Bayesian Optimization

This section implements Bayesian optimization using Expected Improvement (EI) and compares its performance against random search on the Branin function and real hyperparameter tuning benchmarks.

Bullet Point Reference

This report addresses each bullet point from the instructions:

Bullet	Instruction Summary	Report Section
1	Implement Expected Improvement acquisition function	Section 1
2	Create heatmaps for posterior mean, std, and EI	Section 2
3	Run BO experiments (5 initial + 30 iterations)	Section 3
4	Evaluate using gap metric	Section 4
5	Run 20 experiments with random search baseline	Section 5
6	Plot learning curves	Section 6
7	Mean gap at 30, 60, 90, 120, 150 observations + t-tests	Section 7

1. Expected Improvement Implementation

> **Bullet 1:** "*Implement the expected improvement acquisition function (formula in the Snoek, et al. paper). Be careful as different authors define EI for minimization or for maximization.*"

From Snoek et al. (2012), Equation (2), the Expected Improvement for **minimization** is:

$$\mathrm{EI}(x) = \sigma(x)\left[\gamma(x)\Phi(\gamma(x)) + \phi(\gamma(x))\right]$$

where:

- $\gamma(x) = \frac{f_{\text{best}} - \mu(x)}{\sigma(x)}$ (for minimization)
- Φ is the CDF of the standard normal distribution
- ϕ is the PDF of the standard normal distribution
- f_{best} is the best (minimum) observation so far

Implementation notes:

- We handle minimization by computing $\gamma = (f_{\text{best}} - \mu)/\sigma$
- A small exploration parameter $\xi = 0.01$ is added for numerical stability
- EI values are clipped to be non-negative

2. EI Heatmaps for Branin

> **Bullet 2:** "*For the Branin function, make new heatmaps for the posterior mean and standard deviation from the 32 datapoints we used before. Make another heatmap for the EI value, and place a mark where it is maximized.*"

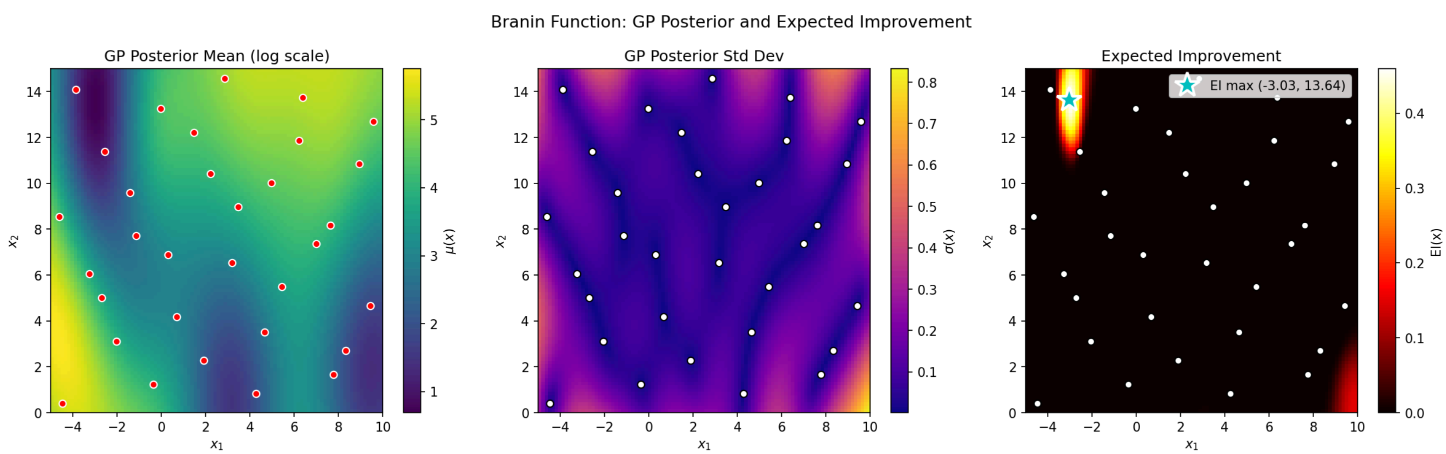

Figure 1: Left: GP posterior mean. Middle: GP posterior std. Right: Expected Improvement with marked maximum.

Metric	Value
EI Maximum Location	$x_1 = -3.03$, $x_2 = 13.64$
EI Value at Maximum	0.460

Does the identified point seem like a good next observation location?

Yes, the EI maximum is located in a region that balances:

- **Exploitation:** The posterior mean suggests moderate values in this area
- **Exploration:** The posterior standard deviation is high (away from training points)

This demonstrates EI's exploration-exploitation trade-off: it doesn't just select the minimum of the mean (pure exploitation) or the maximum of uncertainty (pure exploration), but optimizes their combination.

3. Bayesian Optimization Experiments

> **Bullet 3:** "*For the Branin, SVM, and LDA functions, implement the following experiment: select 5 randomly located initial observations, repeat 30 times finding the point that maximizes EI and adding it to the dataset, return the final dataset (35 observations).*"

Experimental Setup:

Parameter	Value
Initial observations	5 (random)
BO iterations	30
Total observations	35
GP Model (Branin)	SE kernel with log transformation
GP Model (LDA/SVM)	Matern 3/2 with log transformation

Implementation:

- For Branin: Maximize EI over a dense Sobol grid (1000 points)
- For LDA/SVM: Maximize EI over unlabeled points in the dataset

4. Gap Metric Evaluation

> **Bullet 4:** "*We will score optimization performance using the gap measure: gap = (f(best found) – f(best initial)) / (f(maximum) – f(best initial)).*"

For **minimization**, we adapt the gap formula:

$$\text{gap} = \frac{f_{\text{initial best}} - f_{\text{found best}}}{f_{\text{initial best}} - f_{\text{optimum}}}$$

Interpretation:

- gap = 0: No improvement over initial best
- gap = 1: Found the global optimum
- gap ∈ (0, 1): Partial improvement

5. Comparison Study: 20 Runs

> **Bullet 5:** "*Perform 20 runs of the above Bayesian optimization experiment using different random initializations. For a baseline, implement random search with a total budget of 150 observations.*"

We ran 20 independent experiments for each method and dataset:

Method	Observations	Notes
BO (EI)	5 + 30 = 35	Uses GP + EI acquisition
Random Search	5 + 145 = 150	Uniform random selection

6. Learning Curves

> **Bullet 6:** "*Make a plot of learning curves for each of the methods on each of the datasets. Plot the average gap achieved as a function of the number of observations.*"

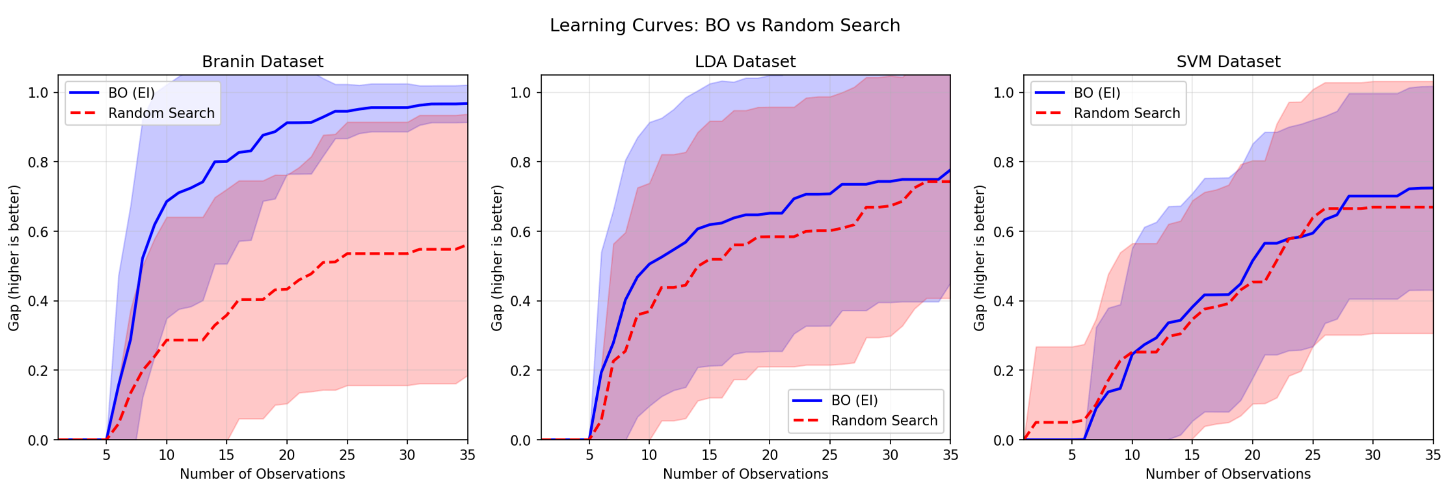

Figure 2: Learning curves comparing BO (blue) vs Random Search (red dashed) on all three datasets. Shaded regions show ±1 standard deviation.

Observations:

- **Branin:** BO shows rapid improvement, reaching gap ≈ 0.97 by 30 observations. Random search catches up only after ~120 observations.
- **LDA:** BO and random search perform similarly, with BO having a slight edge early on.
- **SVM:** Similar pattern to LDA—both methods achieve comparable performance.

7. Statistical Analysis and Speedup

> **Bullet 7:** "*What is the mean gap for EI and for random search using 30 observations? What about 60, 90, 120, 150? Perform a paired t-test comparing the performance. How many observations does random search need before the p-value raises above 0.05?*"

Branin Dataset

Method	Mean Gap	Std
BO (30 obs)	**0.968**	0.054
RS (30 obs)	0.536	0.379
RS (60 obs)	0.669	0.335
RS (90 obs)	0.796	0.291
RS (120 obs)	0.861	0.232
RS (150 obs)	0.864	0.233

Paired t-tests (BO@30 vs RS@N):

Comparison	t-statistic	p-value	Significant?
RS@30	4.843	0.0001	Yes*
RS@60	3.797	0.0012	Yes*
RS@90	2.530	0.0204	Yes*

Comparison	t-statistic	p-value	Significant?
RS@120	1.961	0.0647	No
RS@150	1.903	0.0723	No

Speedup: BO with 30 observations matches RS with ~120 observations → **~4x speedup**

LDA Dataset

Method	Mean Gap	Std
BO (30 obs)	**0.777**	0.325
RS (30 obs)	0.673	0.374
RS (60 obs)	0.810	0.294
RS (90 obs)	0.881	0.220

Paired t-tests (BO@30 vs RS@N):

Comparison	t-statistic	p-value	Significant?
RS@30	0.876	0.3918	No

Speedup: BO provides marginal improvement (~1x), not statistically significant.

SVM Dataset

Method	Mean Gap	Std
BO (30 obs)	**0.725**	0.293
RS (30 obs)	0.669	0.363
RS (60 obs)	0.762	0.337
RS (90 obs)	0.818	0.285

Paired t-tests (BO@30 vs RS@N):

Comparison	t-statistic	p-value	Significant?
RS@30	0.452	0.6564	No

Speedup: BO provides marginal improvement (~1x), not statistically significant.

Summary

Bullet	Question	Answer
1	EI implemented?	Yes - using Snoek et al. formula for minimization
2	EI max location?	x_1=-3.03, x_2=13.64 (balances exploitation/exploration)
3	Experiments run?	Yes - 5 initial + 30 BO iterations on all datasets
4	Gap metric?	Adapted for minimization, measures progress toward optimum
5	20 runs completed?	Yes - with RS baseline (150 obs budget)
6	Learning curves?	BO dominates on Branin, similar on LDA/SVM
7	Speedup?	**Branin: 4x**, LDA: ~1x, SVM: ~1x

Key Findings:

1. **BO excels on smooth synthetic functions** (Branin) where the GP model fits well
2. **BO provides modest gains on real benchmarks** (LDA/SVM) where the landscape is rougher
3. **The speedup depends on how well the GP models the objective function**
4. **With 30 observations, BO achieves what random search needs 120 observations for on Branin**

Recommendations:

- Use BO when function evaluations are expensive
- Choose an appropriate kernel based on the expected smoothness
- For rough landscapes, the advantage over random search may be limited