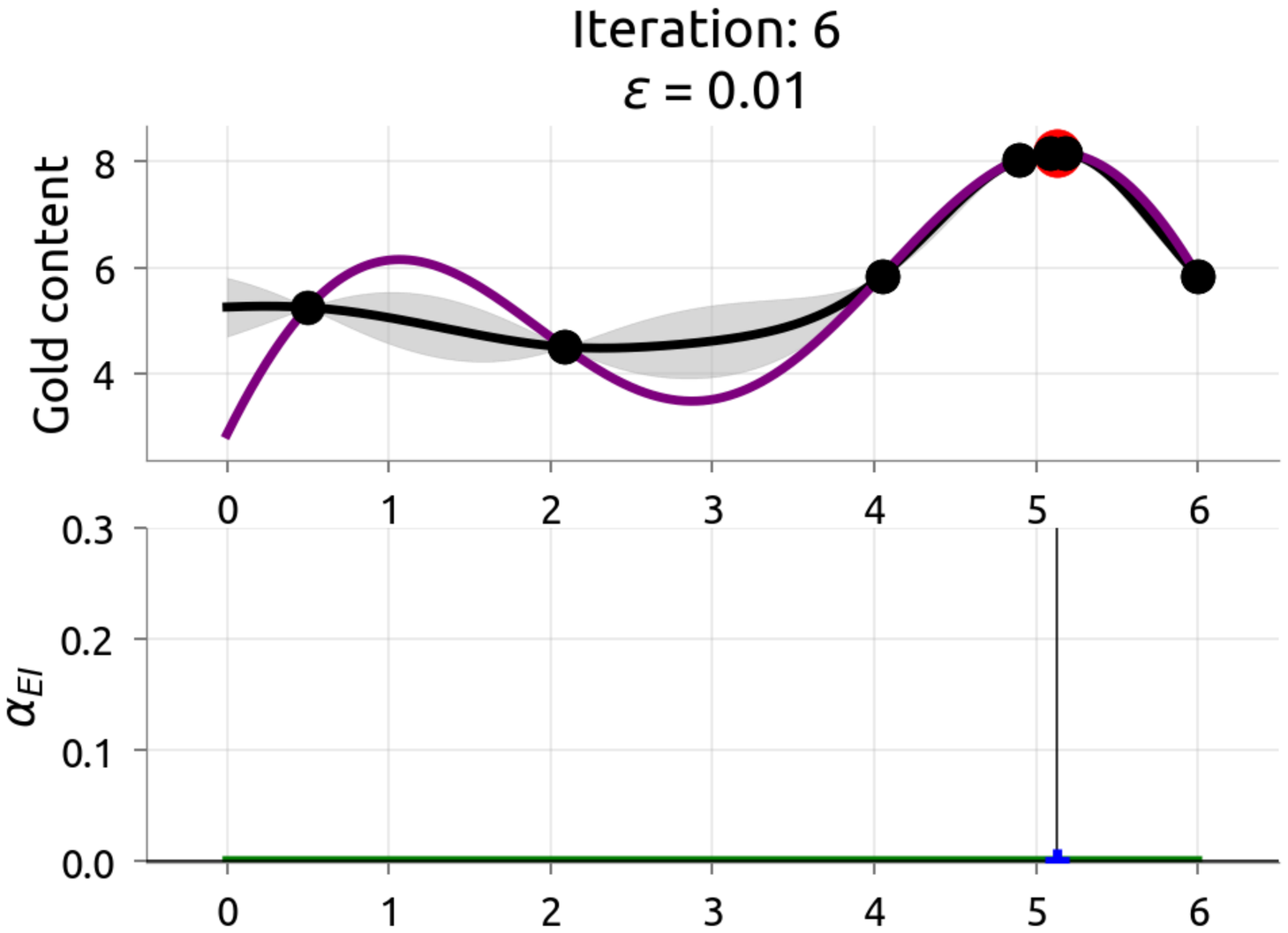

Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point

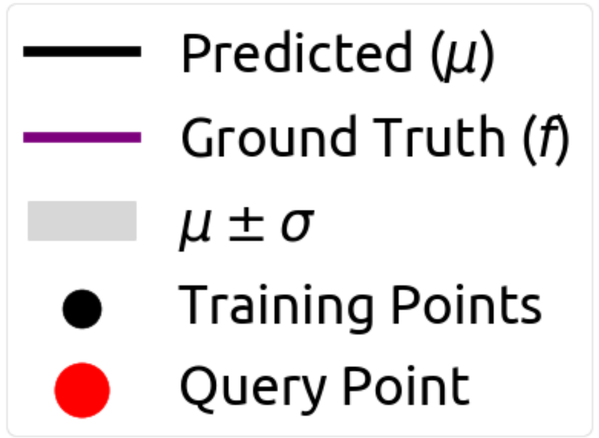

Acquisition function
Maxima

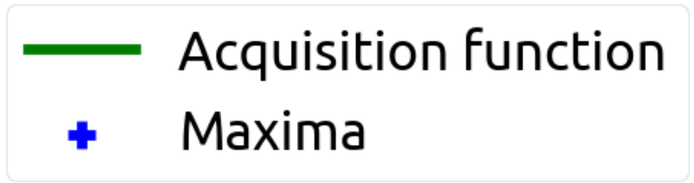

Iteration: 5
ε = 0.01
Gold content
8
6
4
0
1
2
3
4
5
6
α_{EI}
0.3
0.2
0.1
0.0
0
1
2
3
4
5
6

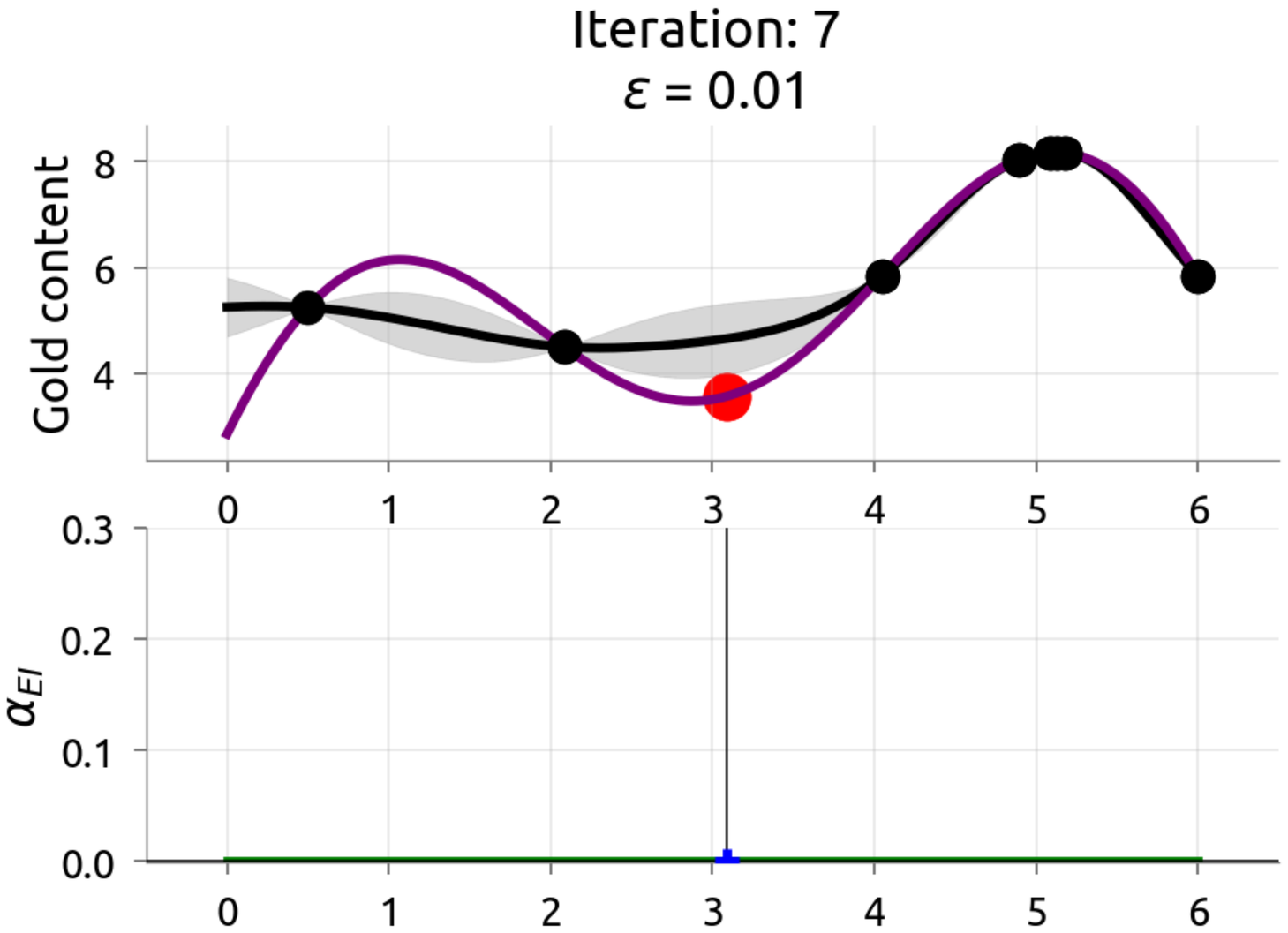

Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point

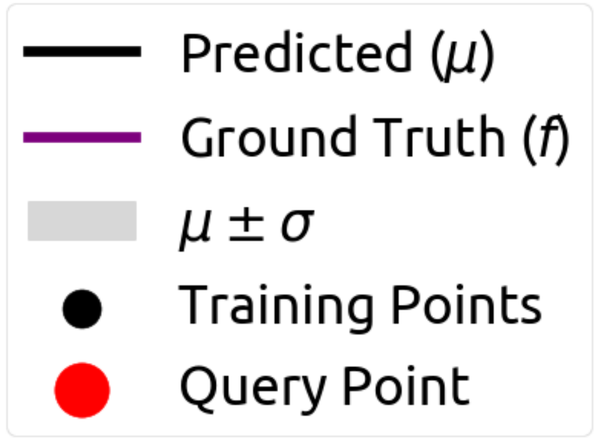

Acquisition function
Maxima

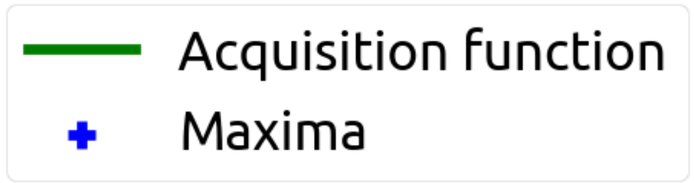

Iteration: 6
ε = 0.01
Gold content
8
6
4
0
1
2
3
4
5
6
0.3
0.2
0.1
0.0
α_{EI}
0
1
2
3
4
5
6

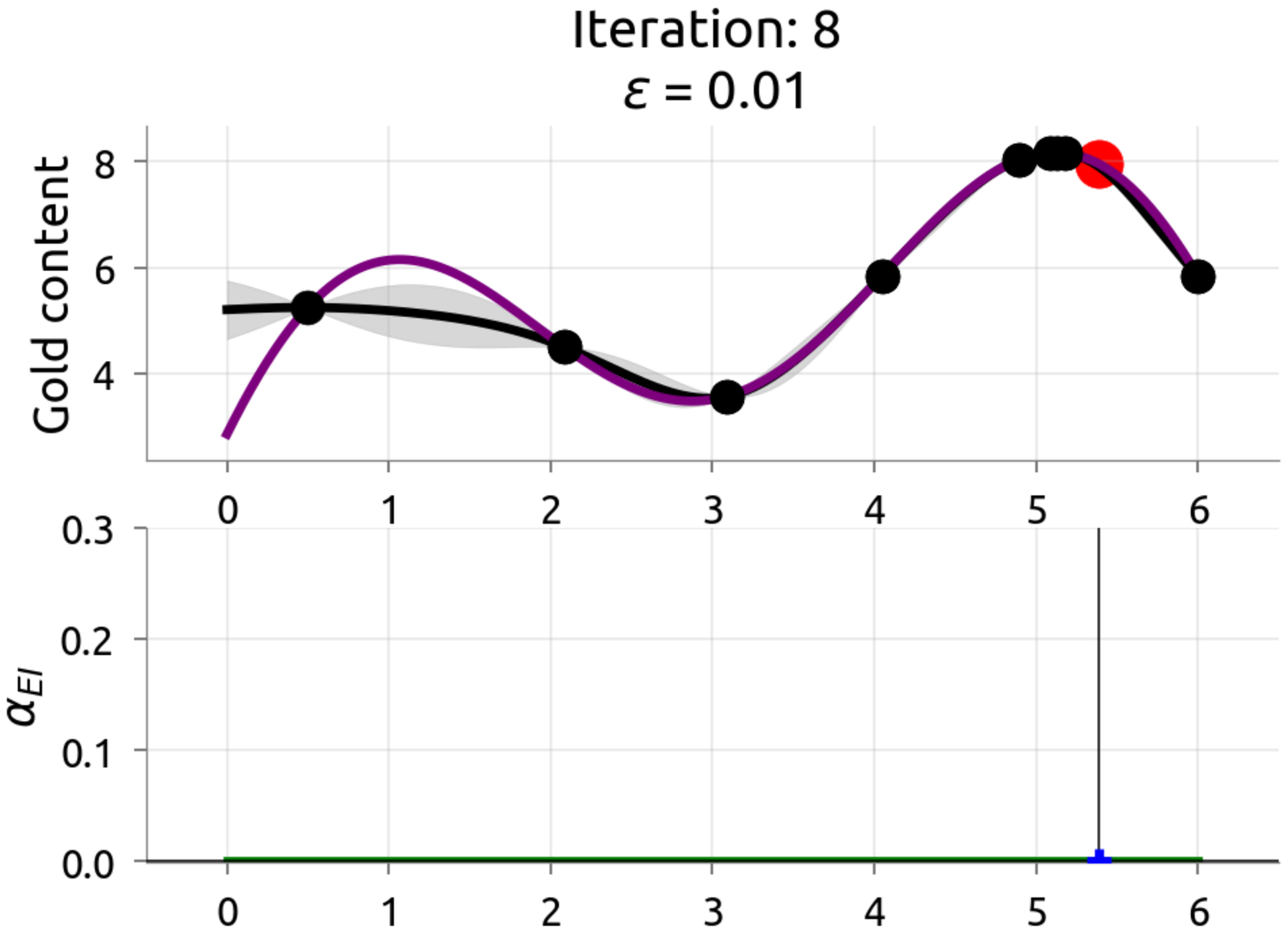

Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point

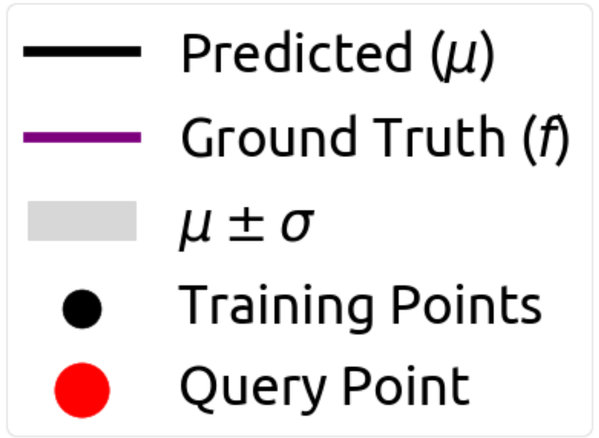

Acquisition function
Maxima

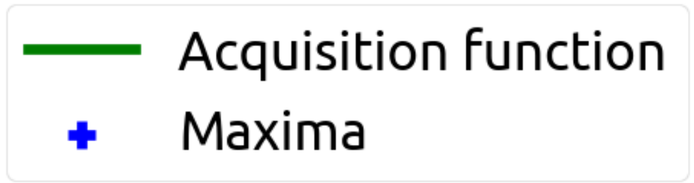

Iteration: 7
ε = 0.01
Gold content
8
6
4
0
1
2
3
4
5
6
α_{EI}
0.3
0.2
0.1
0.0
0
1
2
3
4
5
6

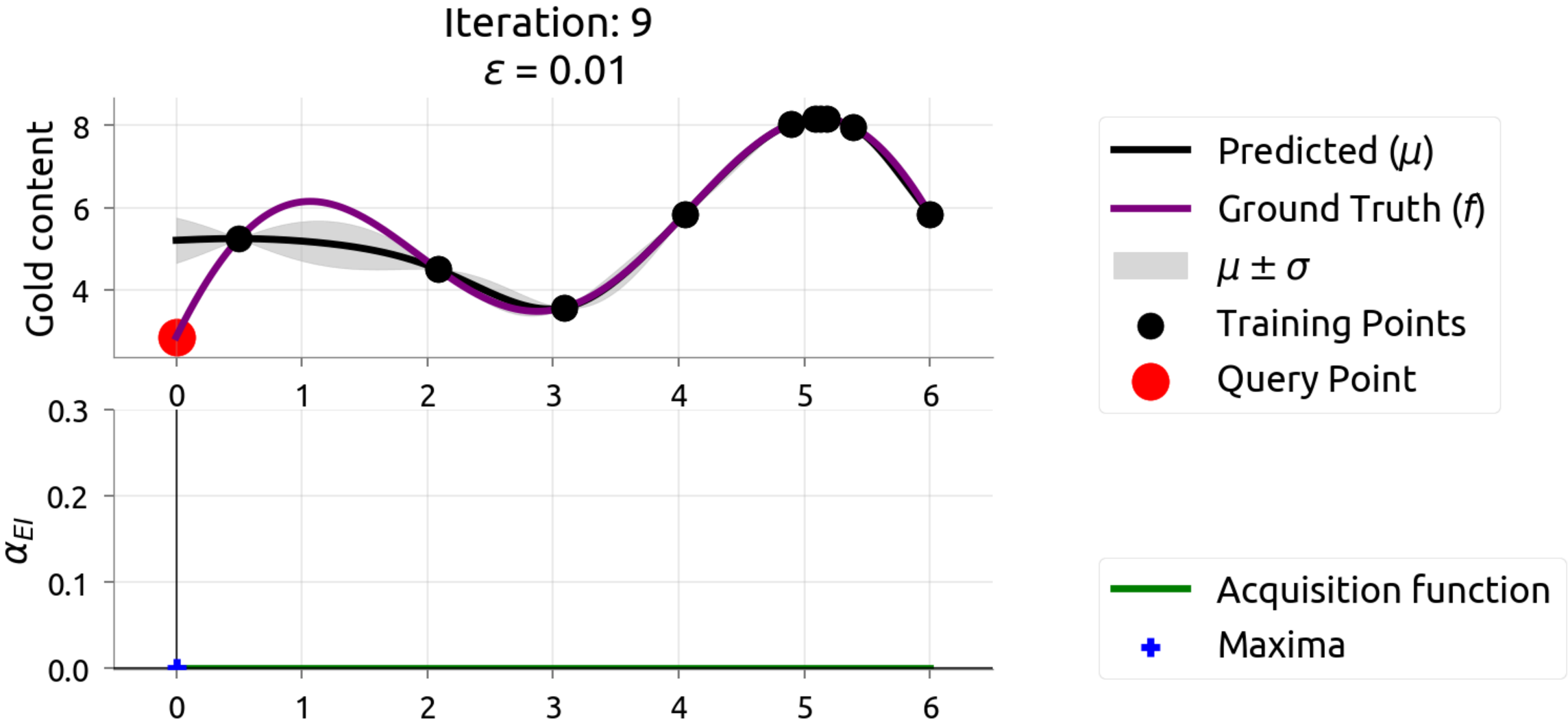

Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point


Acquisition function
Maxima

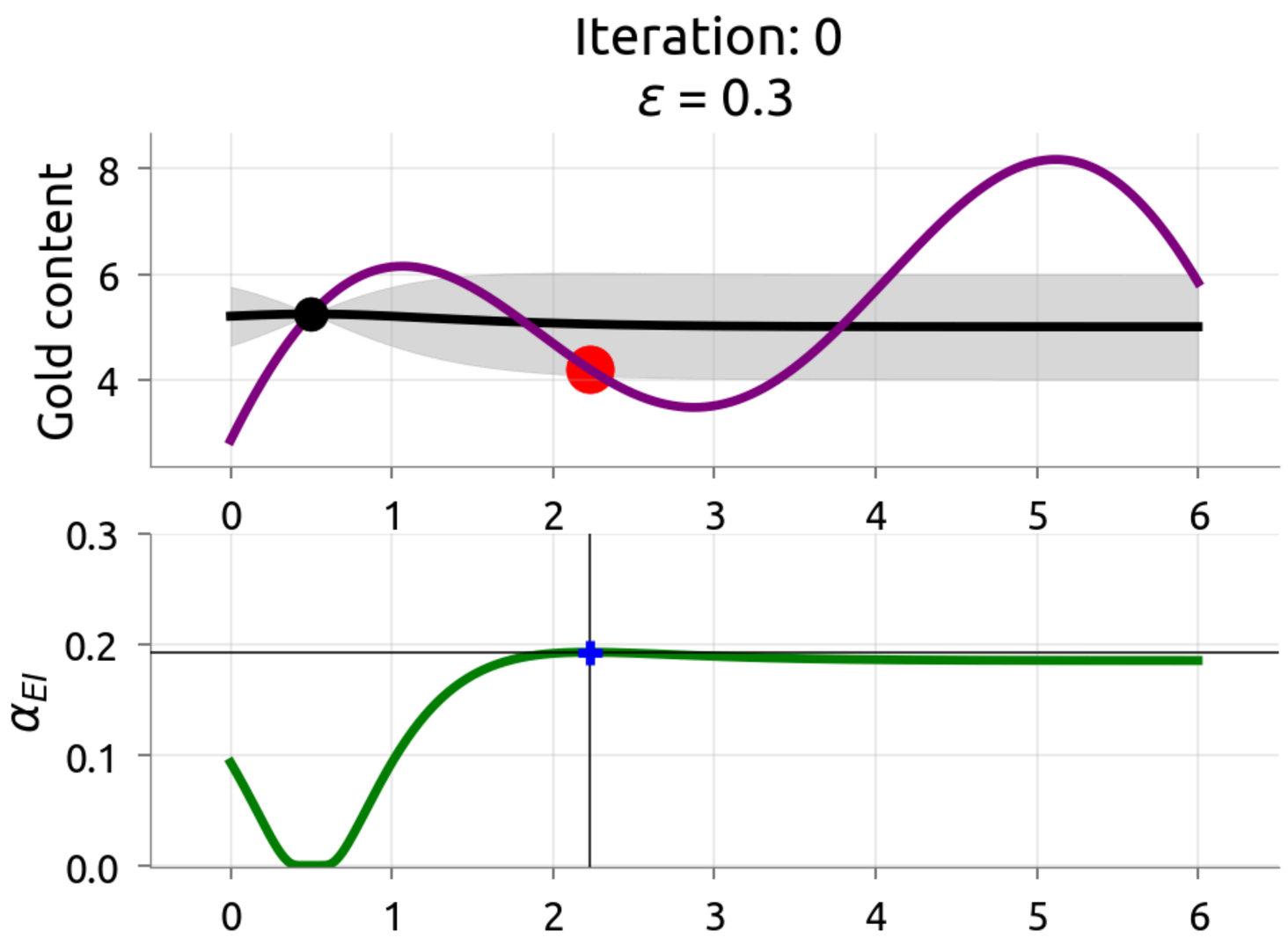

Iteration: 8
ε = 0.01
Gold content
8
6
4
0
1
2
3
4
5
6
0.3
0.2
0.1
0.0
α_{EI}
0
1
2
3
4
5
6

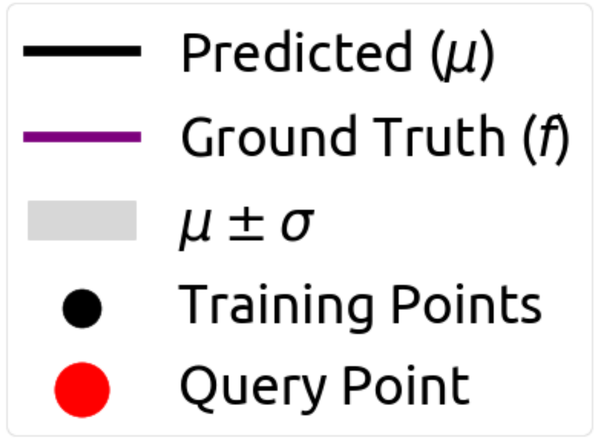

Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point

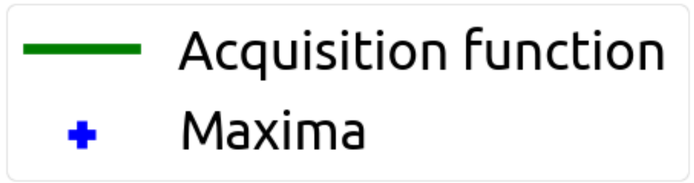

Acquisition function
Maxima

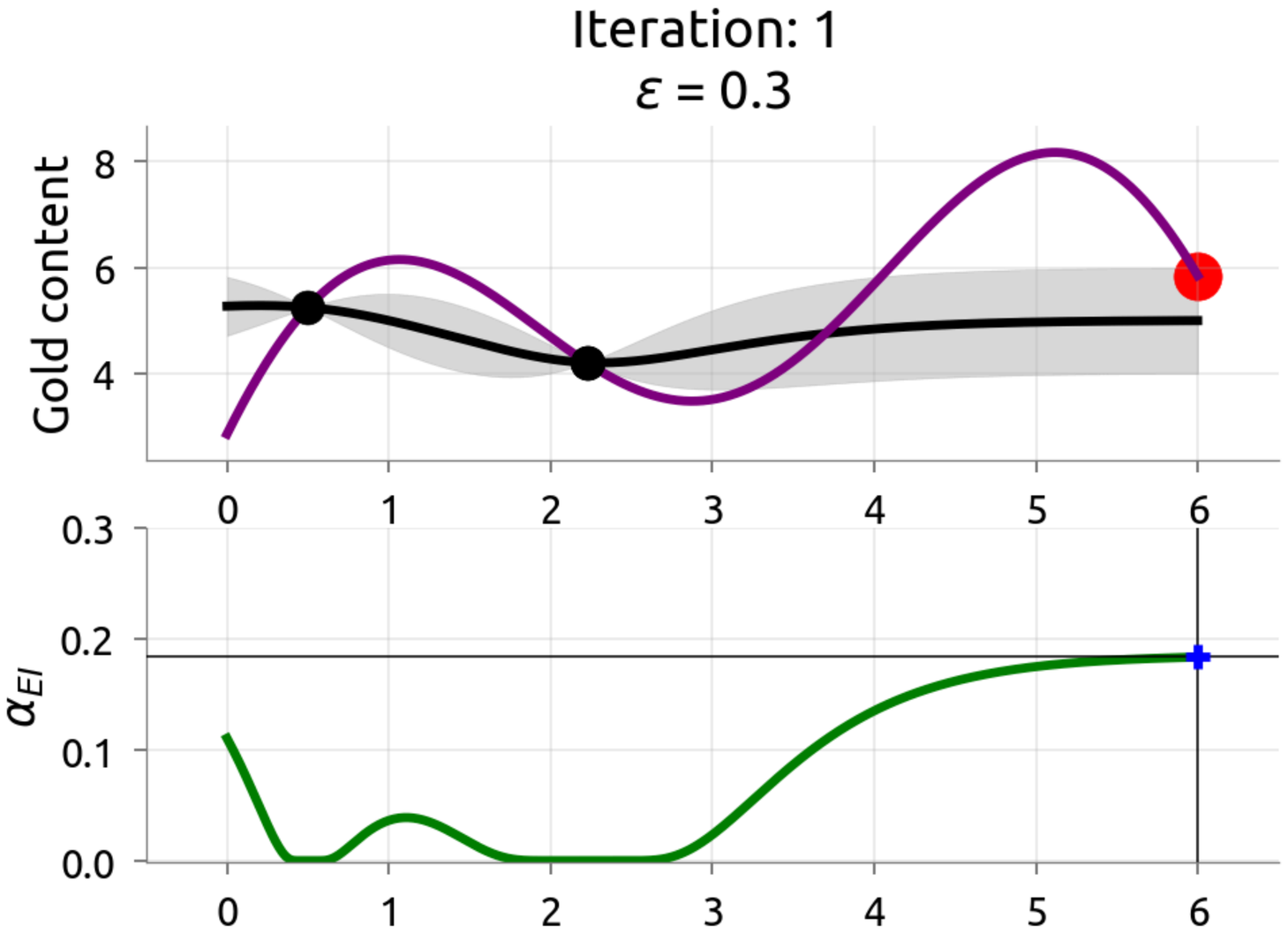

Iteration: 9
ε = 0.01
Gold content
8
6
4
0
1
2
3
4
5
6
α_{EI}
0.3
0.2
0.1
0.0
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

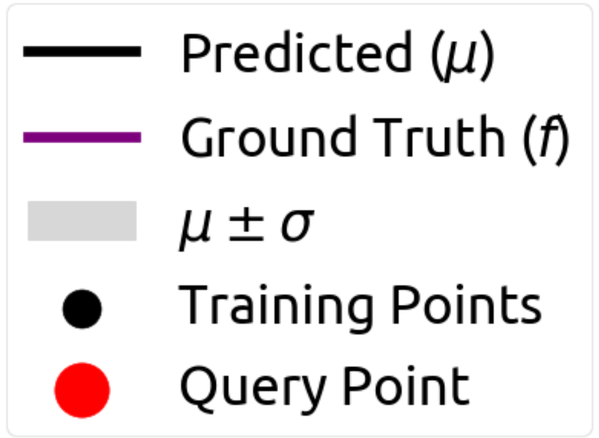
Intuition behind ϵ in PI: ϵ = 0.3

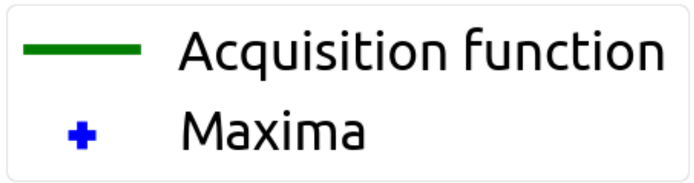


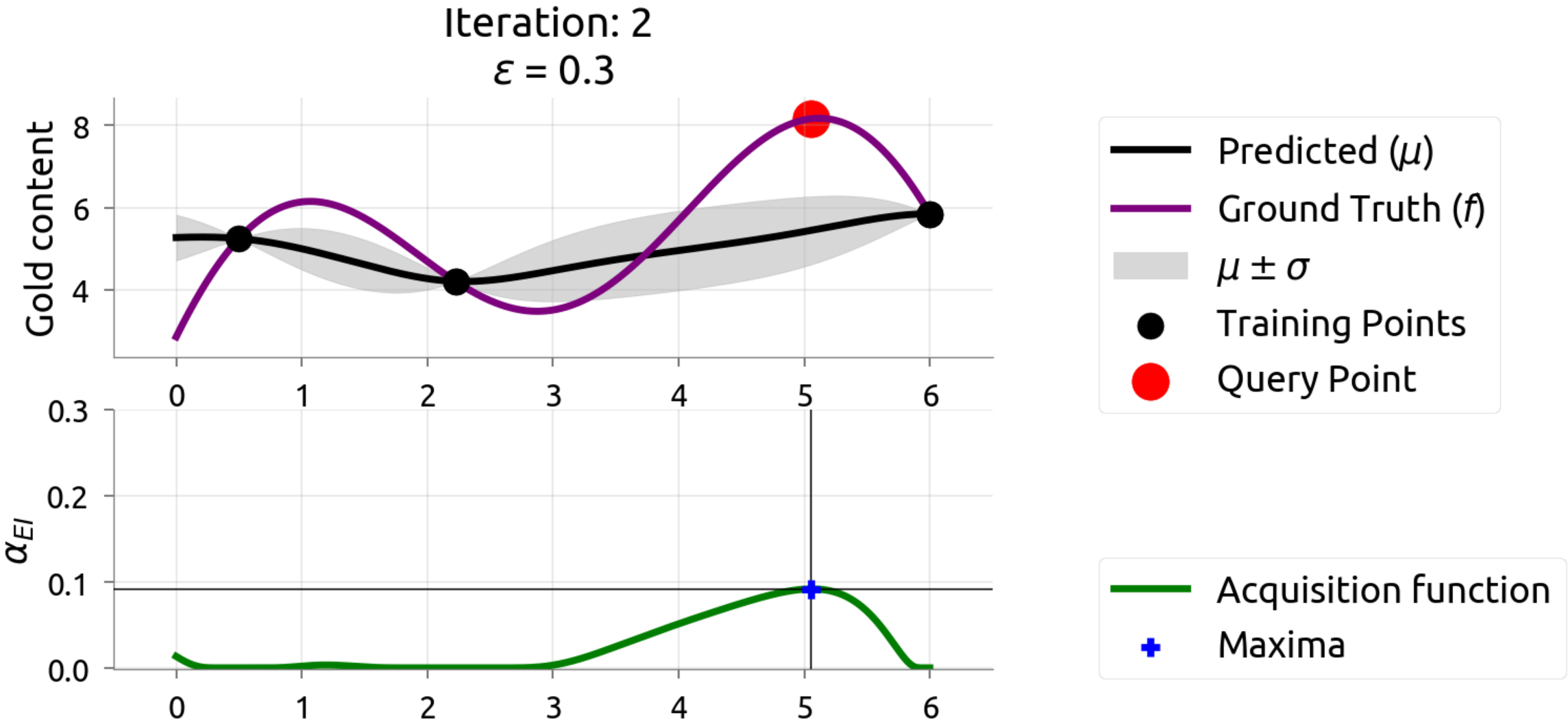


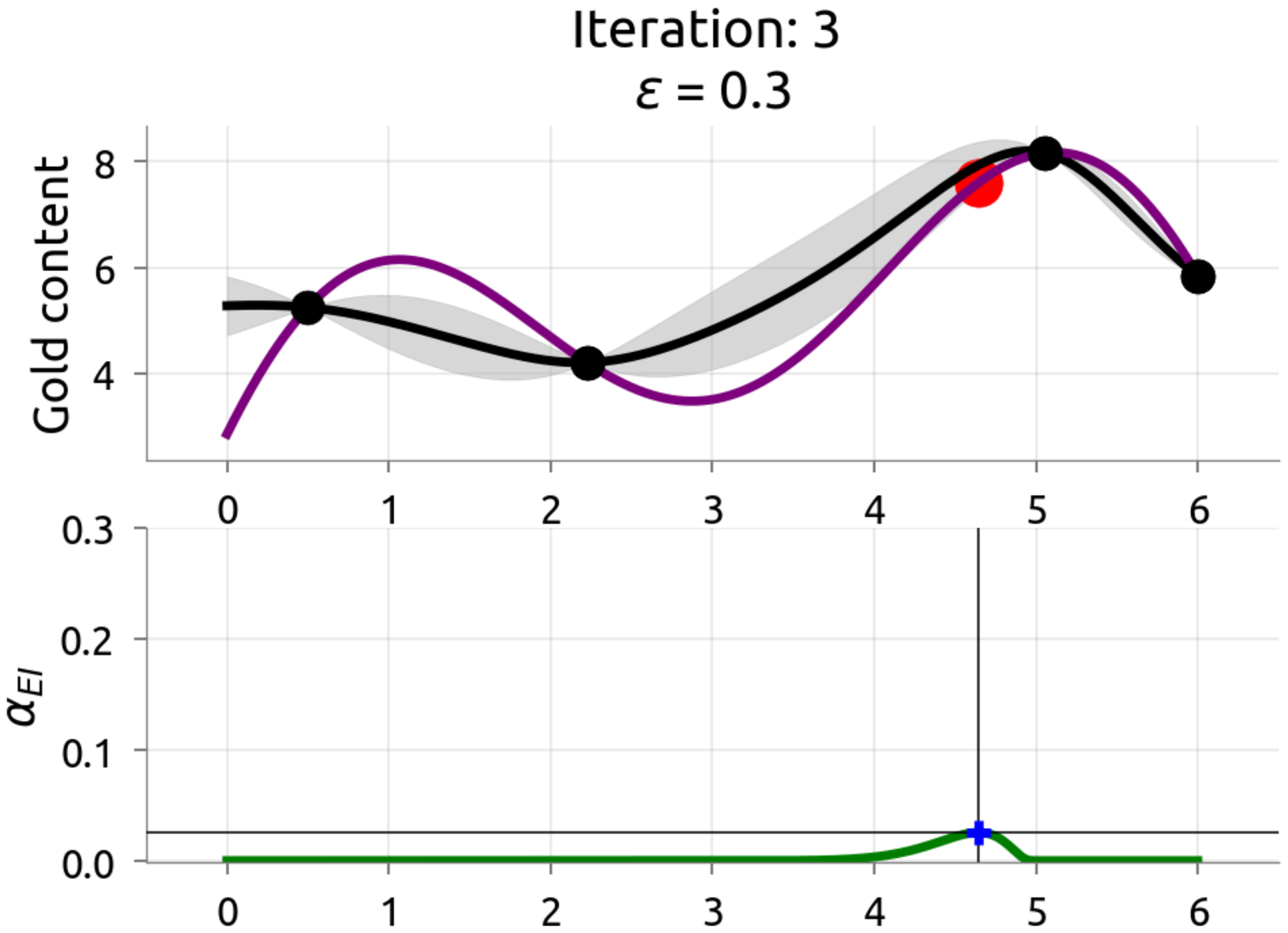


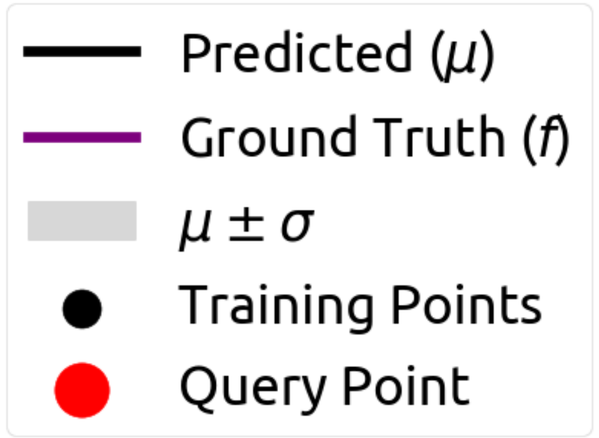

Iteration: 1
$\varepsilon = 0.3$
Gold content
8
6
4
0
1
2
3
4
5
6
0.3
0.2
0.1
0.0
α_{EI}
0
1
2
3
4
5
6

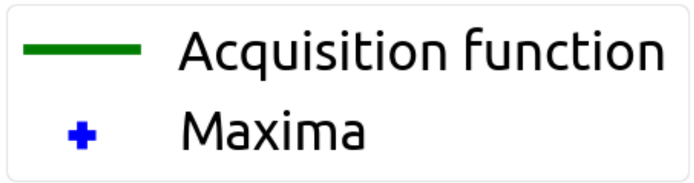

Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point

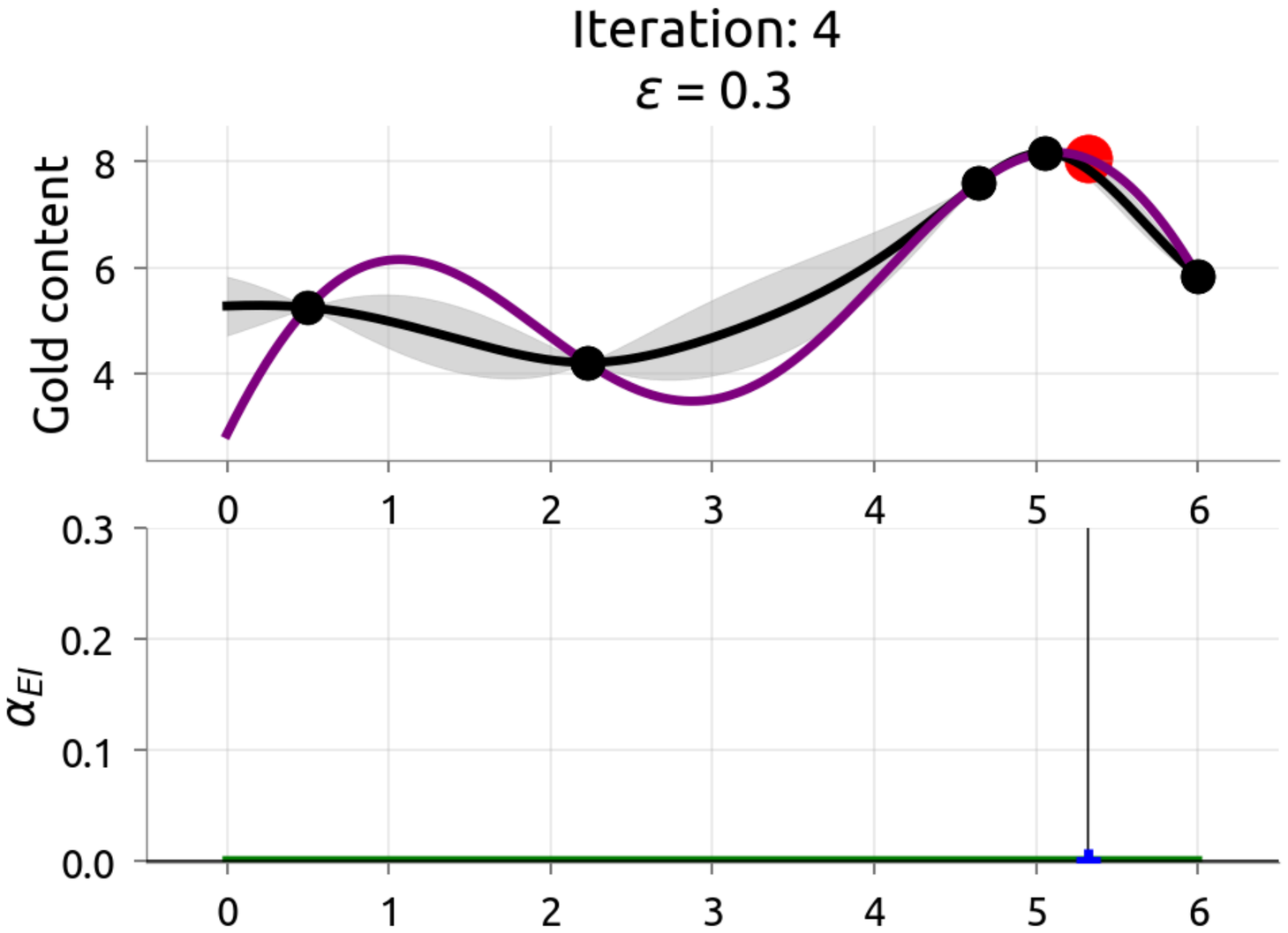

Acquisition function
Maxima

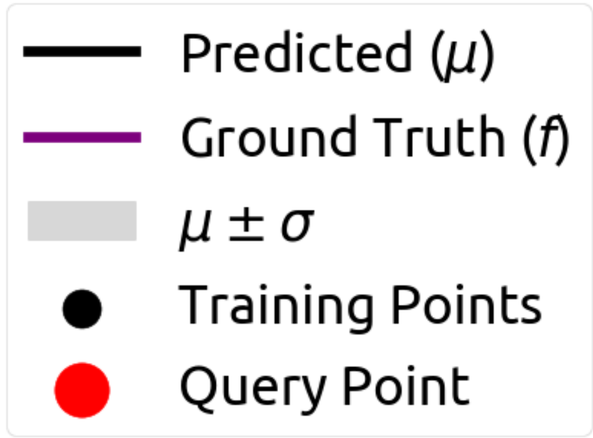

Iteration: 2
$\varepsilon = 0.3$
Gold content
8
6
4
0
1
2
3
4
5
6
α_{EI}
0.3
0.2
0.1
0.0
0
1
2
3
4
5
6
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

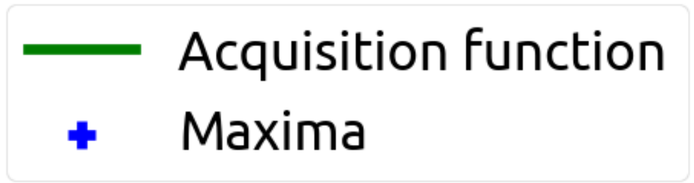

Iteration: 3
$\varepsilon = 0.3$
Gold content
8
6
4
0
1
2
3
4
5
6
0.3
0.2
0.1
0.0
α_{EI}
0
1
2
3
4
5
6

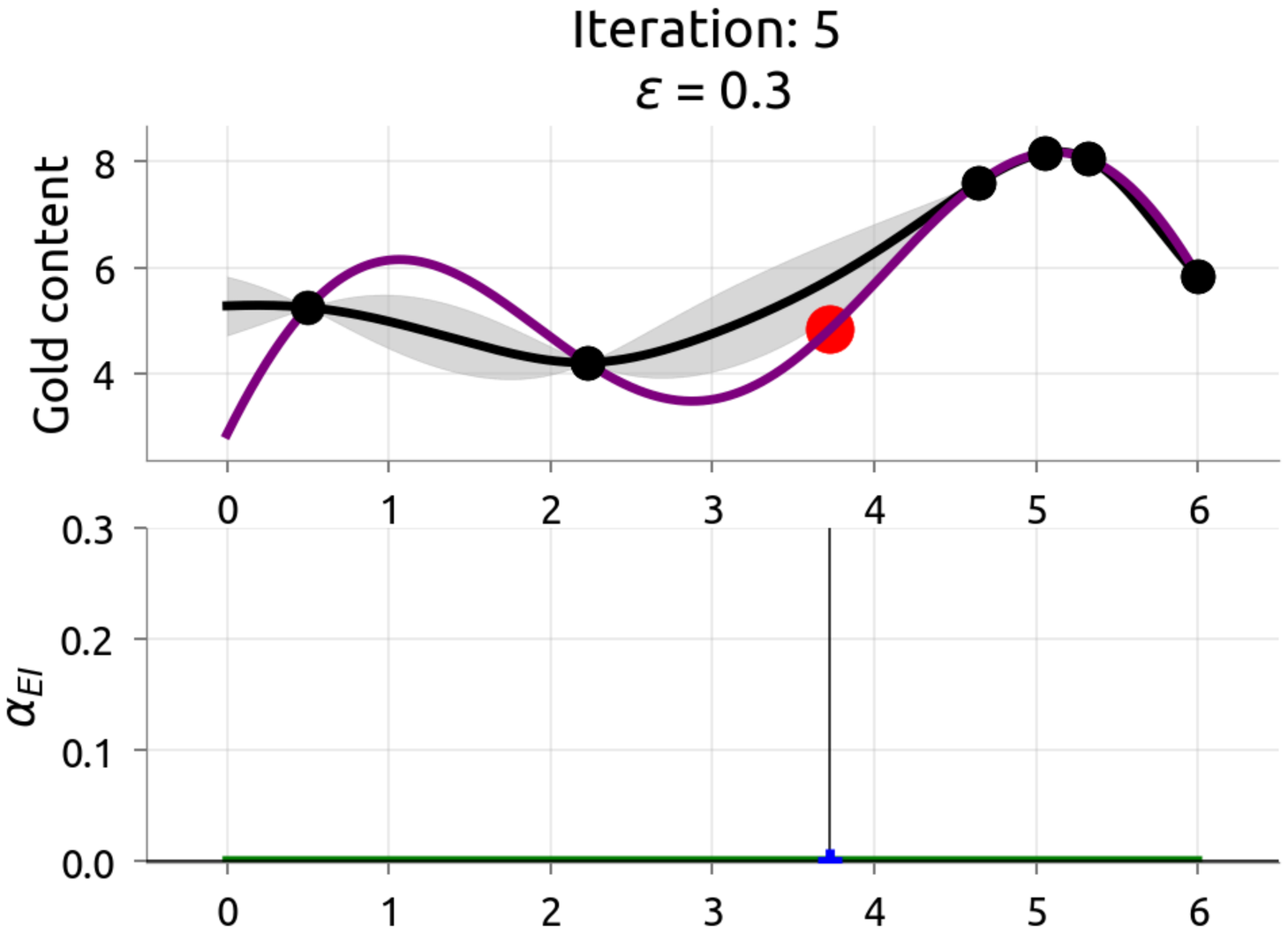

Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point

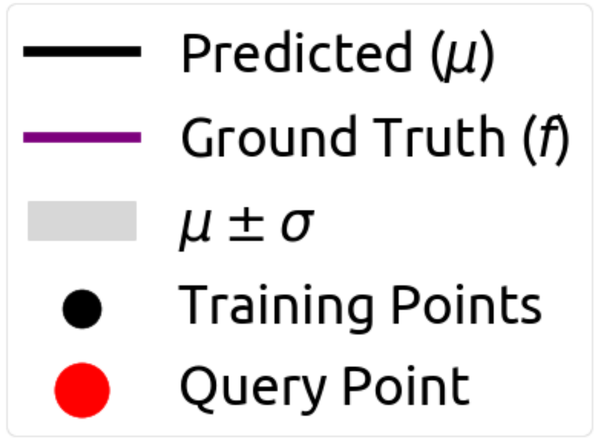

Acquisition function
Maxima

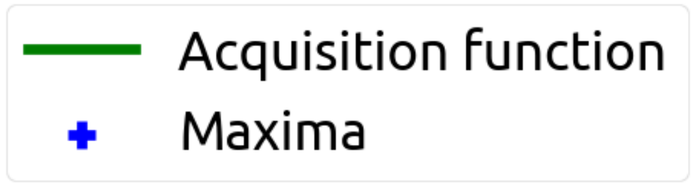

Iteration: 4
$\varepsilon = 0.3$
Gold content
8
6
4
0
1
2
3
4
5
6
α_{EI}
0.3
0.2
0.1
0.0
0
1
2
3
4
5
6

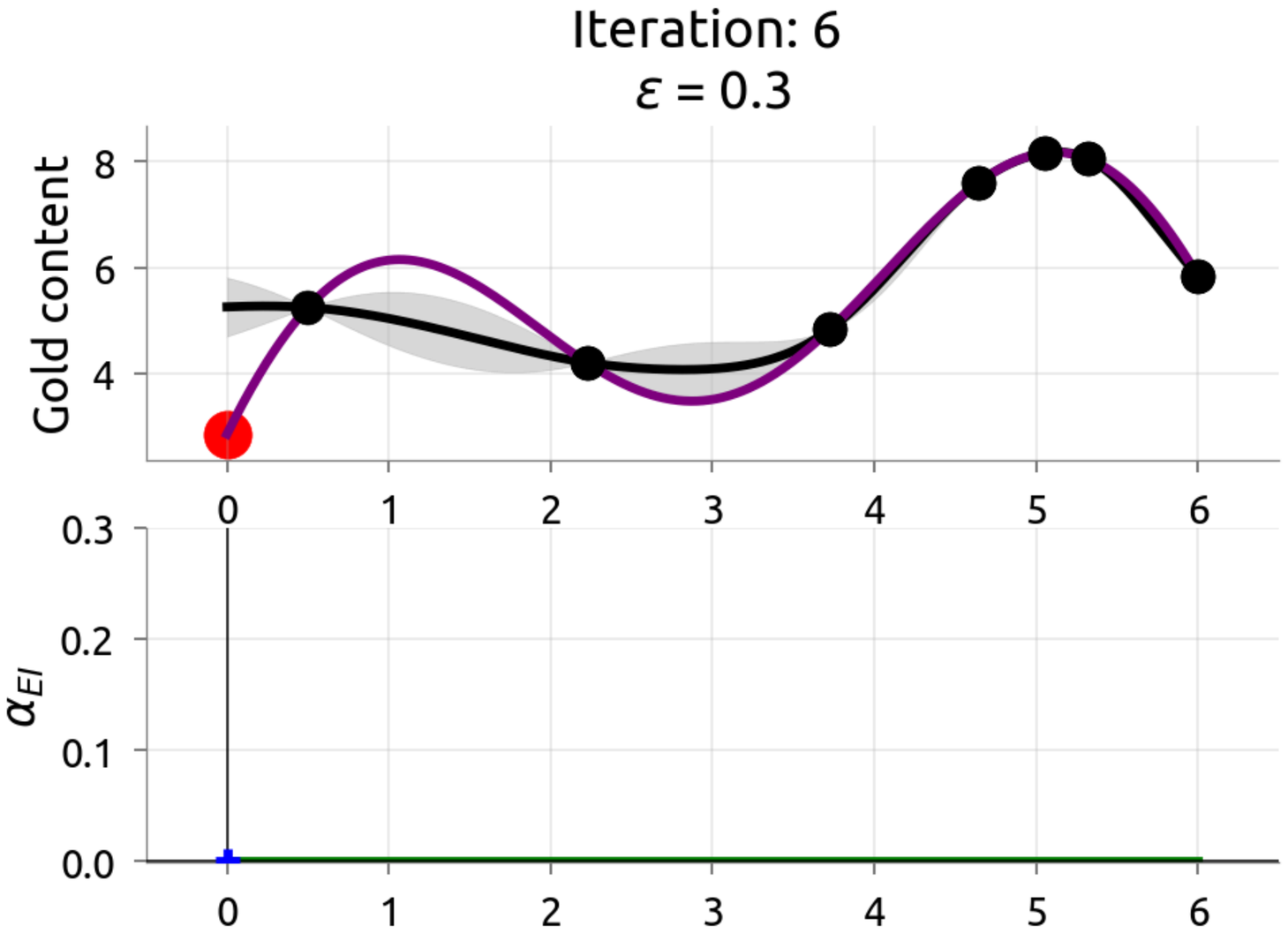

Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point

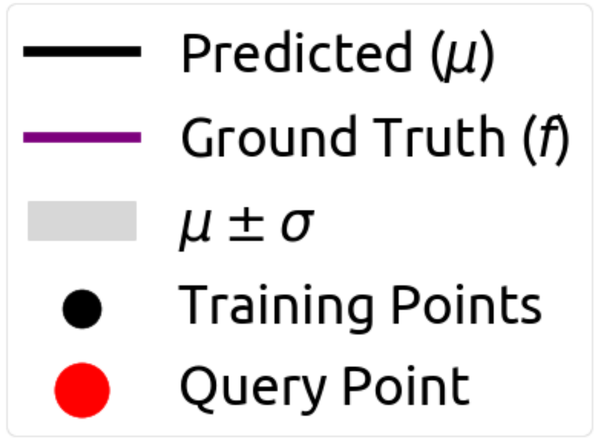

Acquisition function
Maxima

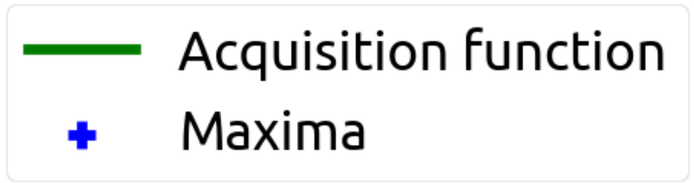

Iteration: 5
$\varepsilon = 0.3$
Gold content
8
6
4
0
1
2
3
4
5
6
0.3
0.2
0.1
0.0
α_{EI}
0
1
2
3
4
5
6

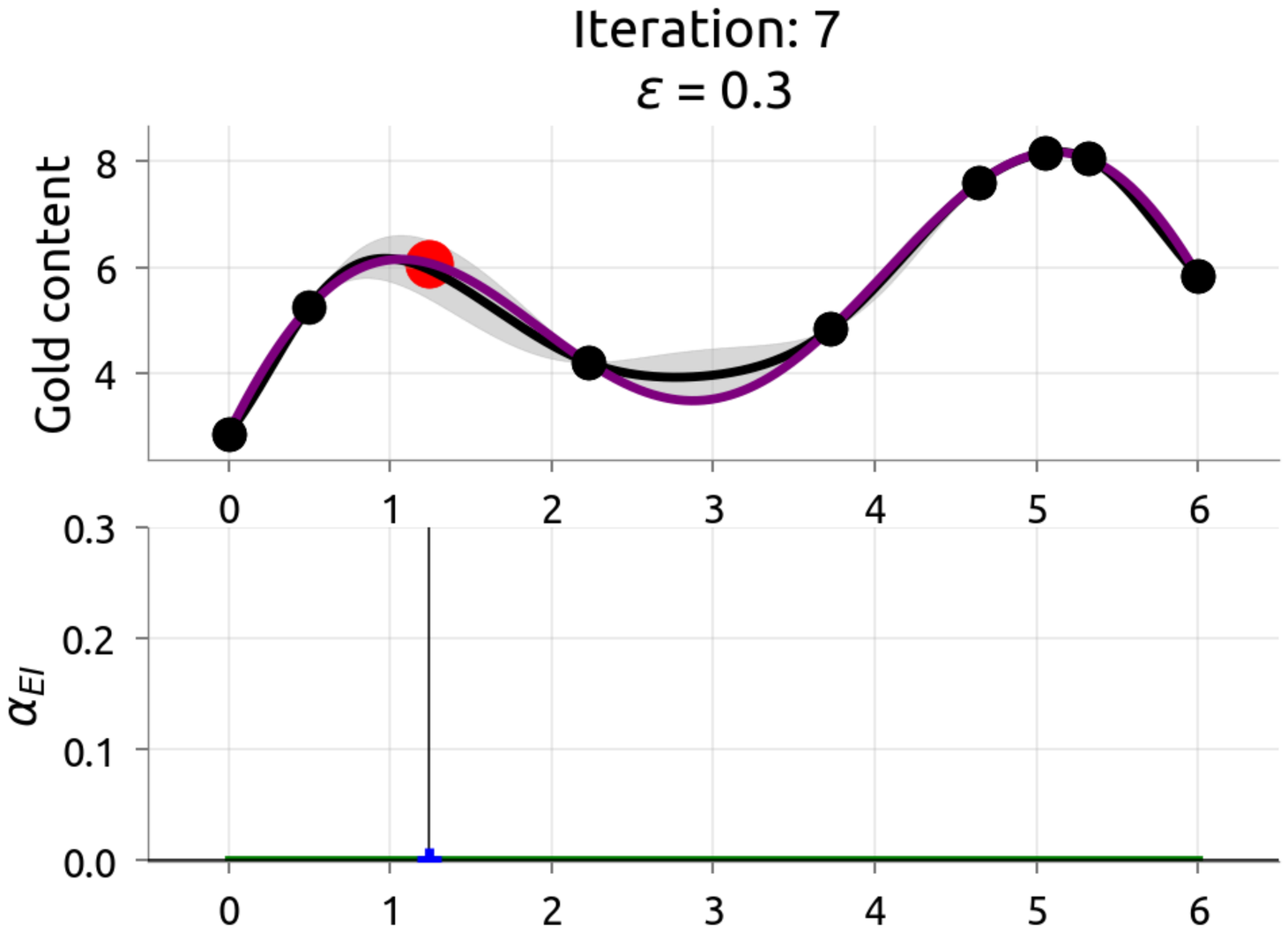

Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point

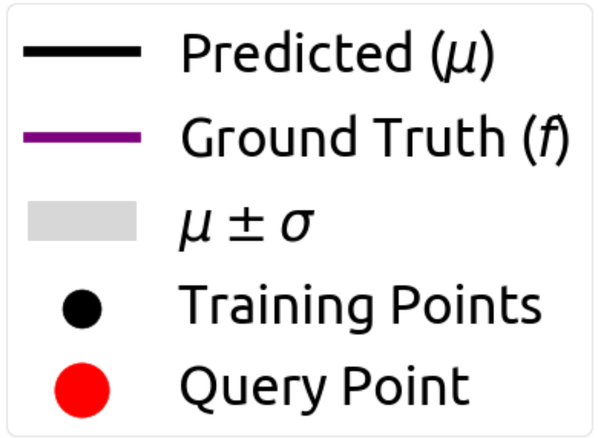

Acquisition function
Maxima

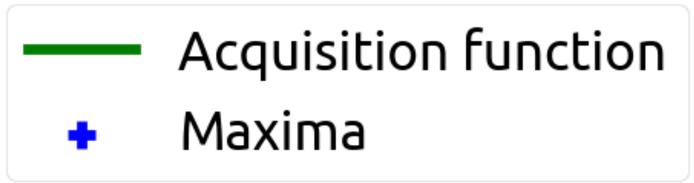

Iteration: 6
ε = 0.3
Gold content
8
6
4
0
1
2
3
4
5
6
0.3
0.2
0.1
0.0
α_{EI}
0
1
2
3
4
5
6

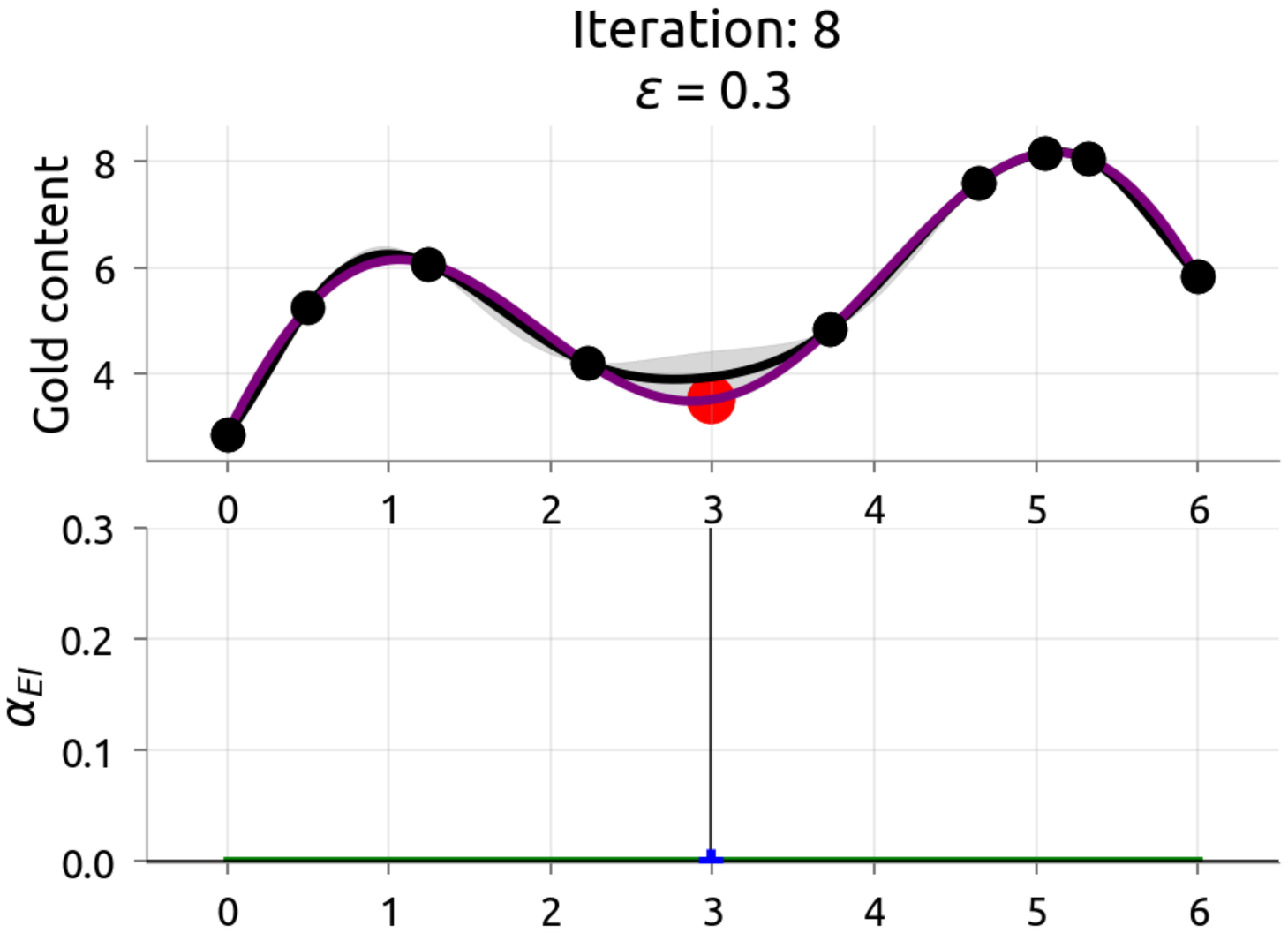

Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point

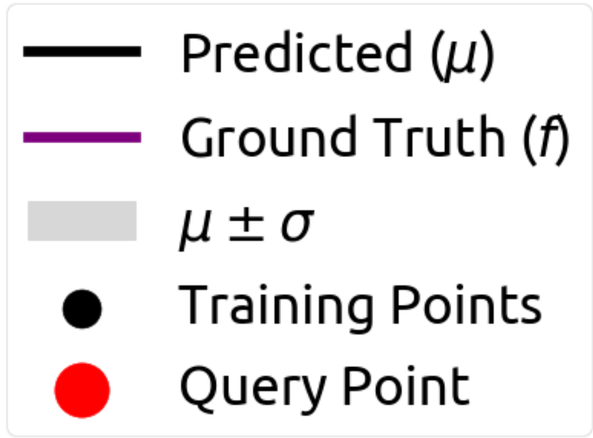

Acquisition function
Maxima

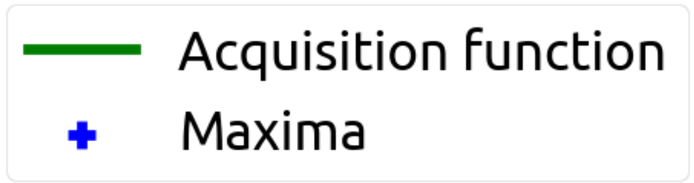

Iteration: 7
$\varepsilon = 0.3$
Gold content
8
6
4
0
1
2
3
4
5
6
0.3
0.2
0.1
0.0
α_{EI}
0
1
2
3
4
5
6

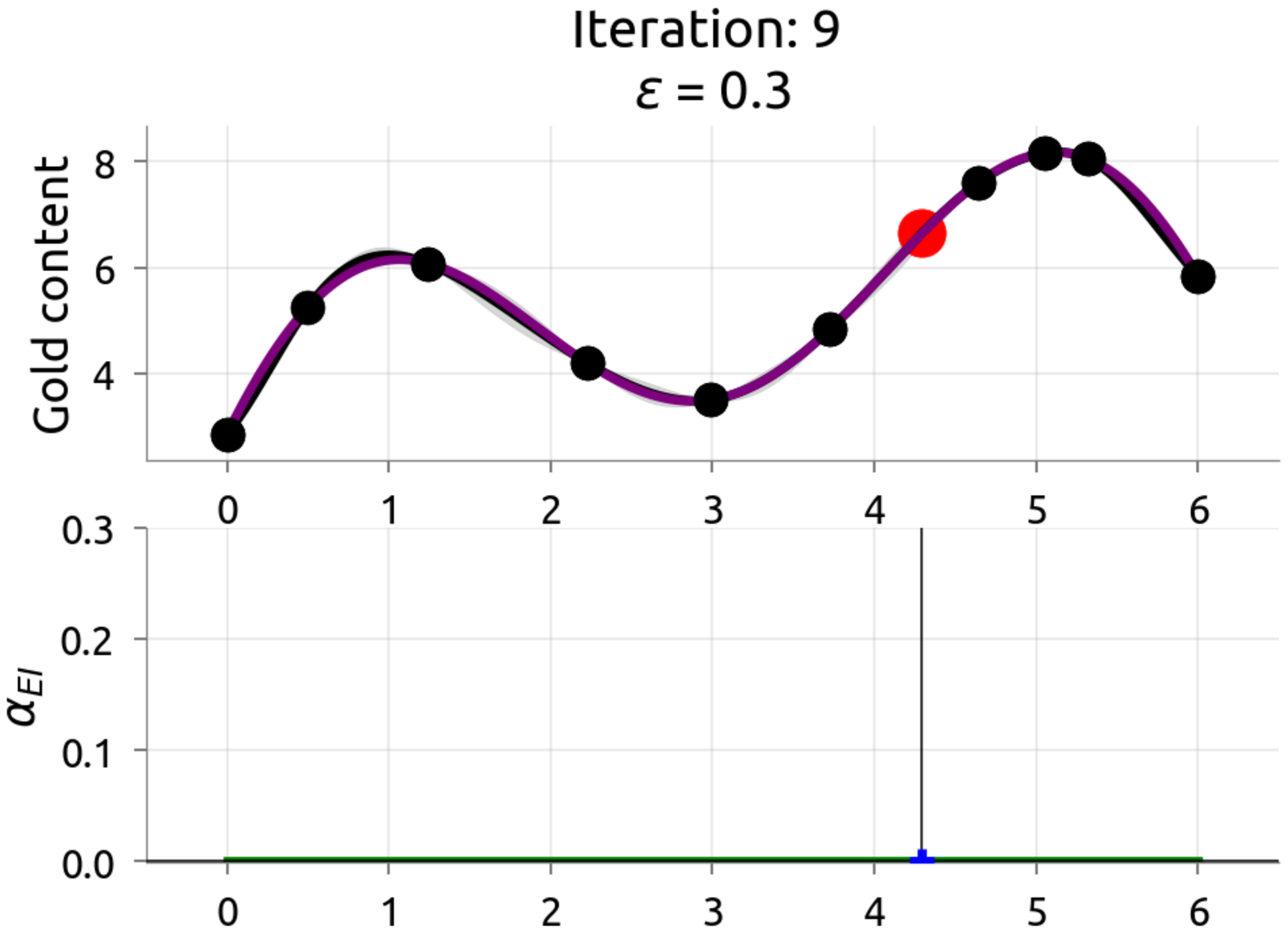

Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point

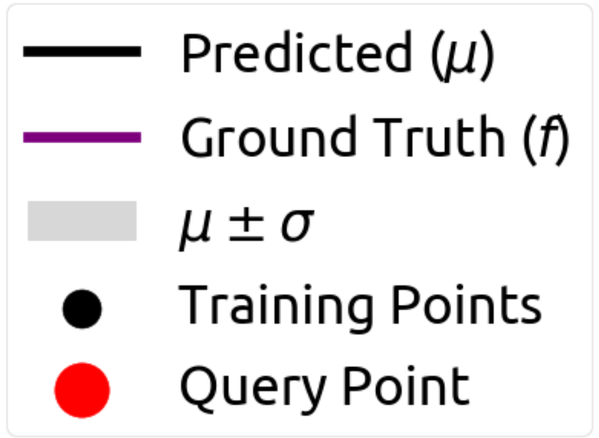

Acquisition function
Maxima

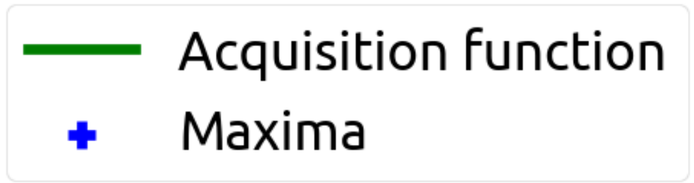

Iteration: 8
$\varepsilon = 0.3$
Gold content
8
6
4
0
1
2
3
4
5
6
α_{EI}
0.3
0.2
0.1
0.0
0
1
2
3
4
5
6


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

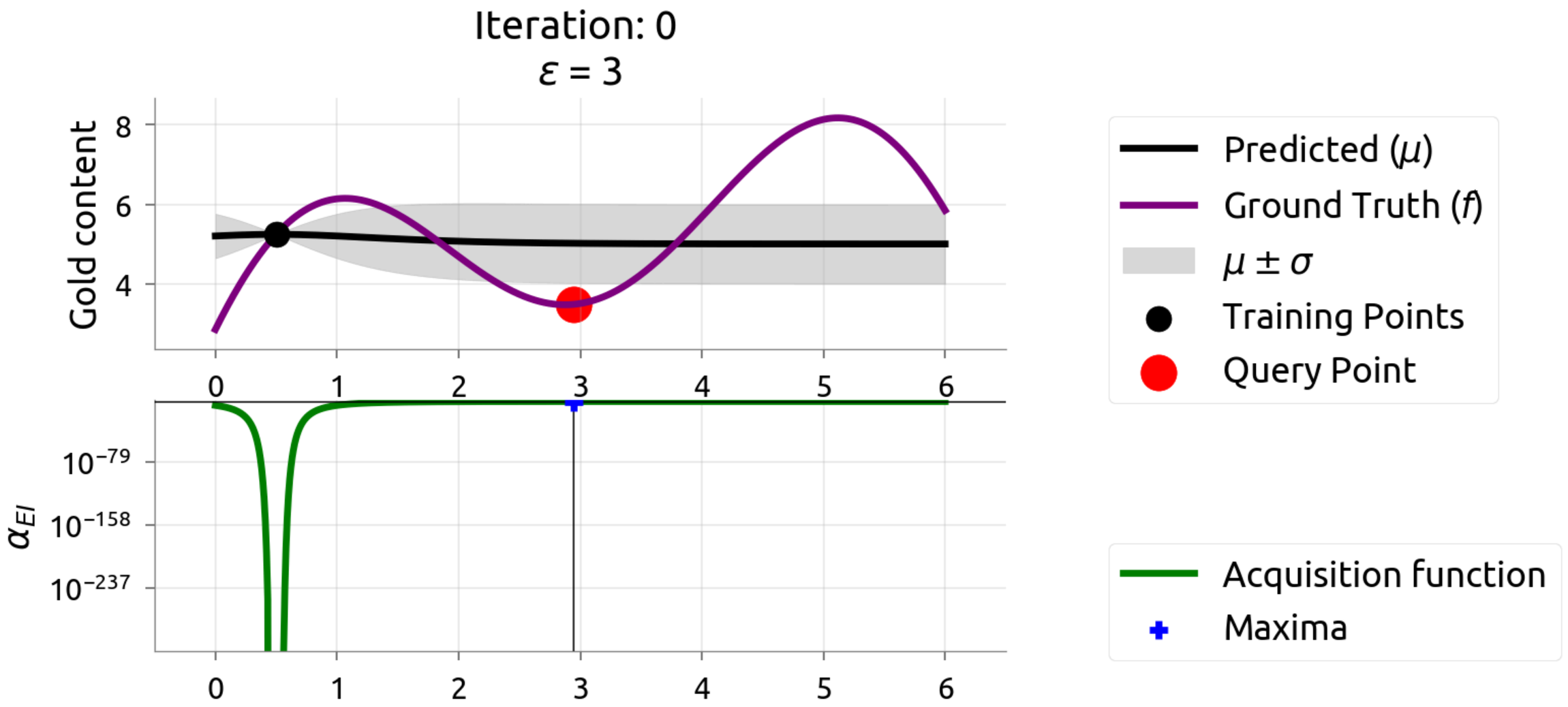

Acquisition function
Maxima

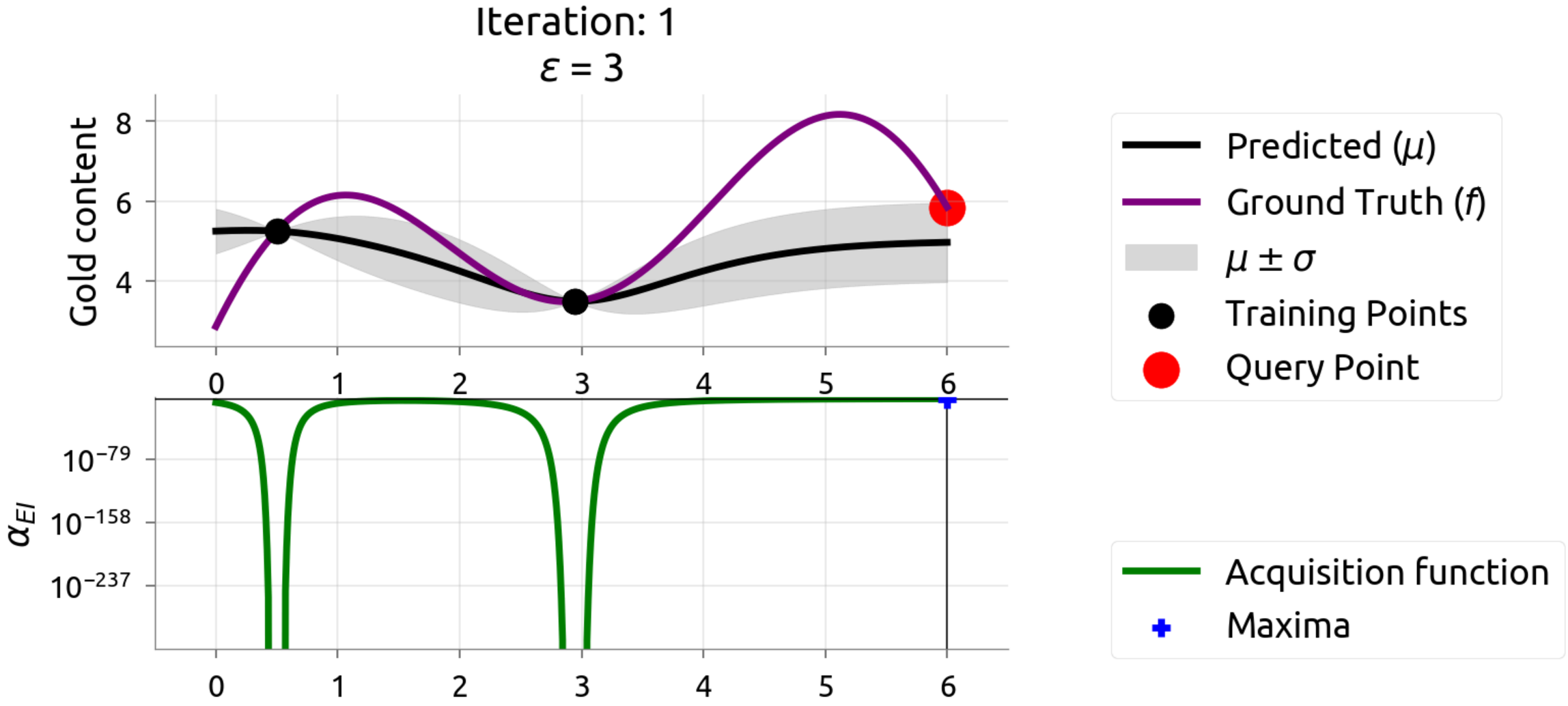

Iteration: 9
$\varepsilon = 0.3$
Gold content
8
6
4
0
1
2
3
4
5
6
0.3
0.2
0.1
0.0
α_{EI}
0
1
2
3
4
5
6

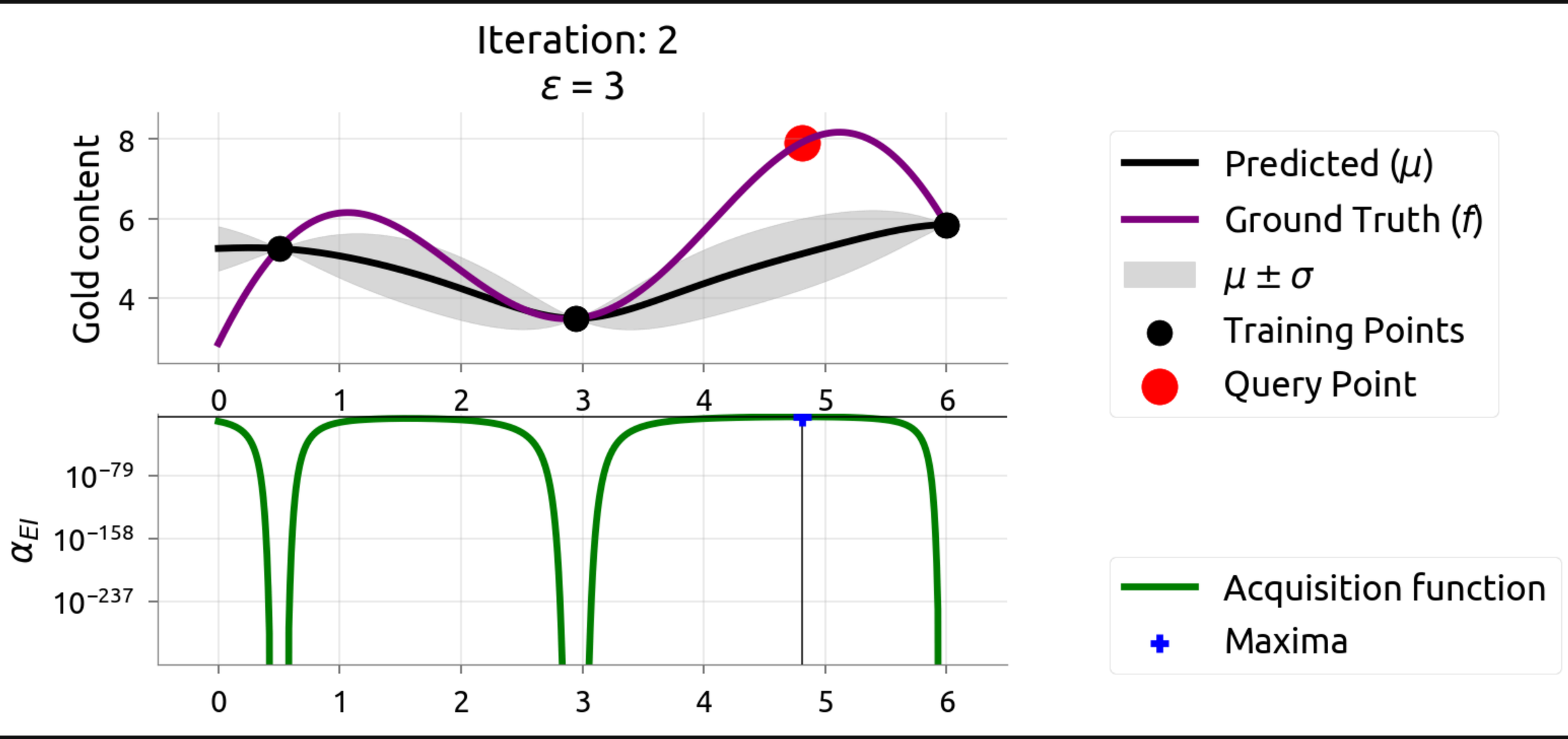

Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point

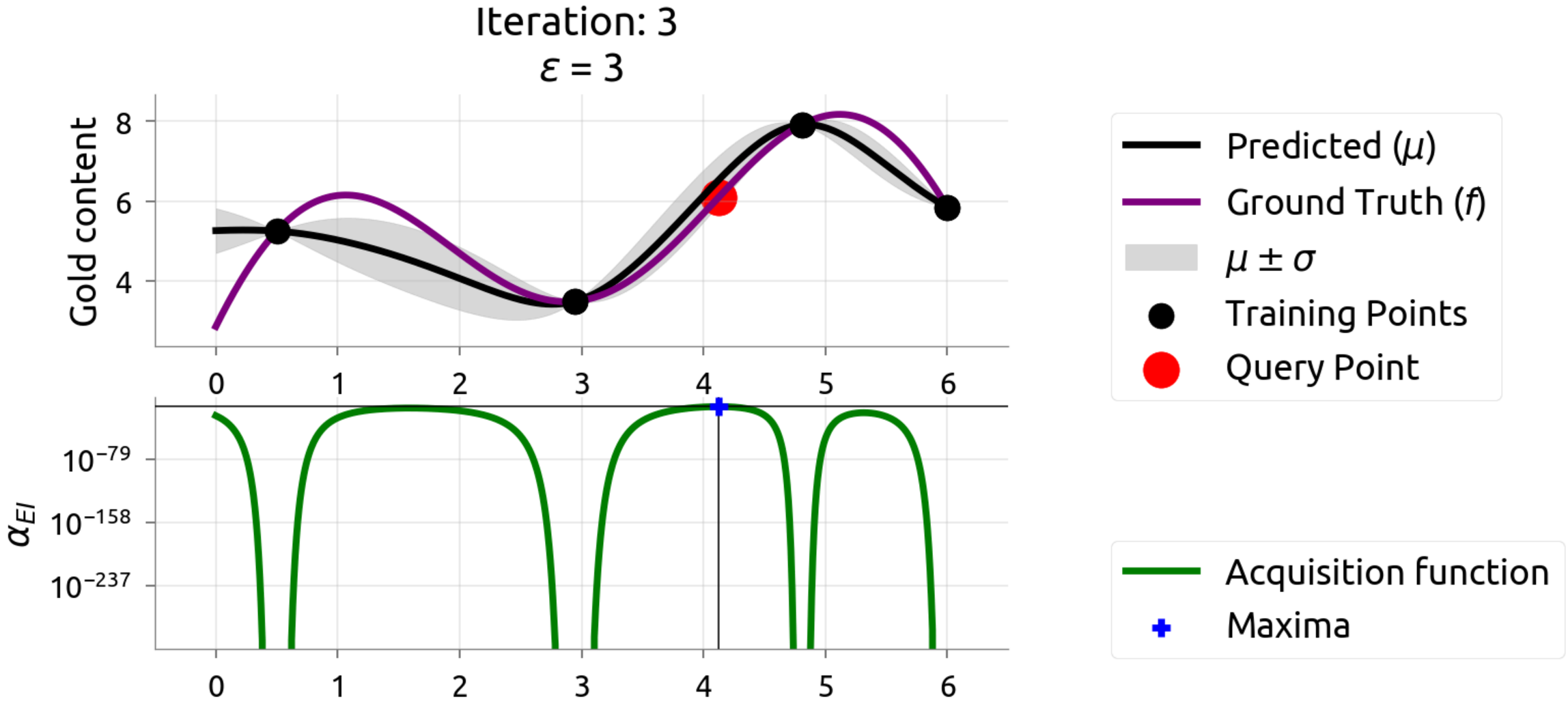

Acquisition function
Maxima

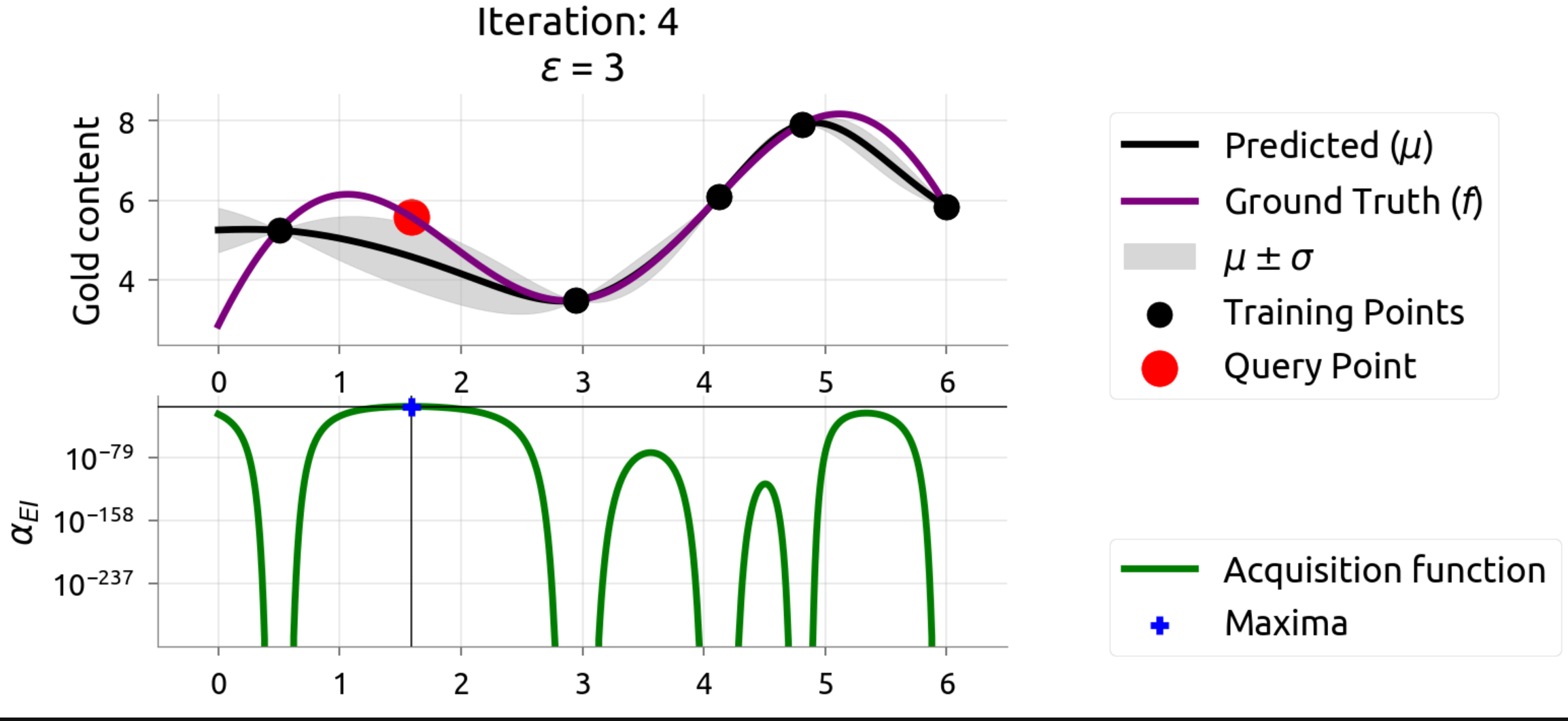
Intuition behind ϵ in PI: ϵ = 3

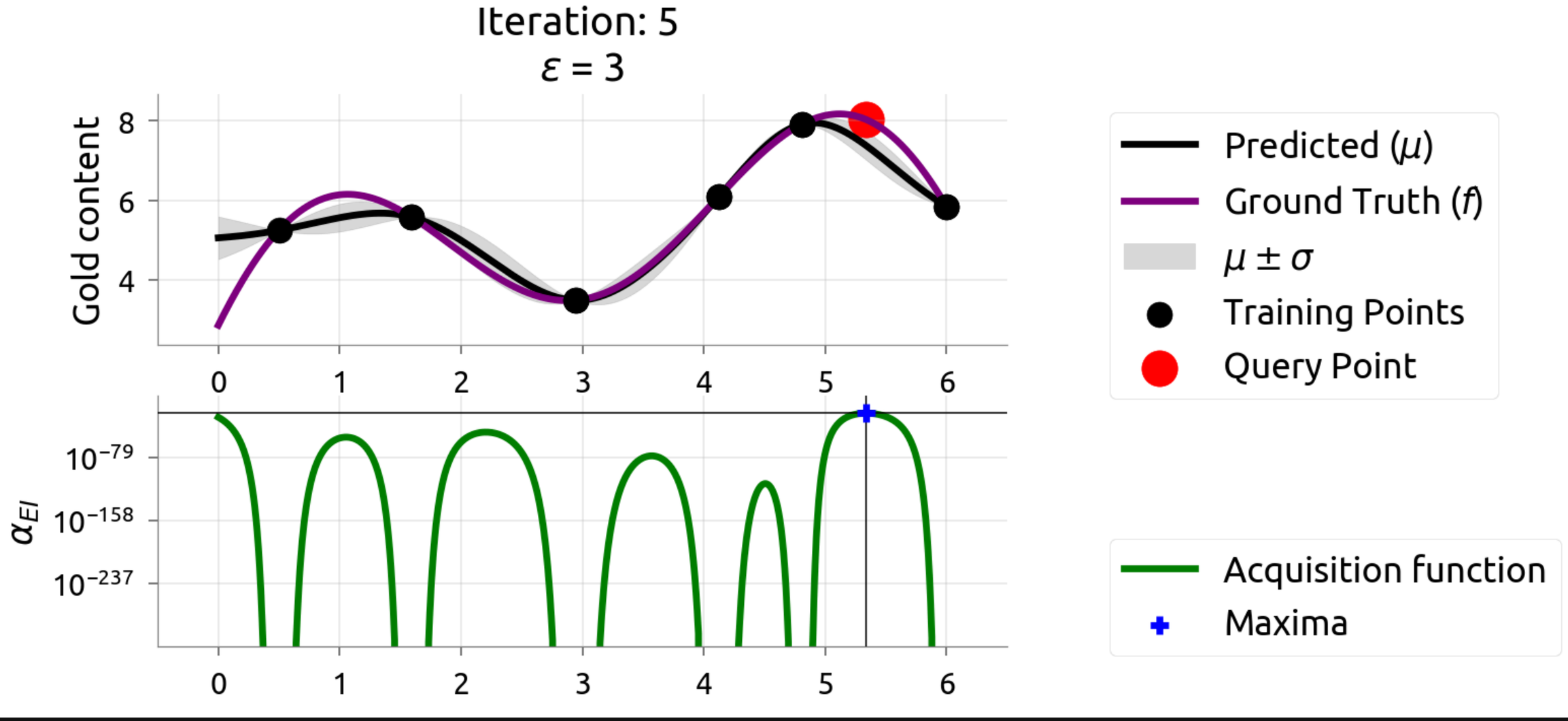


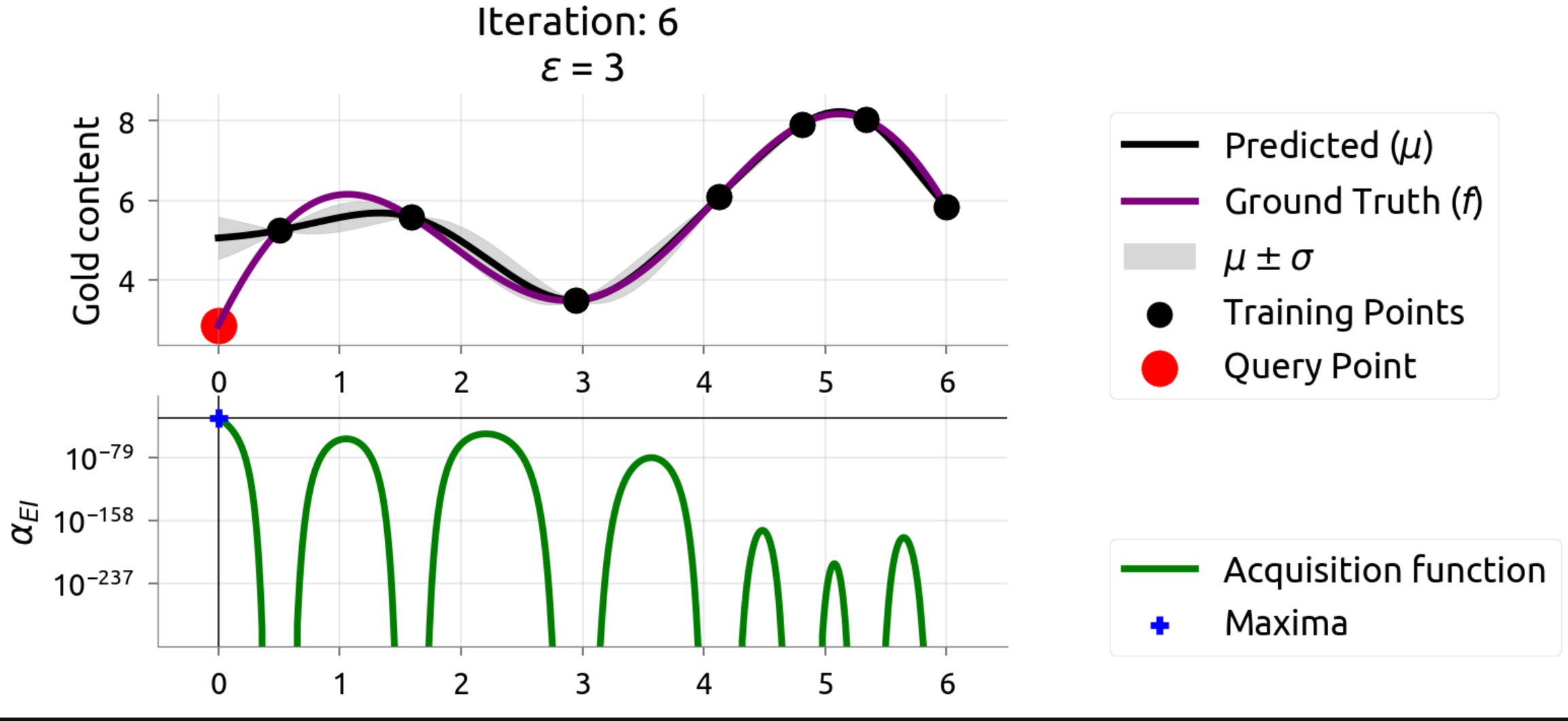

Iteration: 1
$\varepsilon = 3$
Gold content
8
6
4
0
1
2
3
4
5
6
α_{EI}
10^{-79}
10^{-158}
10^{-237}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

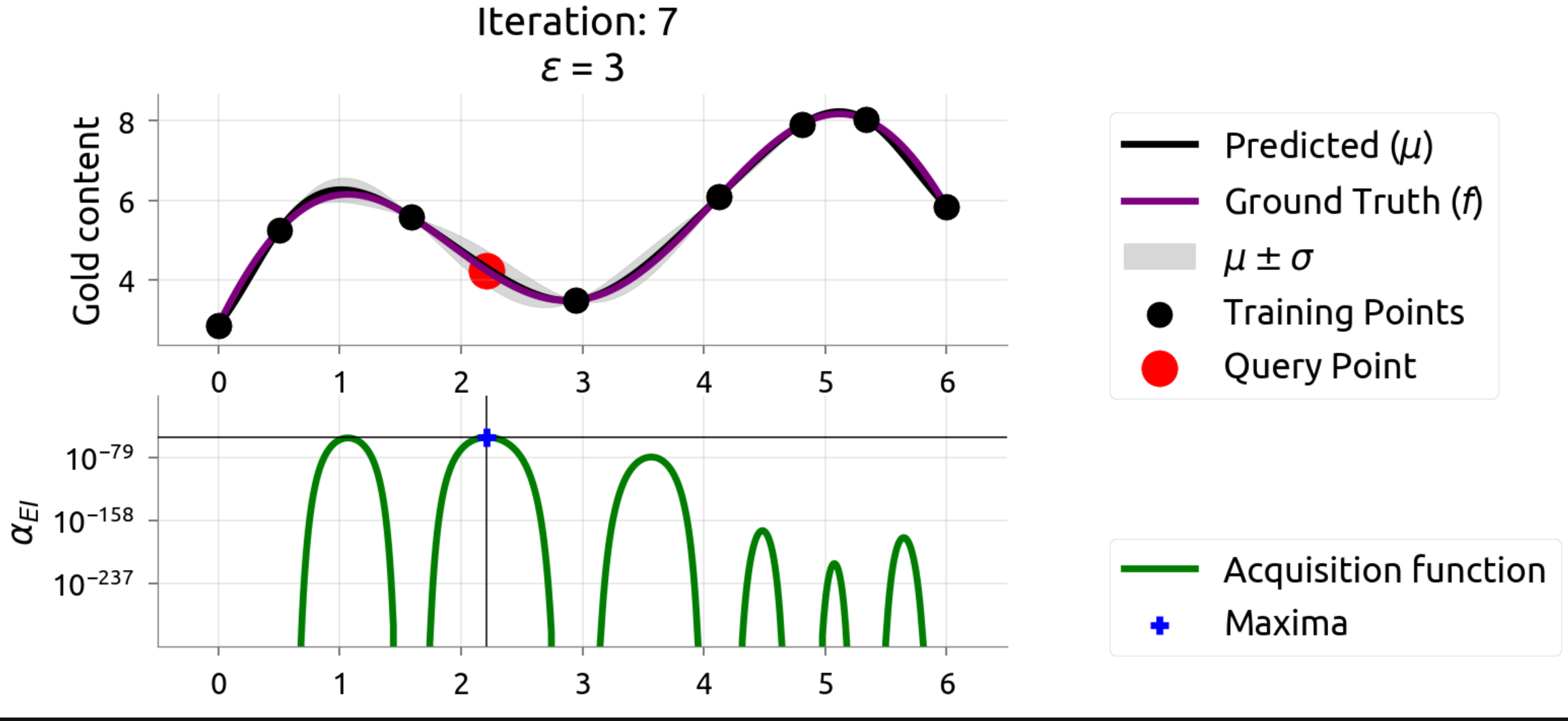

Iteration: 2
$\varepsilon = 3$
Gold content
8
6
4
0
1
2
3
4
5
6
α_{EI}
10^{-79}
10^{-158}
10^{-237}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

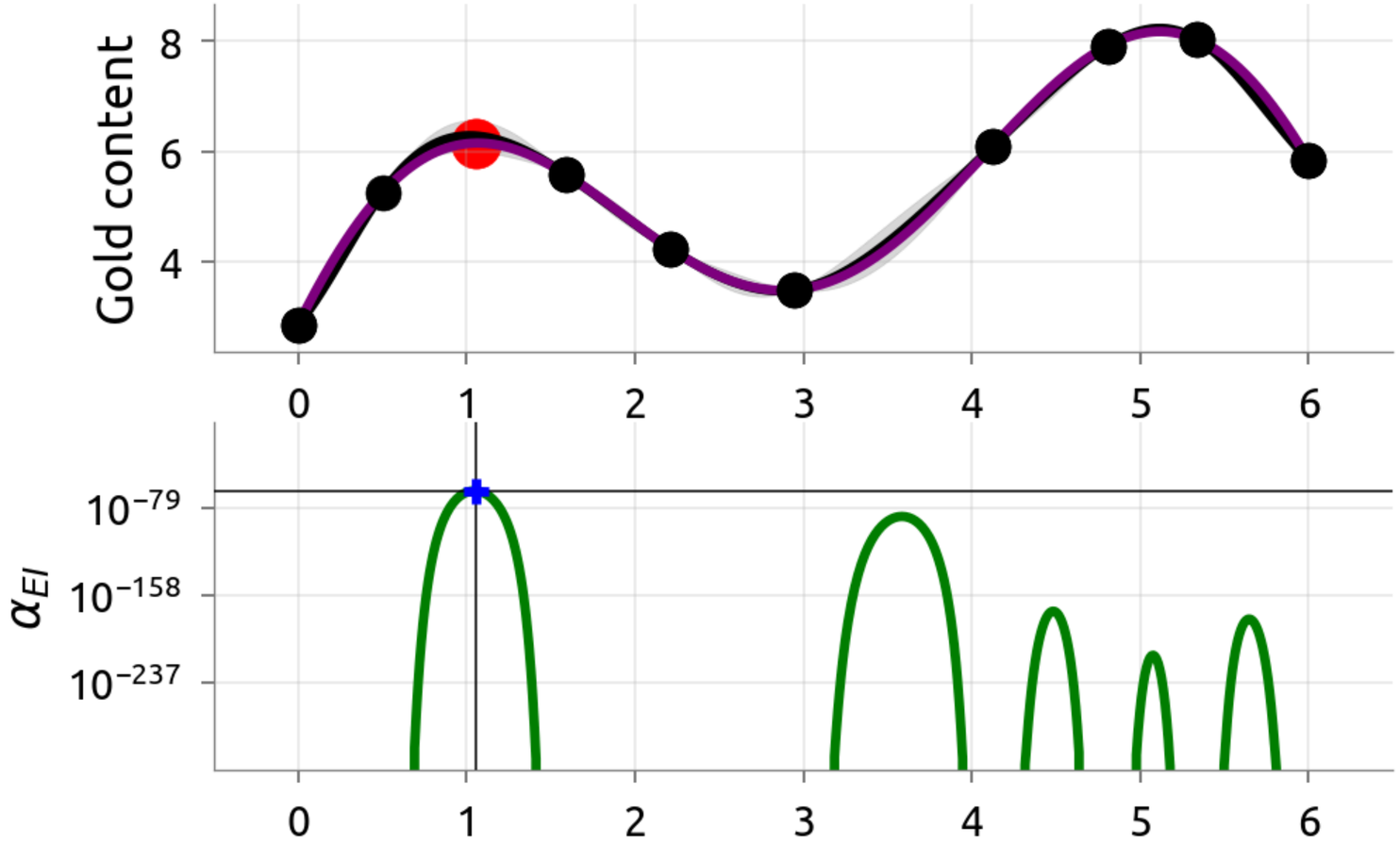

Iteration: 3
$\varepsilon = 3$
Gold content
8
6
4
0
1
2
3
4
5
6
α_{EI}
10^{-79}
10^{-158}
10^{-237}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

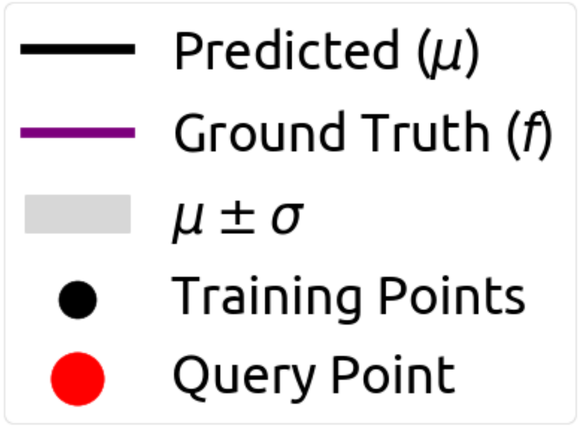

Iteration: 4
ε = 3
Gold content
8
6
4
0
1
2
3
4
5
6
α_{EI}
10^{-79}
10^{-158}
10^{-237}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

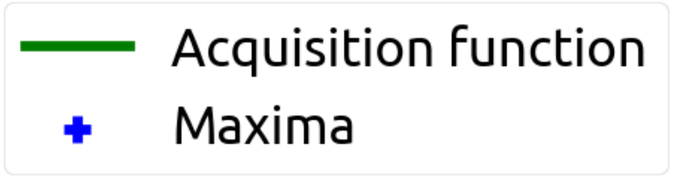

Iteration: 5
$\varepsilon = 3$
Gold content
8
6
4
0
1
2
3
4
5
6
α_{EI}
10^{-79}
10^{-158}
10^{-237}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

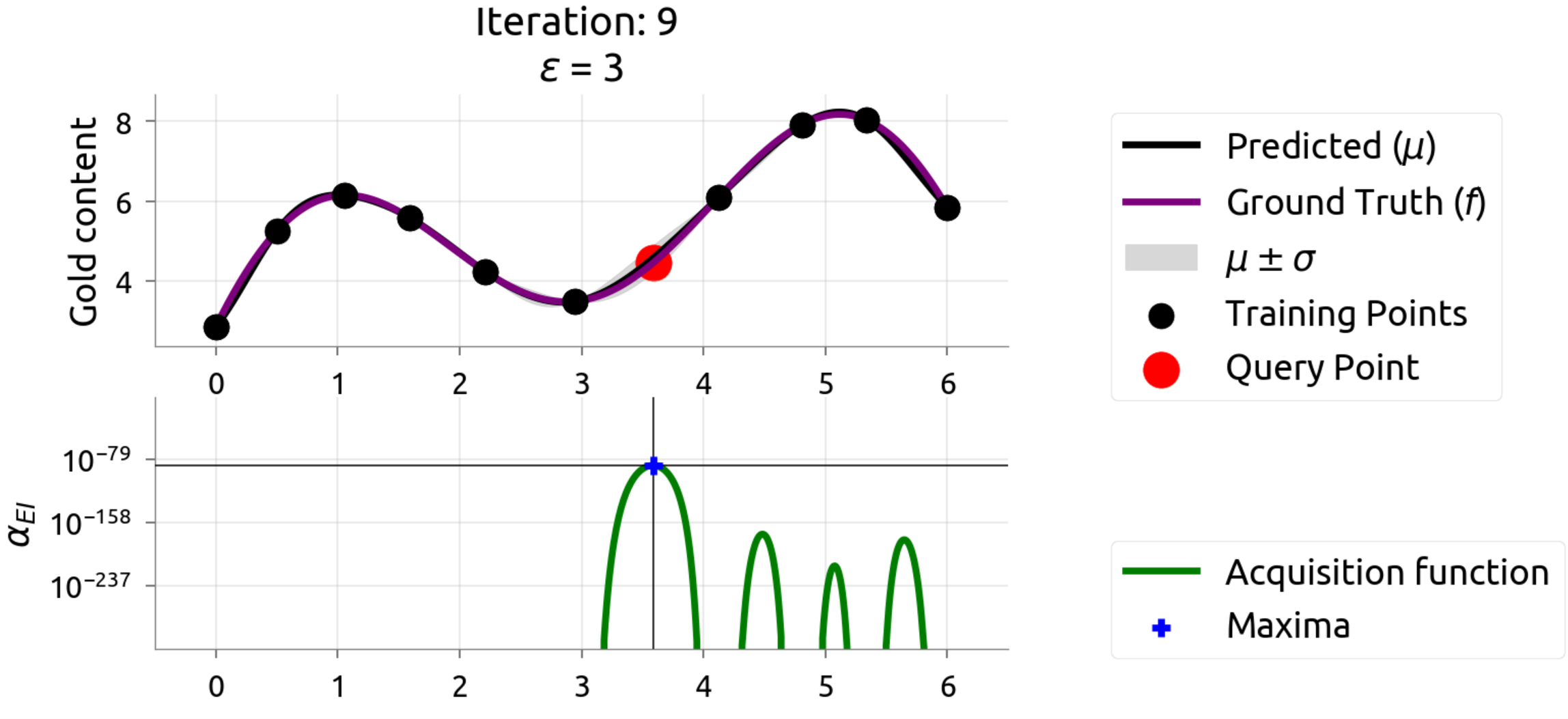

Iteration: 6
$\varepsilon = 3$
Gold content
8
6
4
0
1
2
3
4
5
6
α_{EI}
10^{-79}
10^{-158}
10^{-237}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima


Iteration: 7
$\varepsilon = 3$
Gold content
8
6
4
0
1
2
3
4
5
6
α_{EI}
10^{-79}
10^{-158}
10^{-237}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima


Iteration: 8
ε = 3
Gold content
8
6
4
0
1
2
3
4
5
6
α_{EI}
10^{-79}
10^{-158}
10^{-237}
0
1
2
3
4
5
6

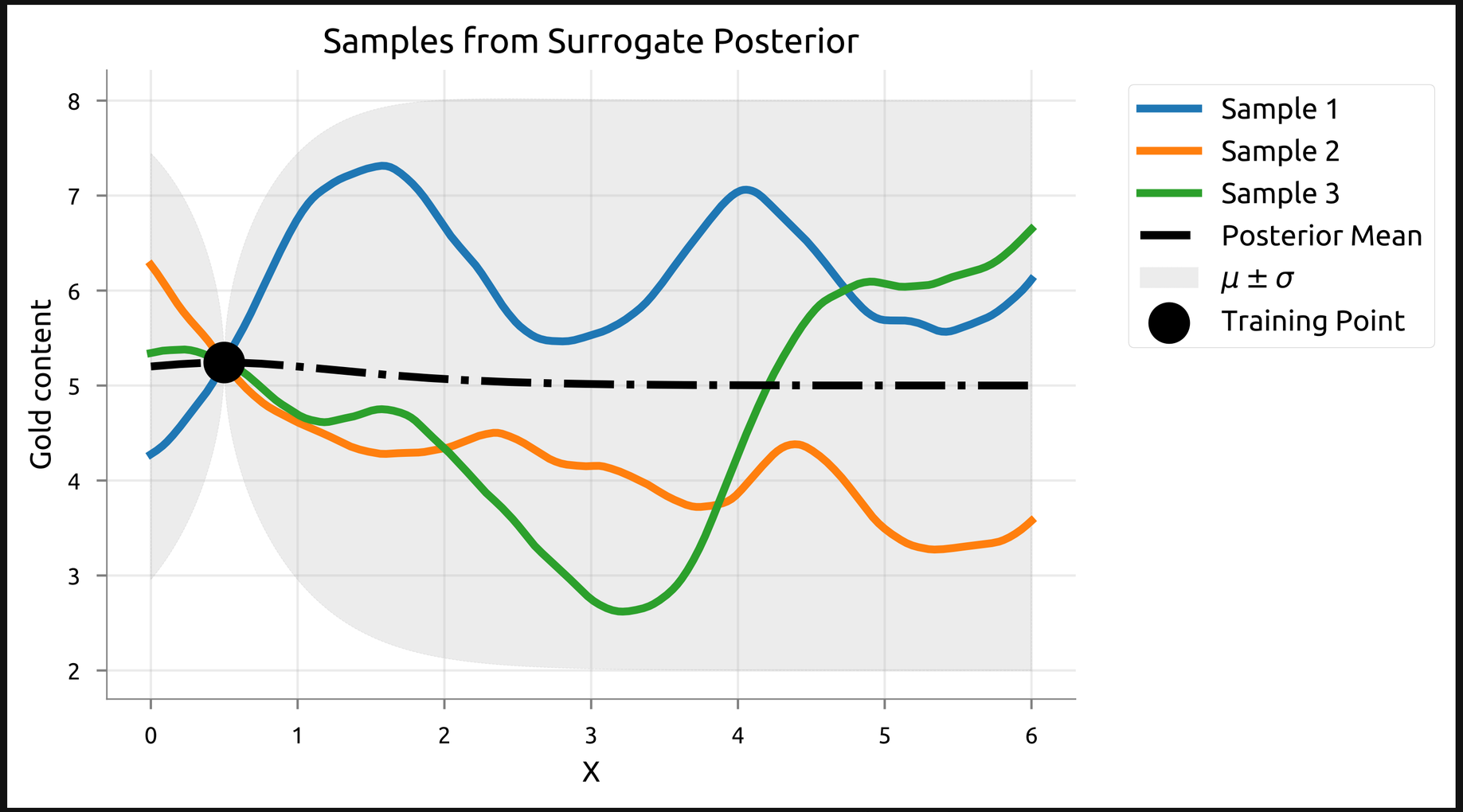

Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point


Acquisition function
Maxima

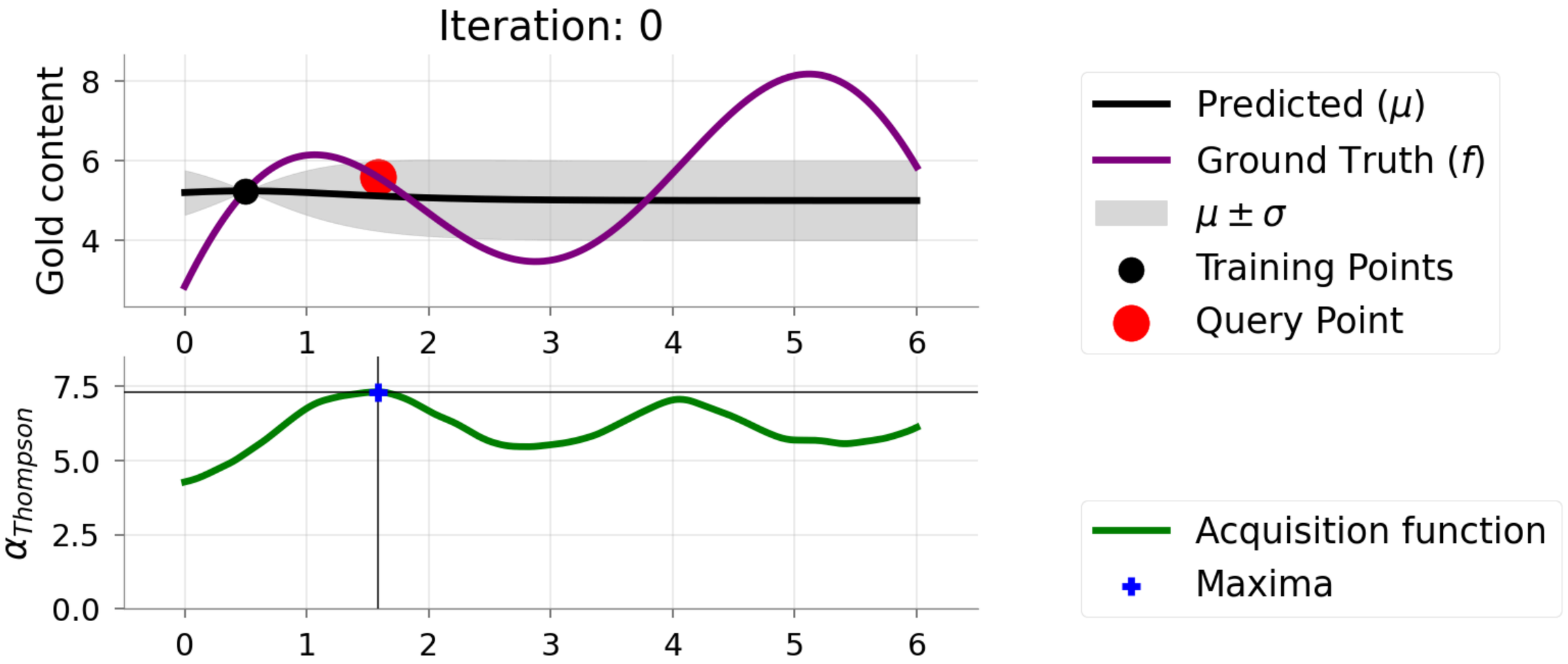

Iteration: 9
$\varepsilon = 3$
Gold content
8
6
4
0
1
2
3
4
5
6
α_{EI}
10^{-79}
10^{-158}
10^{-237}
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

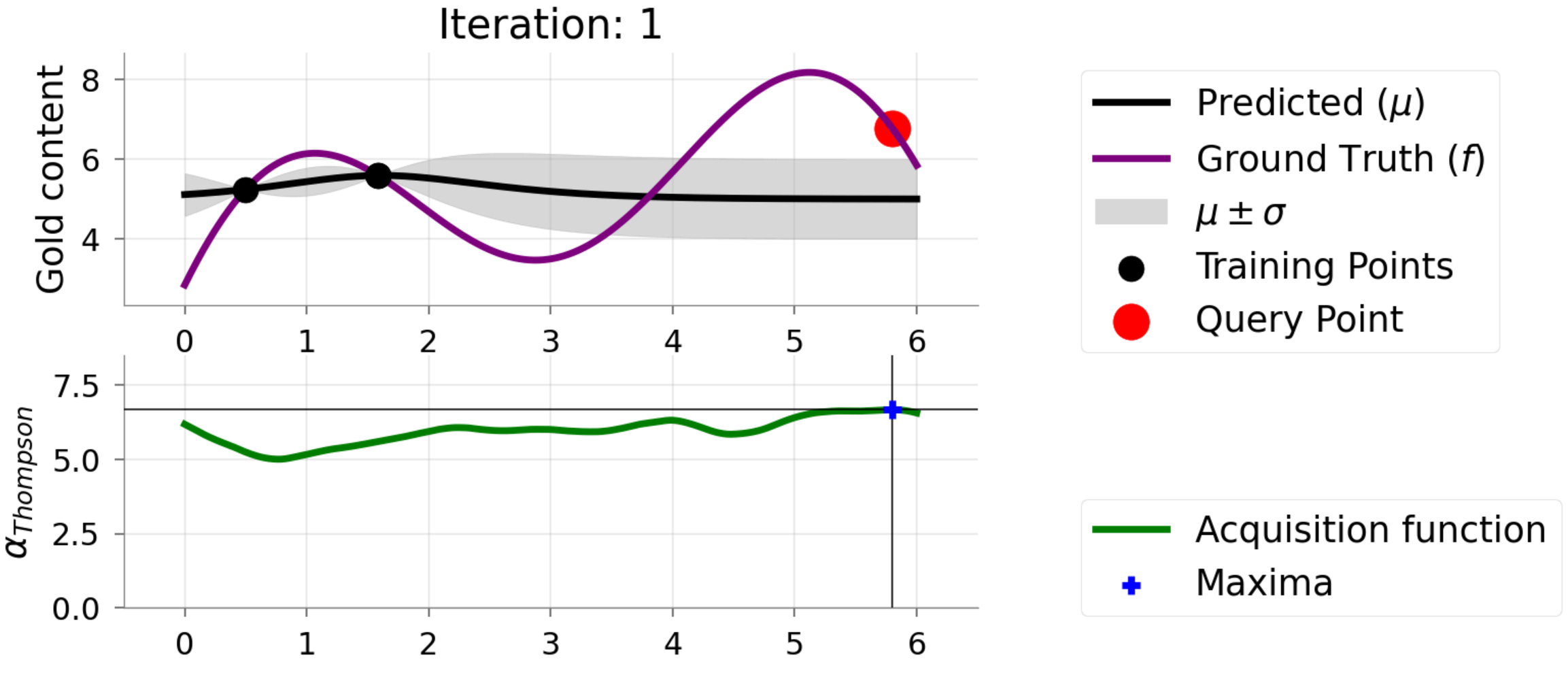
Thompson Sampling

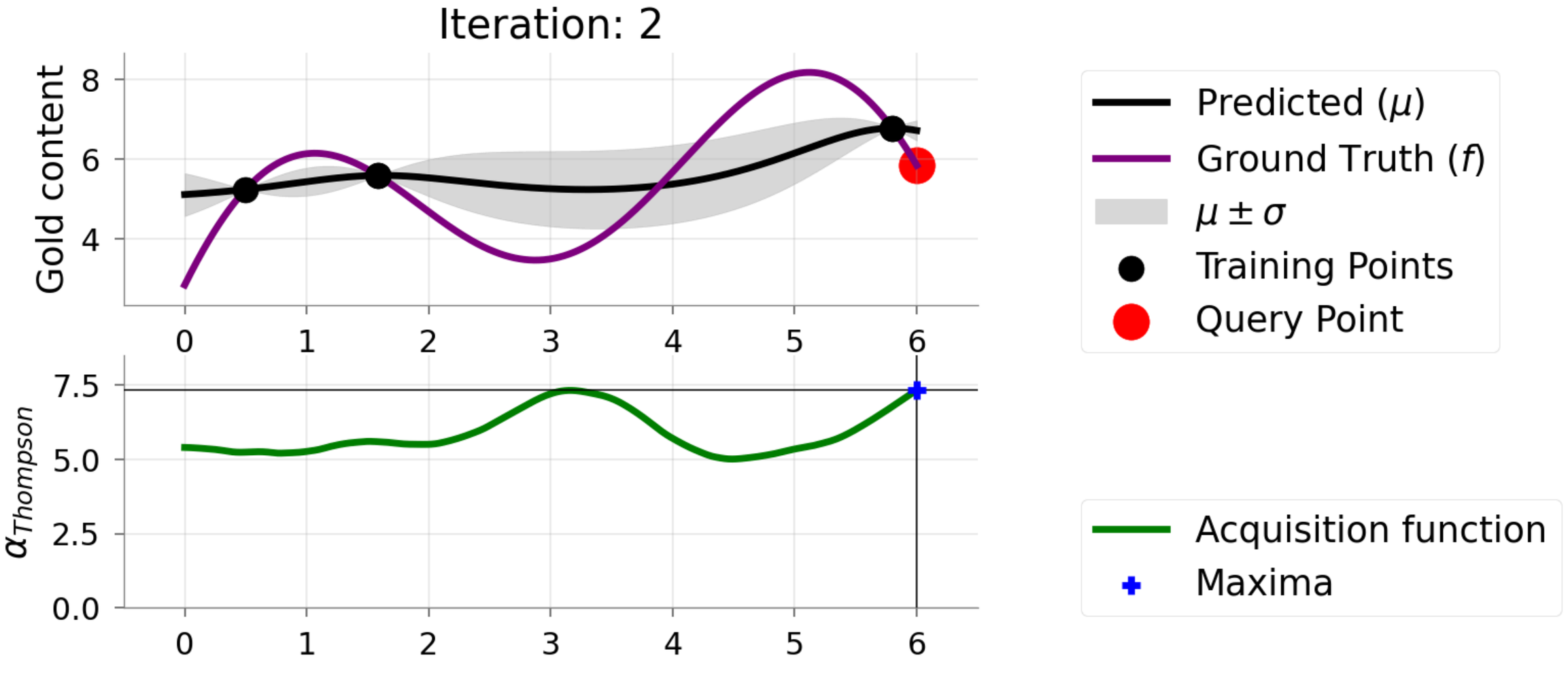
- Samples a function from the surrogate's posterior and optimizes it.
- Balances exploration and exploitation naturally.

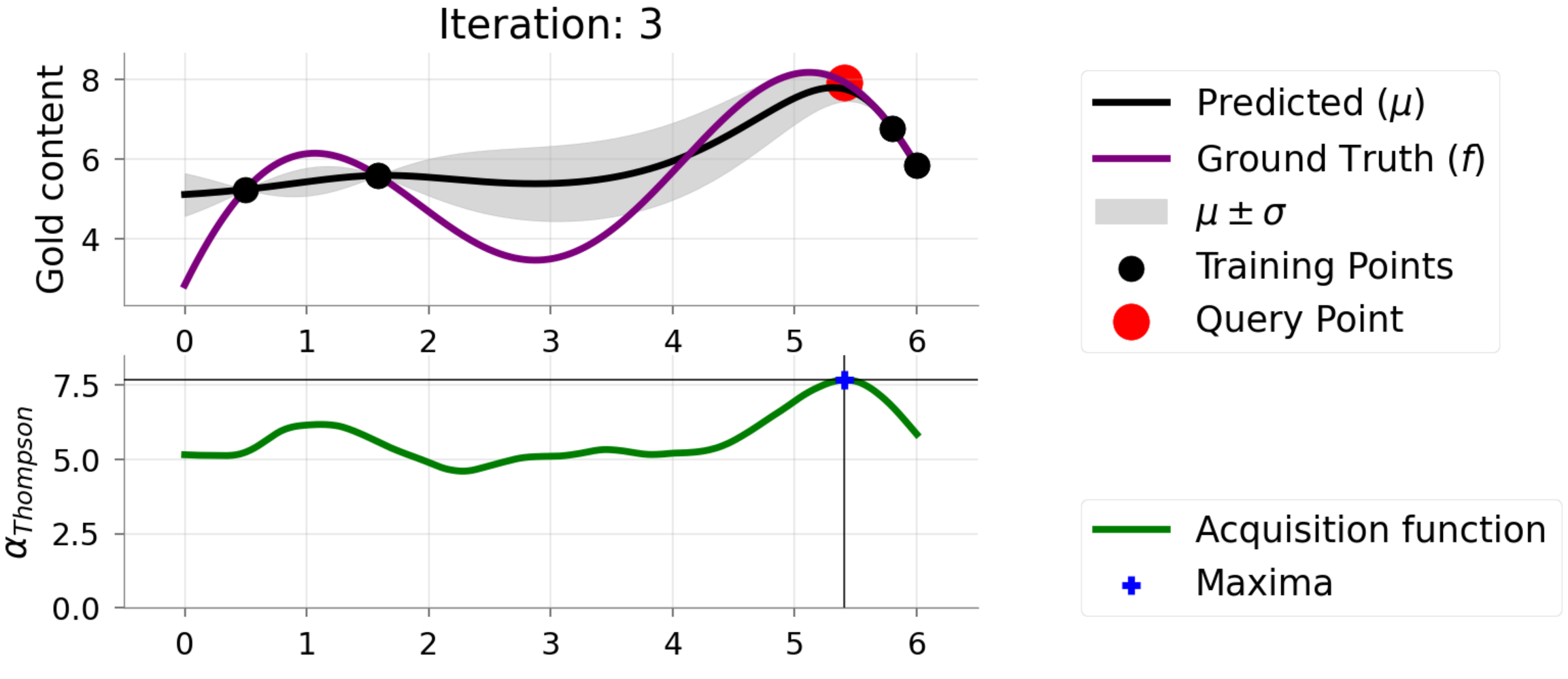


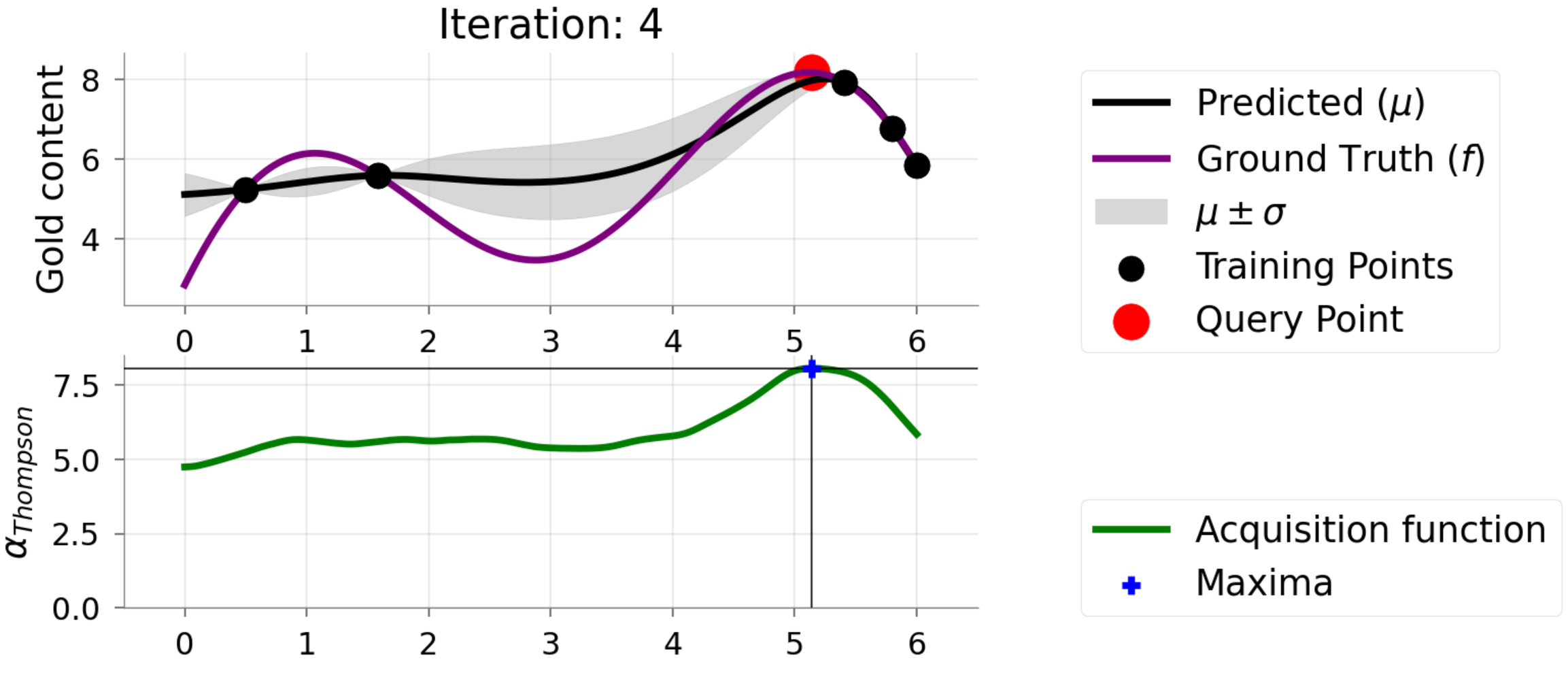
Intuition behind Thompson Sampling 1

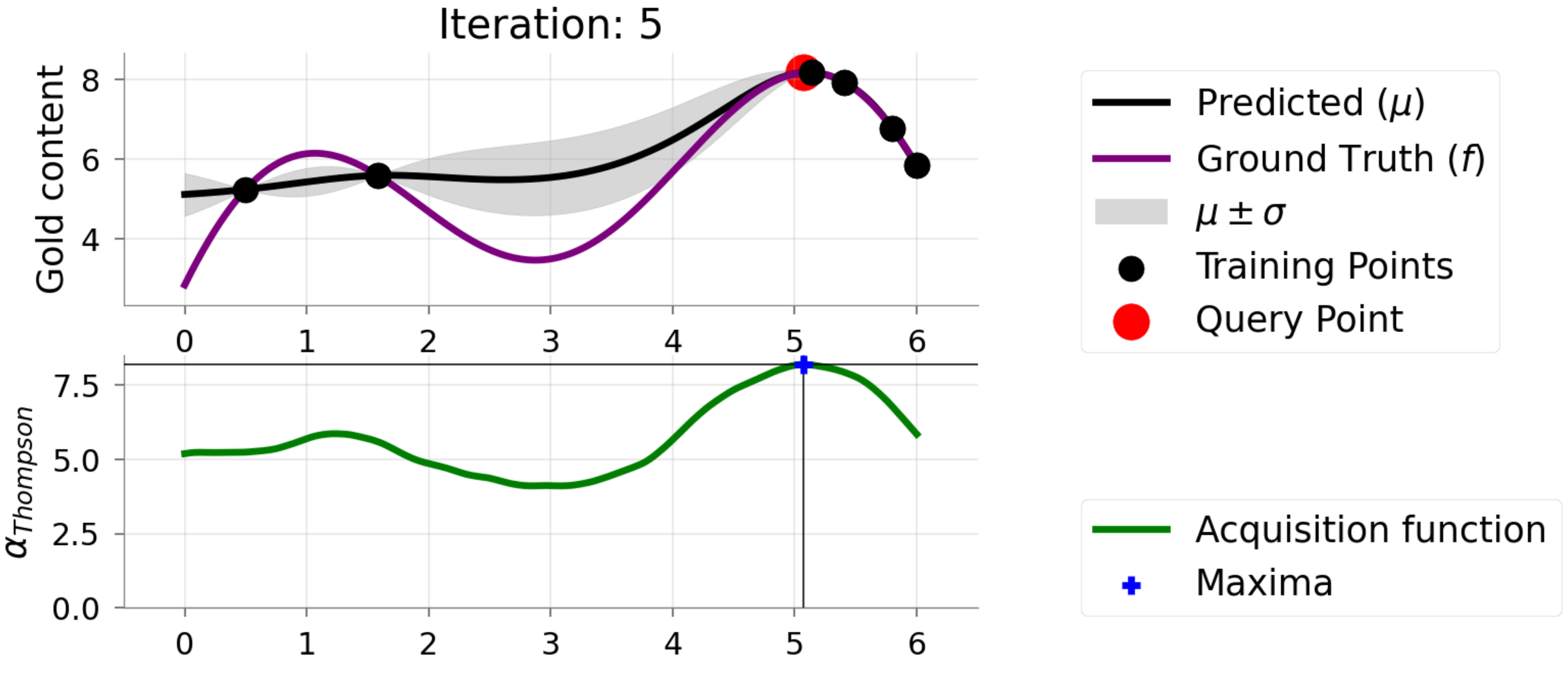


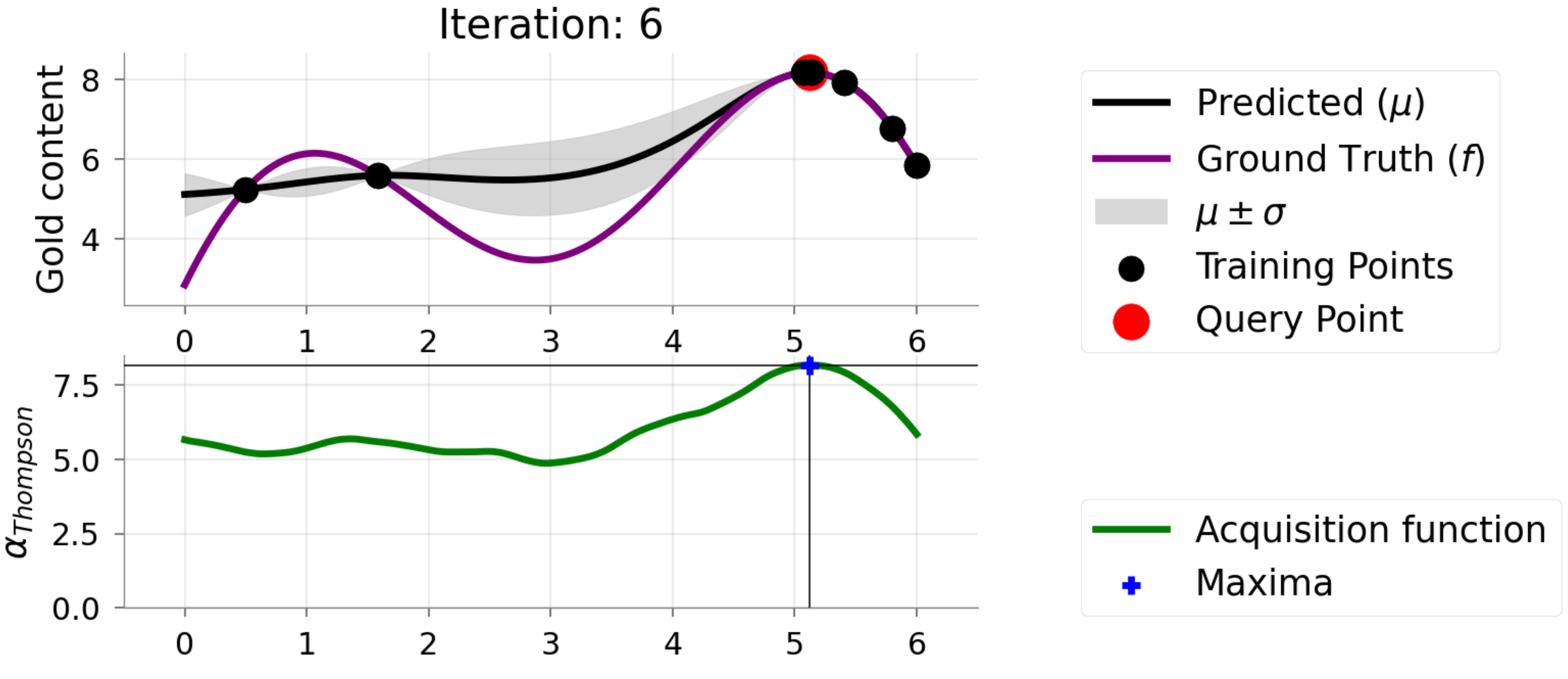

Iteration: 1
Gold content
8
6
4
0
1
2
3
4
5
6
$\alpha_{Thompson}$
7.5
5.0
2.5
0.0
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

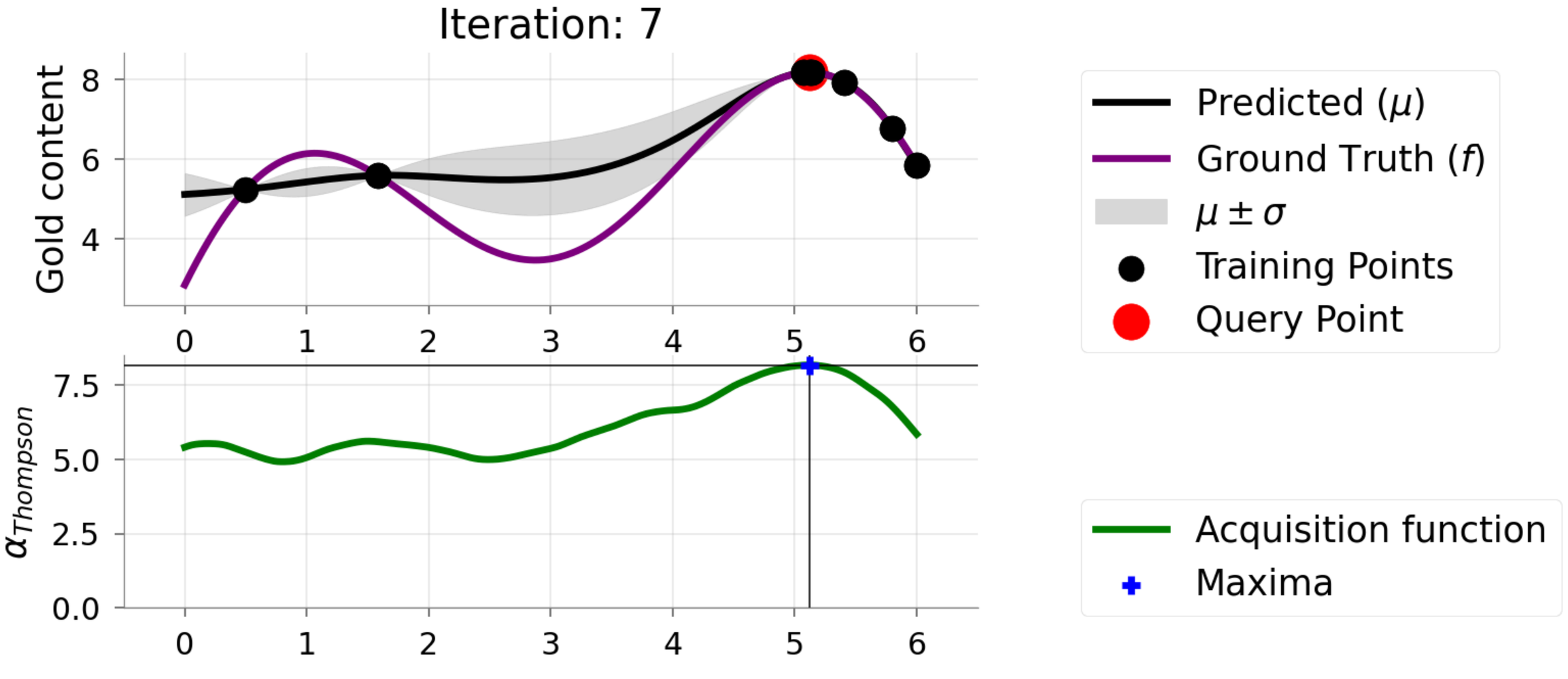

Iteration: 2
Gold content
8
6
4
0
1
2
3
4
5
6
$\alpha_{Thompson}$
7.5
5.0
2.5
0.0
0
1
2
3
4
5
6
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

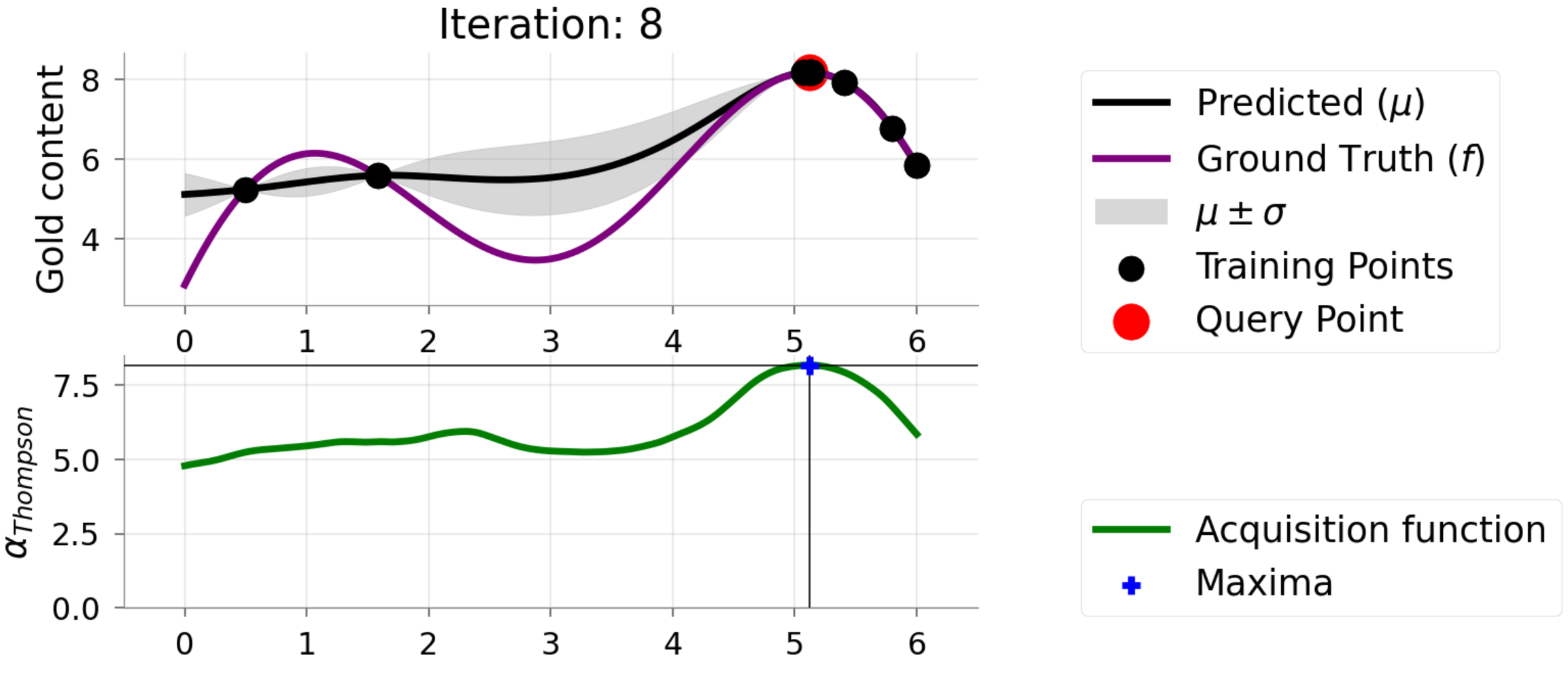

Iteration: 3
Gold content
8
6
4
0
1
2
3
4
5
6
$\alpha_{Thompson}$
7.5
5.0
2.5
0.0
0
1
2
3
4
5
6
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

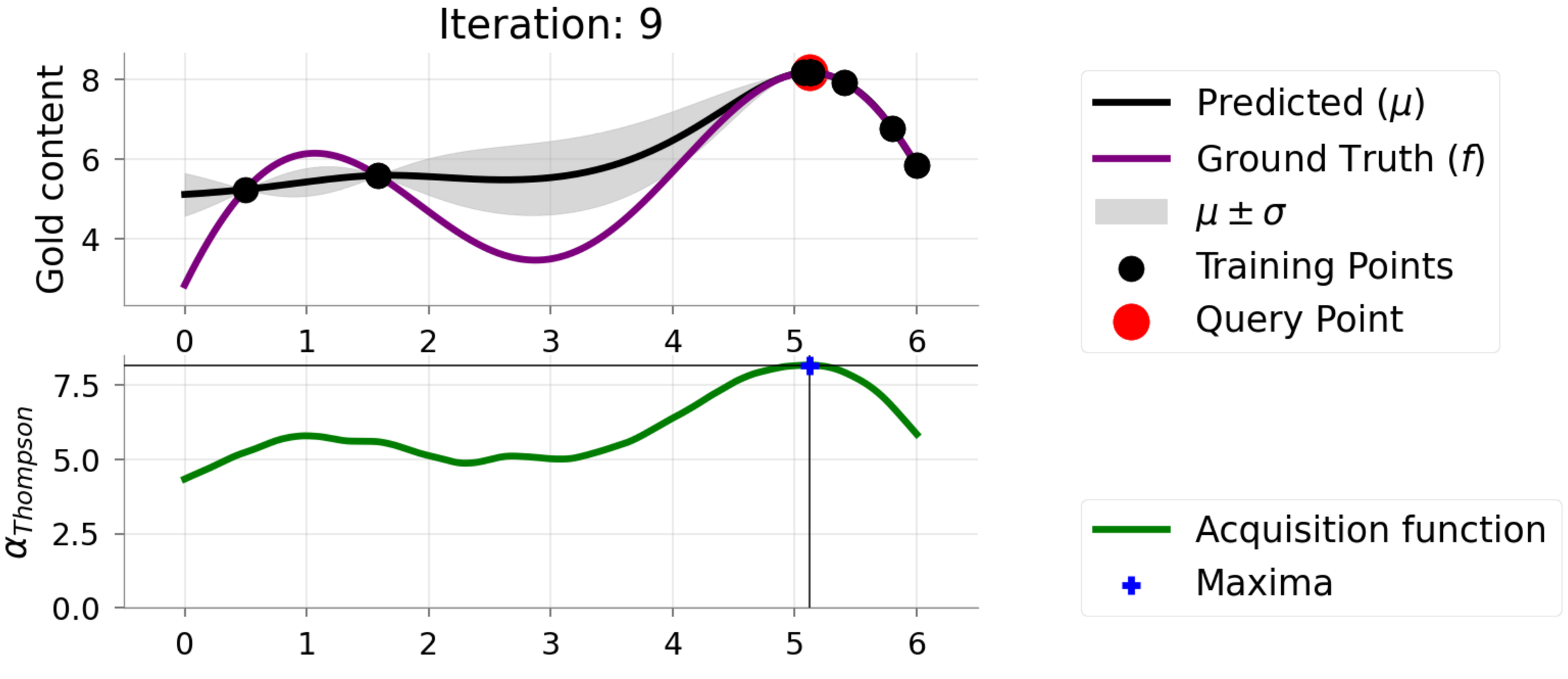

Iteration: 4
Gold content
8
6
4
0
1
2
3
4
5
6
$\alpha_{Thompson}$
7.5
5.0
2.5
0.0
0
1
2
3
4
5
6
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima


Iteration: 5
Gold content
8
6
4
0
1
2
3
4
5
6
$\alpha_{Thompson}$
7.5
5.0
2.5
0.0
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima


Iteration: 6
Gold content
8
6
4
0
1
2
3
4
5
6
$\alpha_{Thompson}$
7.5
5.0
2.5
0.0
0
1
2
3
4
5
6
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima


Iteration: 7
Gold content
8
6
4
0
1
2
3
4
5
6
$\alpha_{Thompson}$
7.5
5.0
2.5
0.0
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima


Iteration: 8
Gold content
8
6
4
0
1
2
3
4
5
6
$\alpha_{Thompson}$
7.5
5.0
2.5
0.0
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima


Iteration: 9
Gold content
8
6
4
0
1
2
3
4
5
6
$\alpha_{Thompson}$
7.5
5.0
2.5
0.0
0
1
2
3
4
5
6
Predicted (μ)
Ground Truth (f)
$\mu \pm \sigma$
Training Points
Query Point
Acquisition function
Maxima

Intuition behind Thompson Sampling 2

- Locations with high uncertainty ($\sigma(x)$) will show a large variance in the functional values sampled from the surrogate posterior
  - There is a non-trivial probability that a sample can take high value in a highly uncertain region
  - Optimizing such samples can aid exploration
- Example:
  - The three samples (sample #1, #2, #3) show a high variance close to $x = 6$
  - Optimizing sample 3 will aid in exploration by evaluating $x = 6$

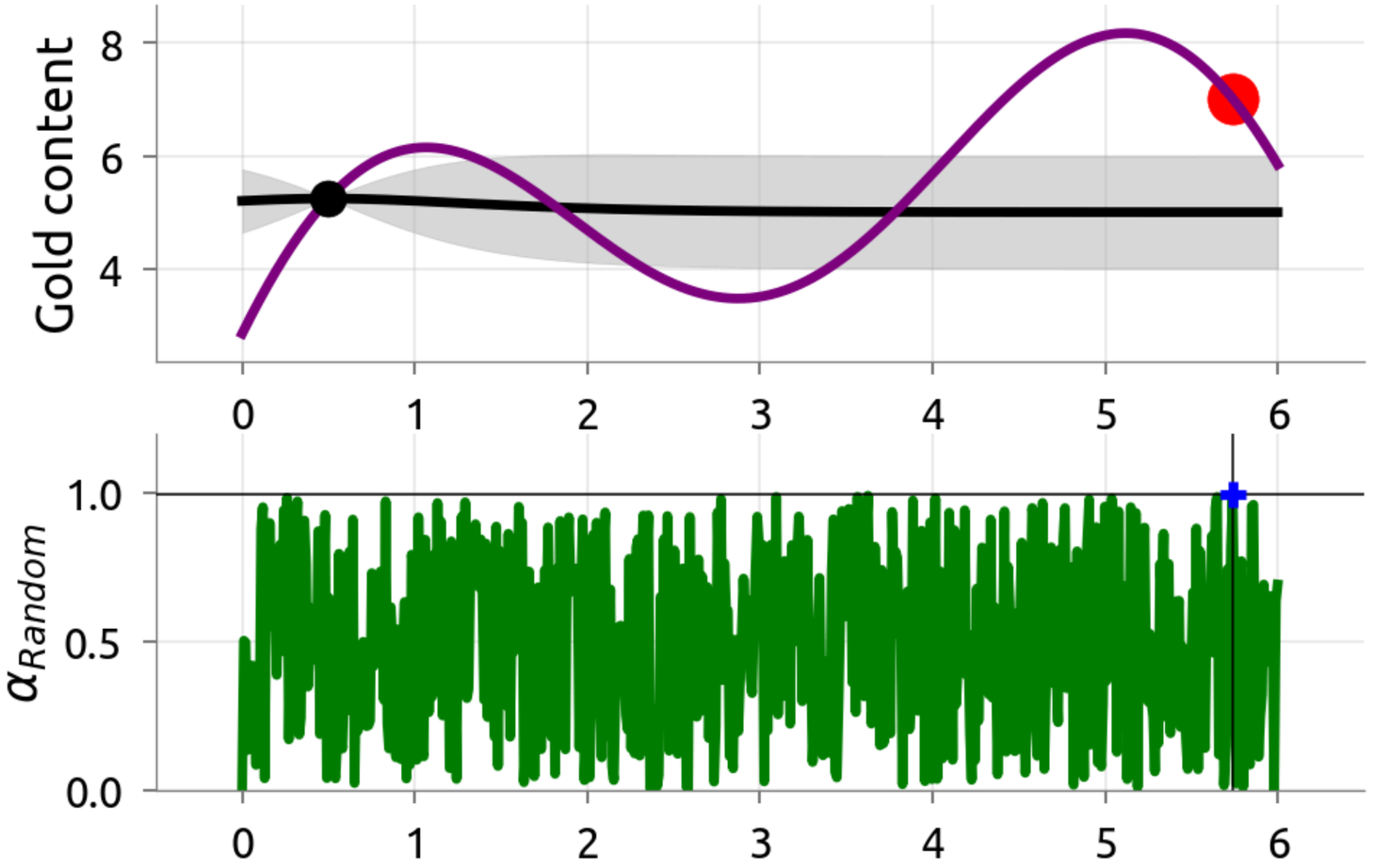
Thompson Sampling: Exploitation Behavior

- Sampled functions must pass through current max value
  - No uncertainty at evaluated locations
  - Optimizing samples from surrogate posterior ensures exploitation
- Example:
  - All sampled functions pass through current max at $x = 0.5$
  - If $x = 0.5$ is near global maxima, this enables:
    - Better exploitation of the region
    - Selection of improved maximum values

Random Sampling

- Randomly sample points from the search space
- No exploitation or exploration
- Useful for initial exploration or when no prior knowledge is available

Intuition behind Random Sampling



Predicted (μ)

Ground Truth (f)

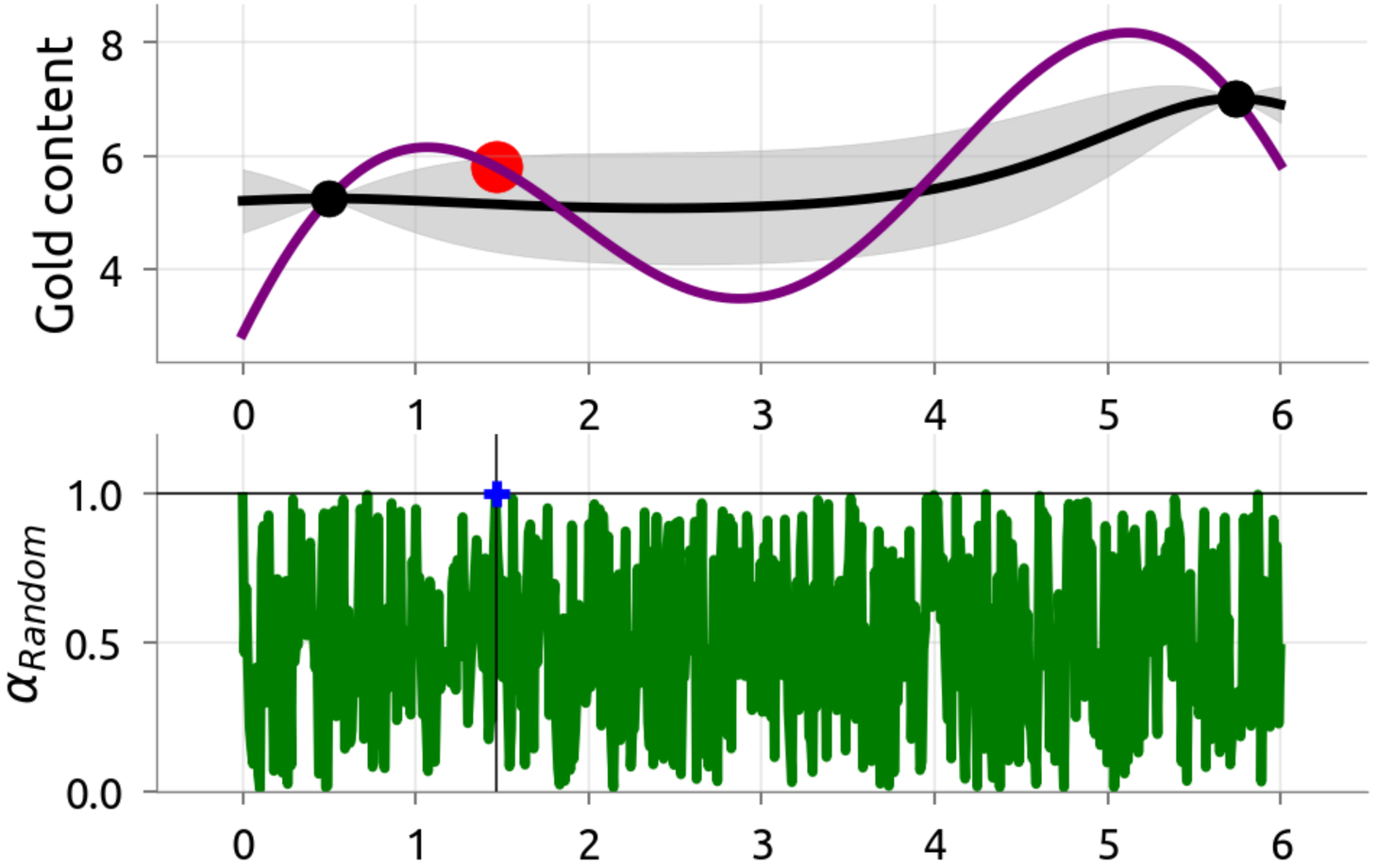
$\mu \pm \sigma$

Training Points

Query Point

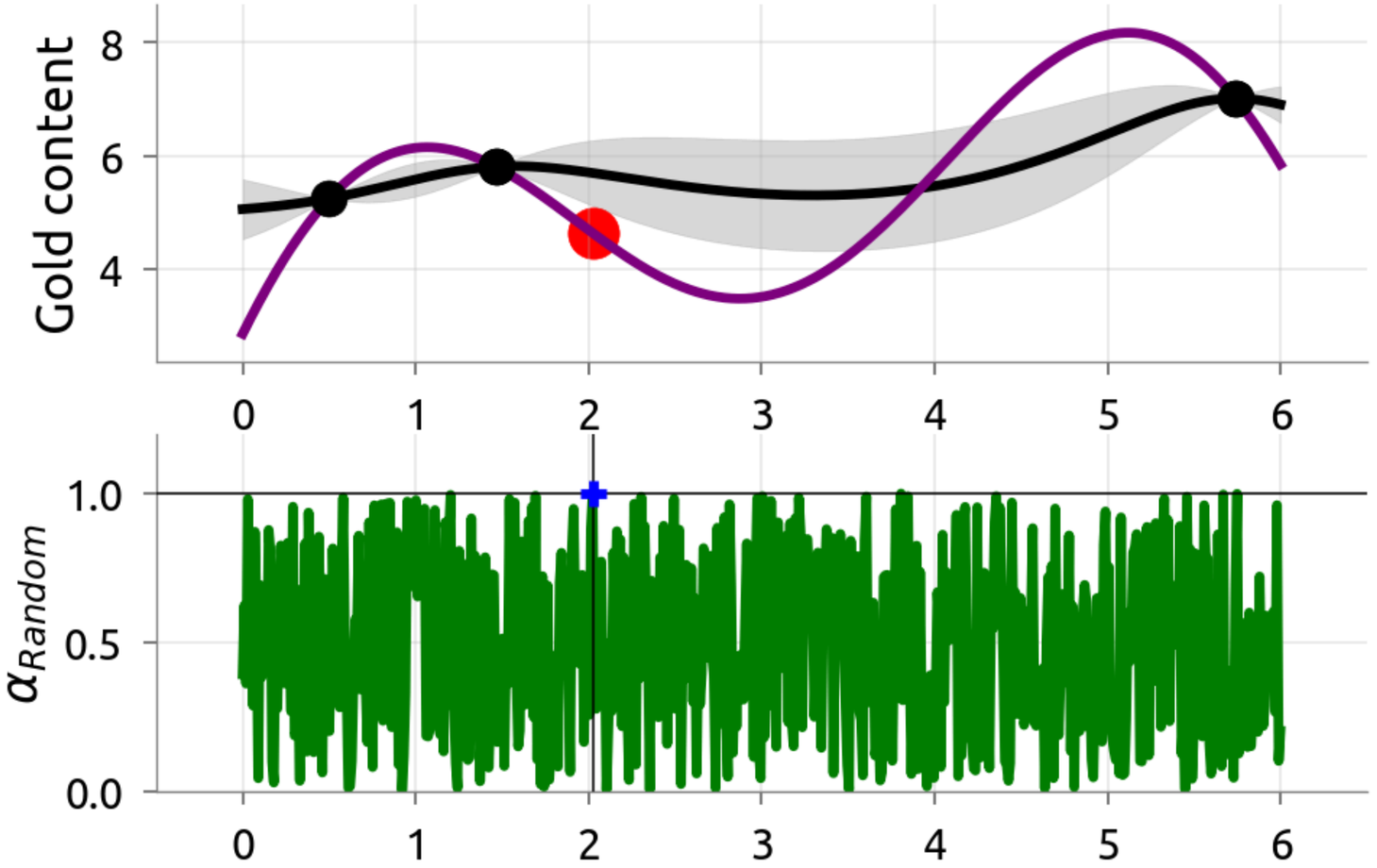
Acquisition function

Maxima



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

Acquisition function
Maxima



Predicted (μ)

Ground Truth (f)

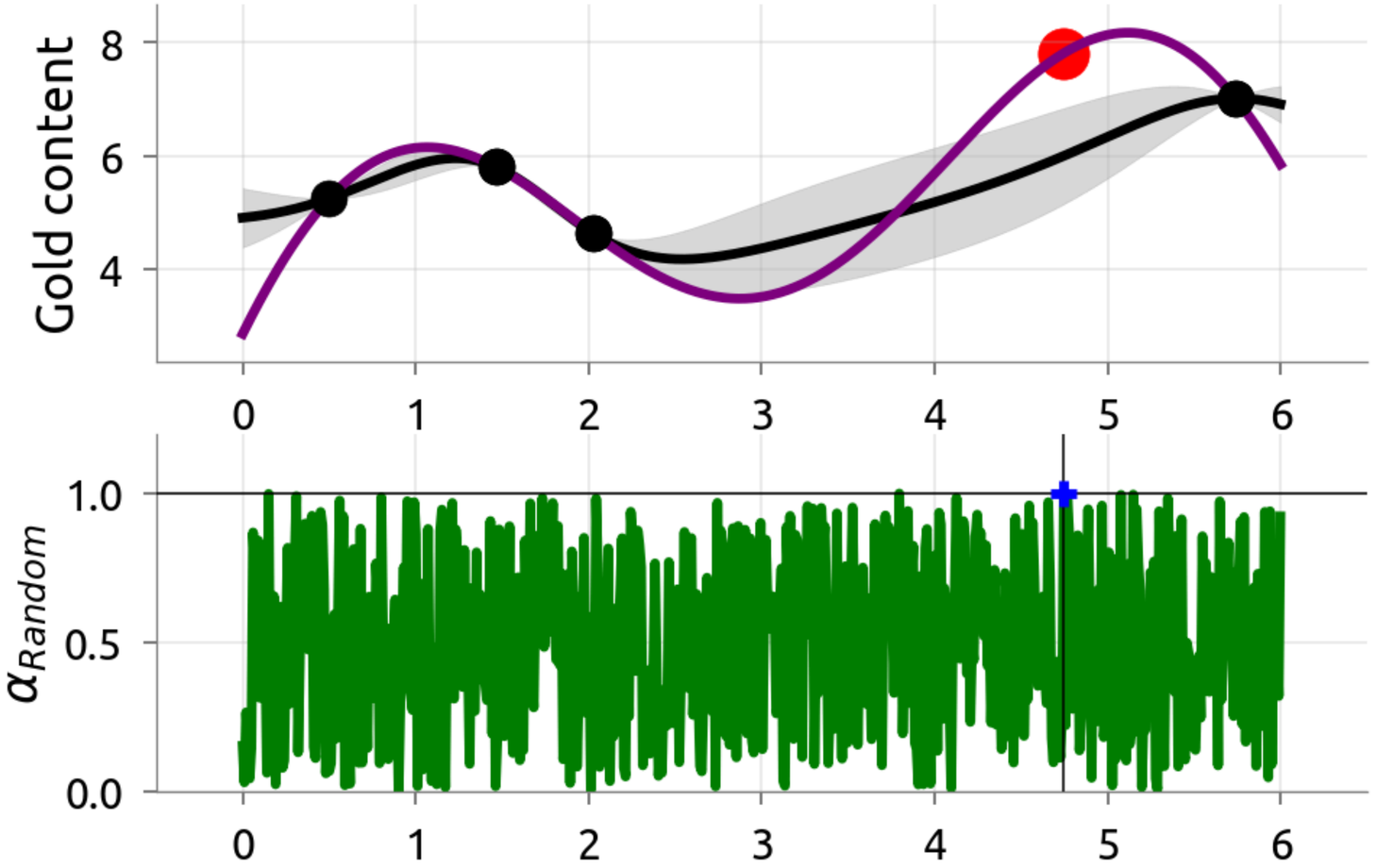
$\mu \pm \sigma$

Training Points

Query Point

Acquisition function

Maxima



Predicted (μ)

Ground Truth (f)

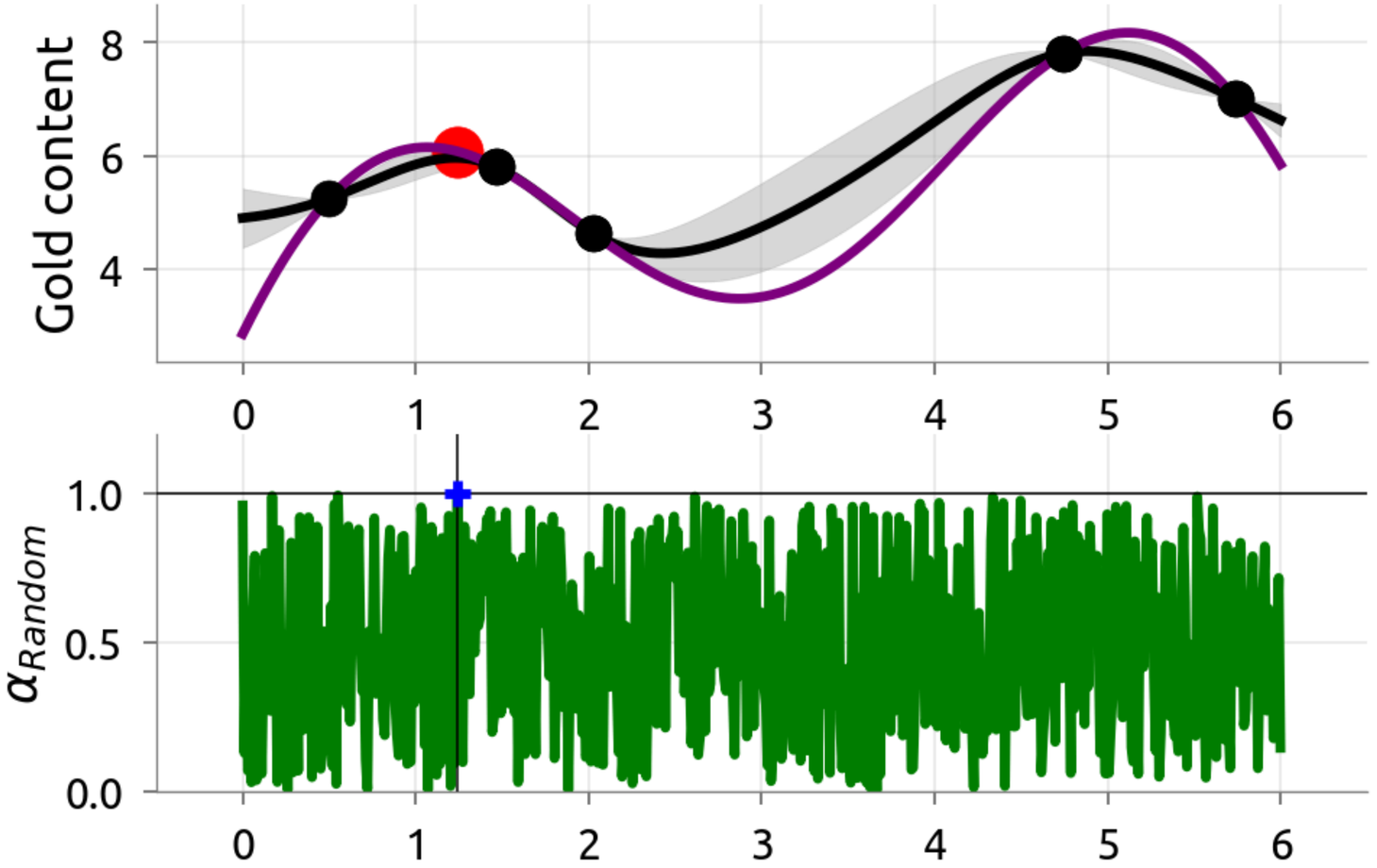
$\mu \pm \sigma$

Training Points

Query Point

Acquisition function

Maxima



Predicted (μ)

Ground Truth (f)

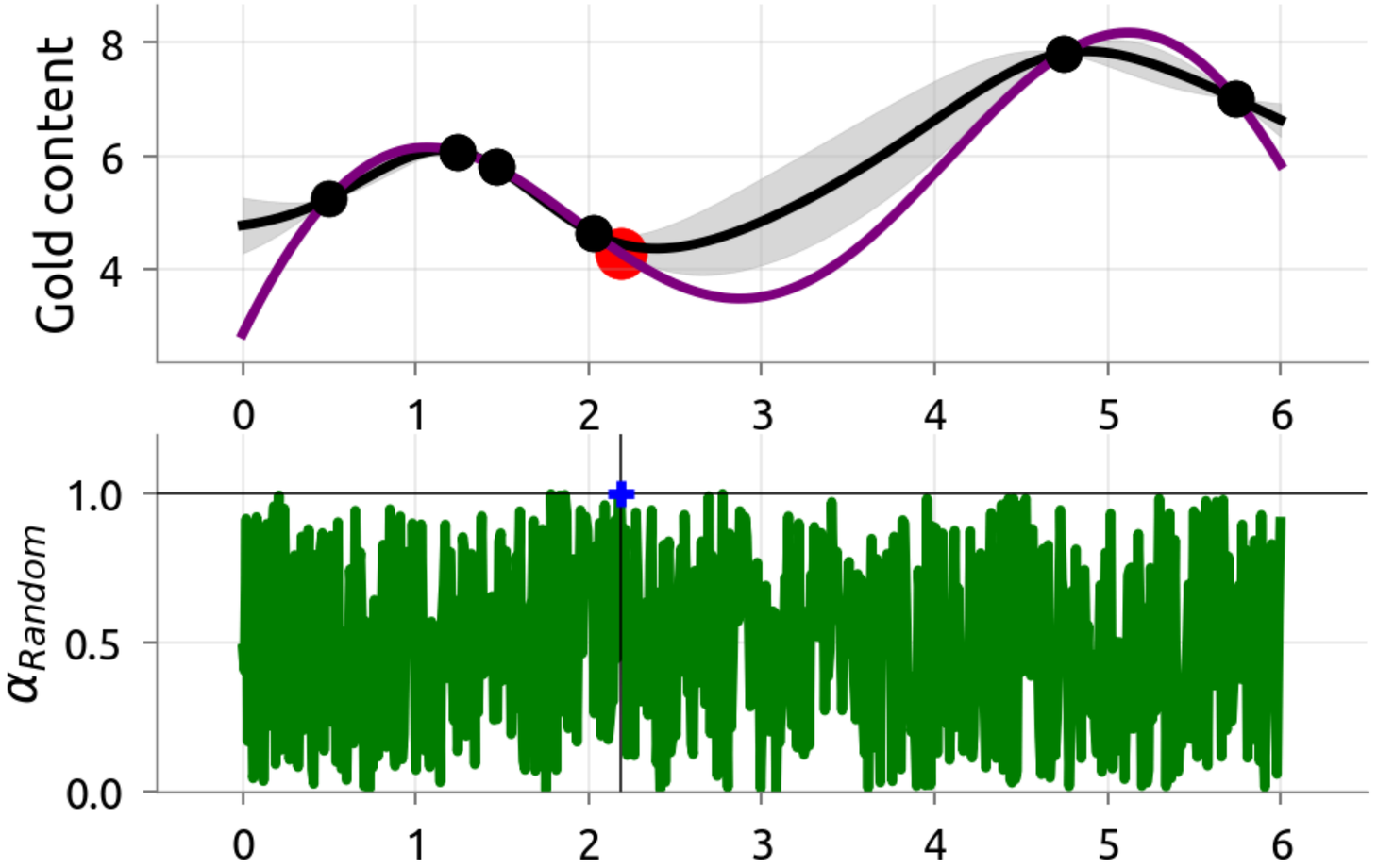
$\mu \pm \sigma$

Training Points

Query Point

Acquisition function

Maxima



Predicted (μ)

Ground Truth (f)

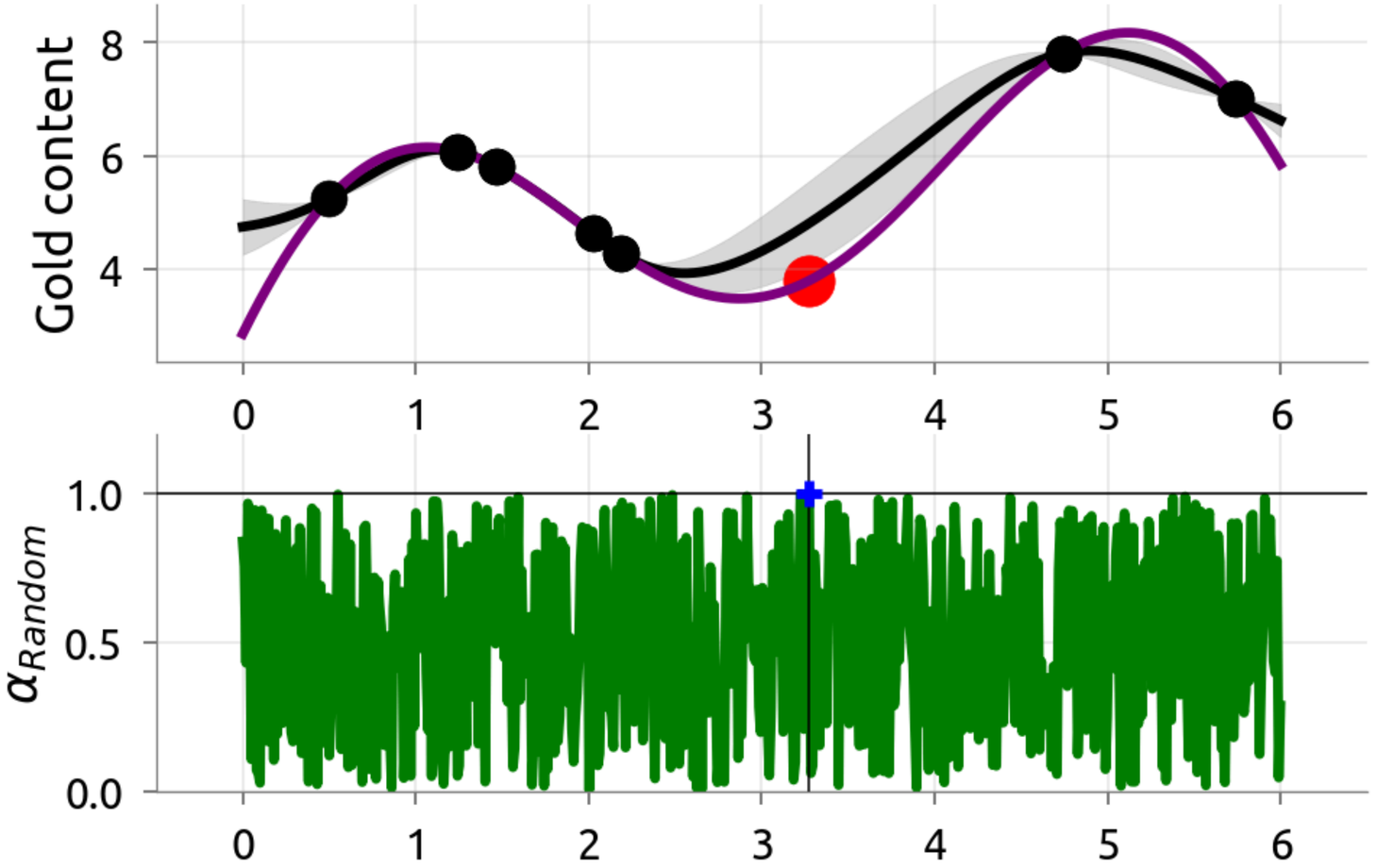
$\mu \pm \sigma$

Training Points

Query Point

Acquisition function

Maxima



Predicted (μ)

Ground Truth (f)

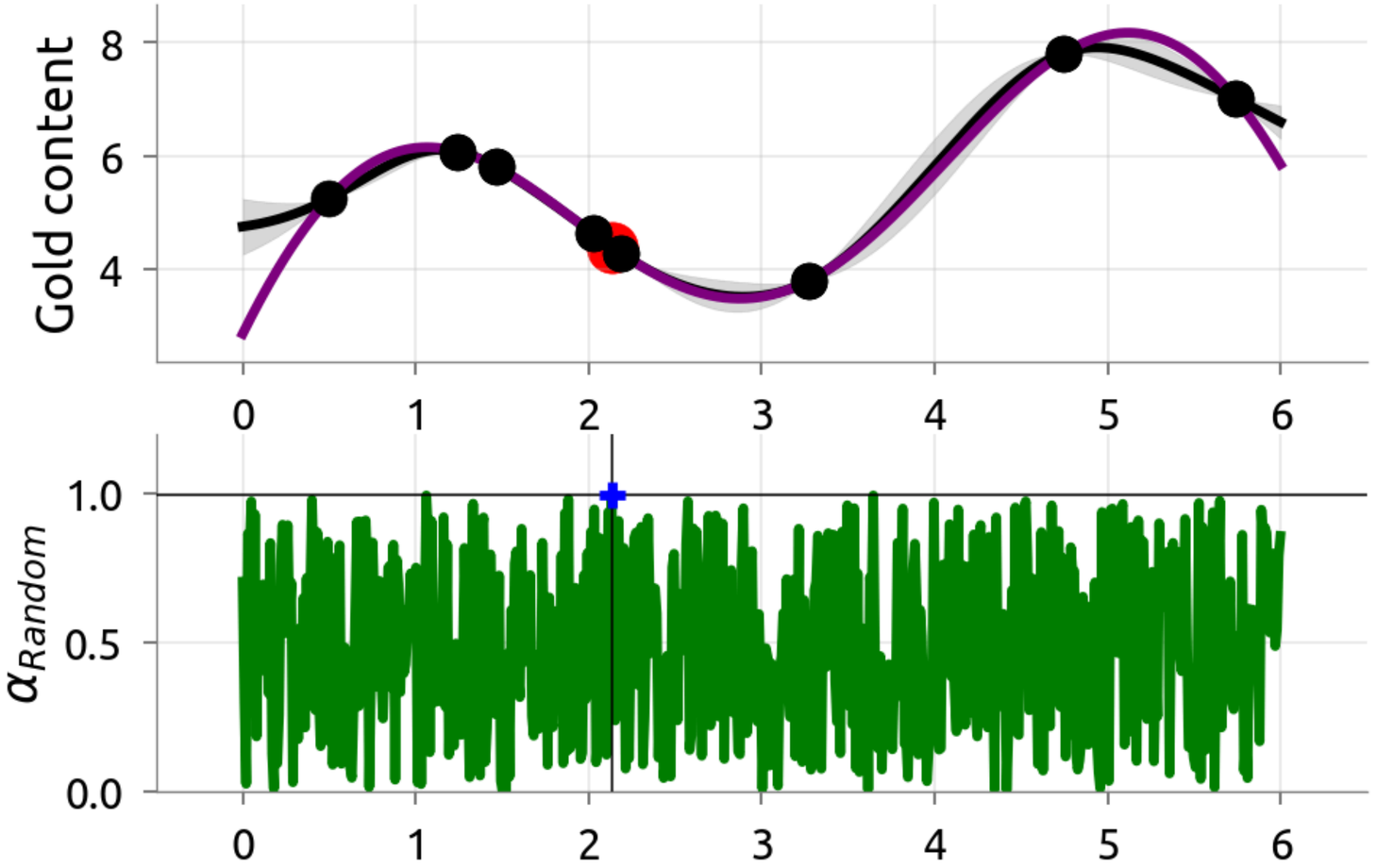
$\mu \pm \sigma$

Training Points

Query Point

Acquisition function

Maxima



Predicted (μ)

Ground Truth (f)

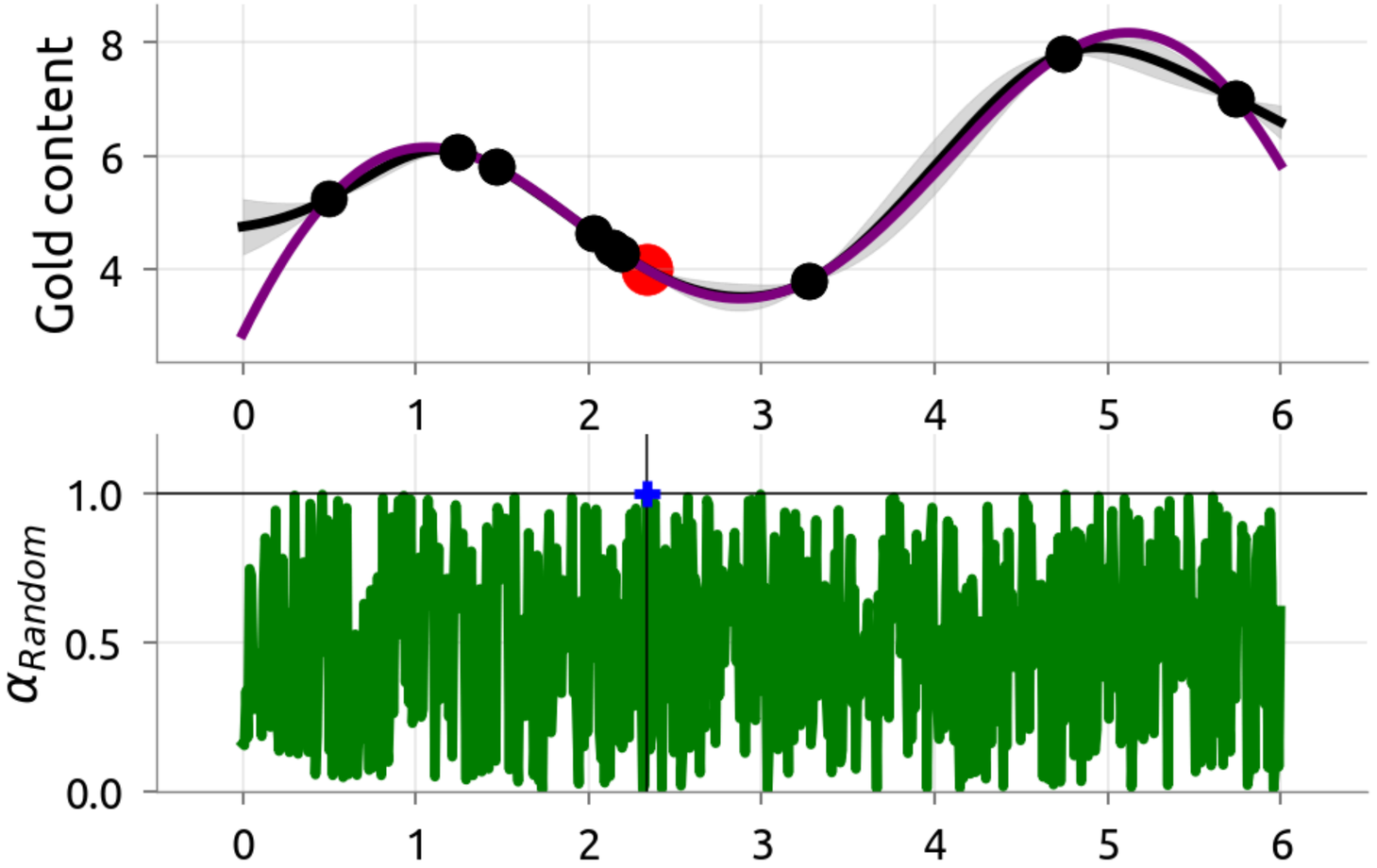
$\mu \pm \sigma$

Training Points

Query Point

Acquisition function

Maxima



Predicted (μ)

Ground Truth (f)

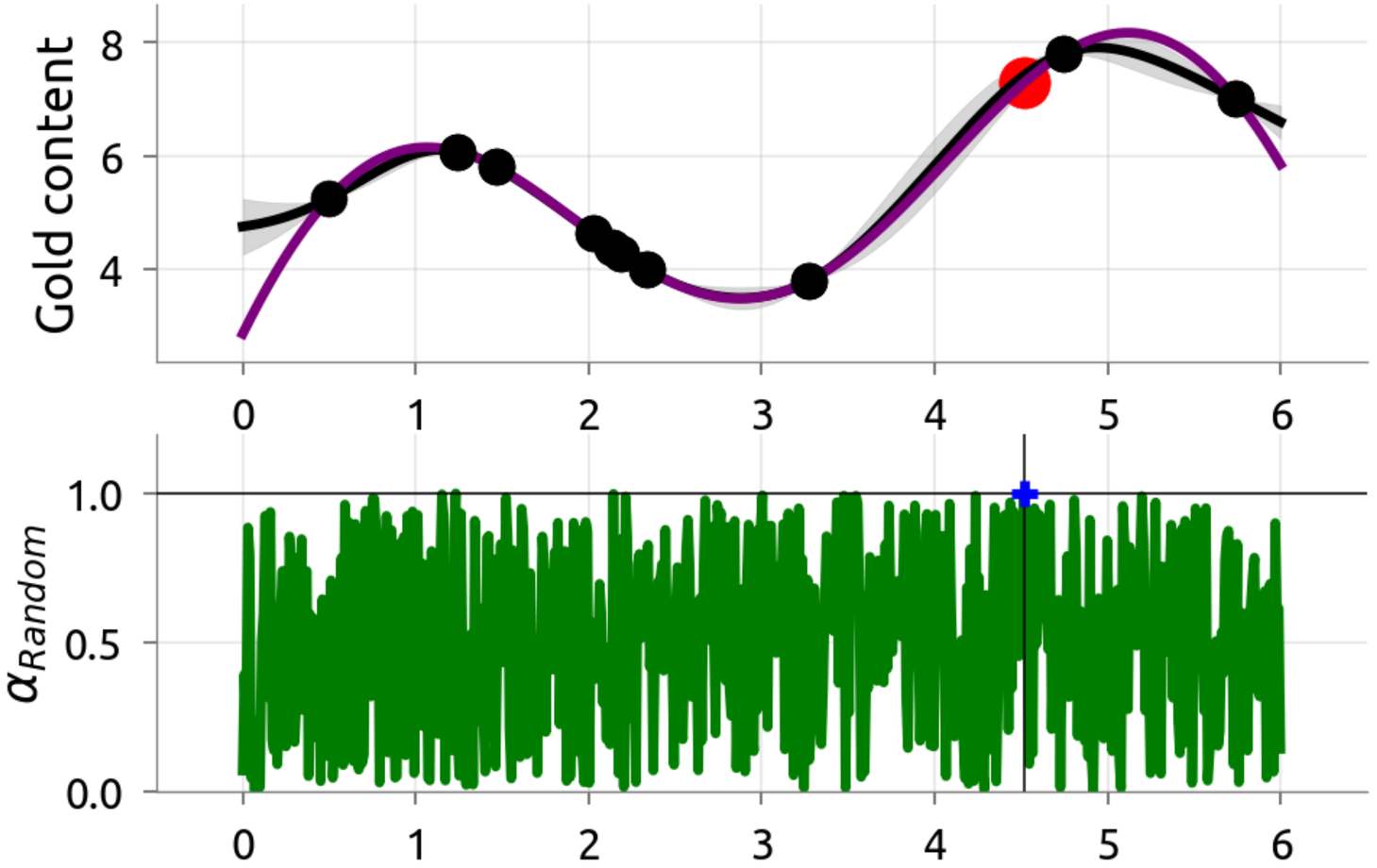
$\mu \pm \sigma$

Training Points

Query Point

Acquisition function

Maxima



Predicted (μ)

Ground Truth (f)

$\mu \pm \sigma$

Training Points

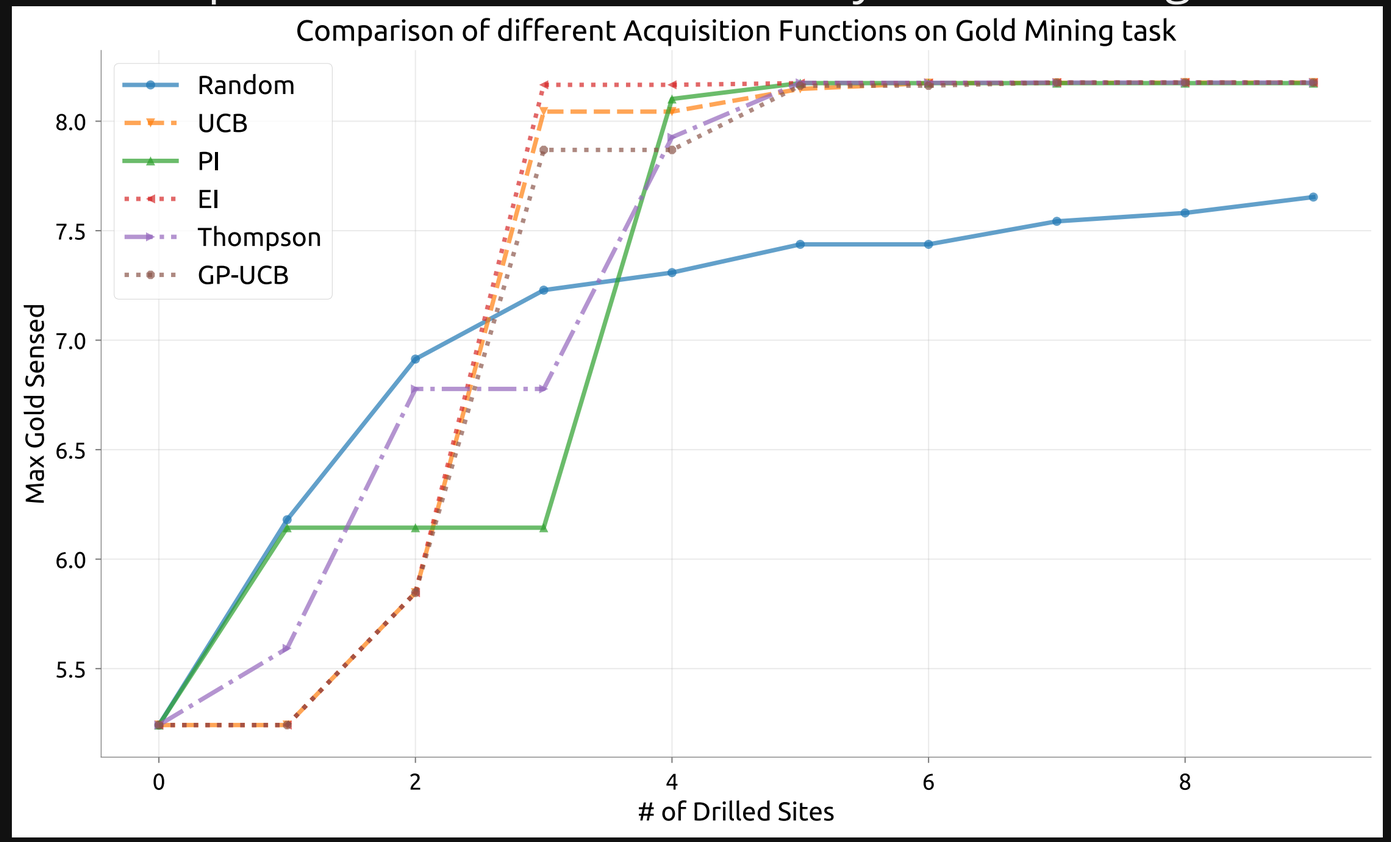
Query Point

Acquisition function

Maxima

Comparison

- random strategy is initially comparable to or better than other acquisition functions
- other acquisition functions can find a good solution in a small number of iterations
- most acquisition functions reach fairly close to the global maxima in as few as three iterations.



Hyperparameters vs Parameters

- **Parameters**: Learned from data during training
- **Hyperparameters**: Set before learning begins

Example: Ridge Regression

- Ridge regression objective:

$$\hat{\theta}_{ridge} = \mathbf{argmin}_{\theta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - x_i^T \theta)^2 + \lambda \sum_{j=1}^{p} \theta_j^2$$

- Parameters: Weight matrix θ
- Hyperparameters:
  - Regularization coefficient $\lambda \geq 0$
  - Learning rate α (if using gradient descent)

Why Bayesian Optimization?

- Common use case: Finding optimal hyperparameters
- Grid search limitations:
  - Impractical for expensive model training
  - Poor scaling with number of hyperparameters
- Bayesian Optimization: Efficient for expensive black-box functions

Conclusion and Summary

- Key Components of Bayesian Optimization:
  - Surrogate function with prior over objective functions
  - Bayesian updating using function evaluations as data
  - Acquisition functions balancing exploration/exploitation
- Ideal Use Cases:
  - Expensive function evaluations
  - When grid/exhaustive search is impractical
  - Hyperparameter optimization for ML models
- Benefits:
  - Sequential optimization is inexpensive
  - Balances exploration and exploitation
  - Efficient for high-dimensional spaces

References

- Distill Article on Bayesian Optimization