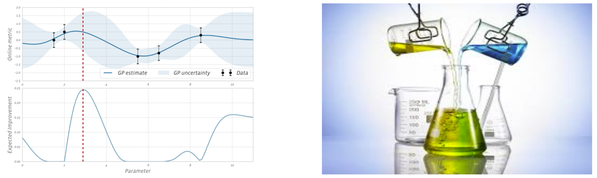
Sparse Bayesian Optimization

Sulin Liu*[,1], Qing Feng*[,2], David Eriksson*[,2], Benjamin Letham[2], Eytan Bakshy[2]

1. Princeton University 2. Meta *equal contribution

Sparsity in Bayesian Optimization Solutions



Bayesian optimization uses Gaussian processes to learn a surrogate model of the black-box objective (for example, maximizing chemistry reaction yield), and perform sequential optimization based on exploitation-exploration trade-off.

Bayesian optimization is used to optimize the objective of complex systems. Sometimes we want **sparsity in solutions**. For example, in chemistry reaction, a sparse solution may require fewer agents and steps to synthesize a compound.

- Sparsity means **Interpretability**
  - helps understand the complex systems
- Sparsity means **Simplicity**
  - easier to deploy and maintain

Regularization in Bayesian Optimization

Black box function modeled by Gaussian processes posterior distribution:

$$f(\mathbf{x} \mid \mathcal{D}) \sim \mathcal{N}(\mu(\mathbf{x}), \sigma^2(\mathbf{x}))$$

A **penalty** term to target feature-level sparsity:

$$\xi(\mathbf{x}) = \|\mathbf{x} - \mathbf{x}^s\|_0$$

A user-defined point to regularize towards

Sparsity-inducing norm

Acquisition function with **External Regularization** (ER):

$$\alpha_{\text{ER}}(\mathbf{x}; \lambda) = \mathbb{E}_f[\text{utility}(f(\mathbf{x}))] - \lambda\xi(\mathbf{x})$$

Acquisition function with **Internal Regularization** (IR):

$$\alpha_{\text{IR}}(\mathbf{x}; \lambda) = \mathbb{E}_f[\text{utility}(f(\mathbf{x}) - \lambda\xi(\mathbf{x}))]$$

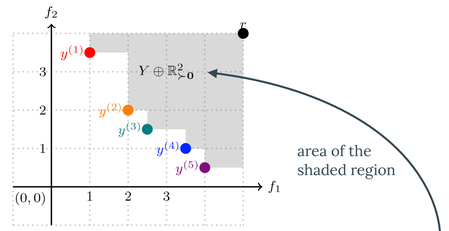
- ER and IR act differently for different acquisitor functions:
  - ER and IR are effectively different for Expected Improvement
  - ER and IR are identical for Upper Confidence Bound

Limitations of Using Regularization

- **Proposition 1 (negative result for ER, informal)**
  - If acquisition function is 0 for every $\mathbf{x}$ where $\mathbf{x}$'s sparsity $\xi(\mathbf{x})$ is less than k, then for any λ, every maximizer will either have sparsity greater than k or equal to $\mathbf{x}^s$.
- **Proposition 2 (negative result for IR, informal)**
  - Consider $h(k) = \max_{\mathbf{x} \in \mathcal{B}} f(\mathbf{x})$ subject to $\xi(\mathbf{x}) = k$, if $h(k)$ is strictly convex over some interval across $\hat{k}$, then there is no maximizer of IR with $\xi(\mathbf{x}) = \hat{k}$ for any λ.

Sparsity Exploring Bayesian Optimization (SEBO)

- **Multi objective approach to optimize sparsity and objective**
  - Consider both objective f and sparsity ξ as objectives

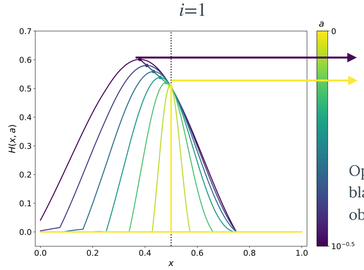


- Use Expected Hypervolume Improvement (EHVI) as acquisition function to effectively optimize multiple objectives simultaneously

$$\alpha_{\text{SEBO}}(\mathbf{x}) = \mathbb{E}_f\left[V(X^{\text{obs}} \cup \{\mathbf{x}\}) - V(X^{\text{obs}})\right]$$

- Directly optimize L_0 sparsity
  - Use the idea of homotopy, a from $a_{\text{start}} \to 0$

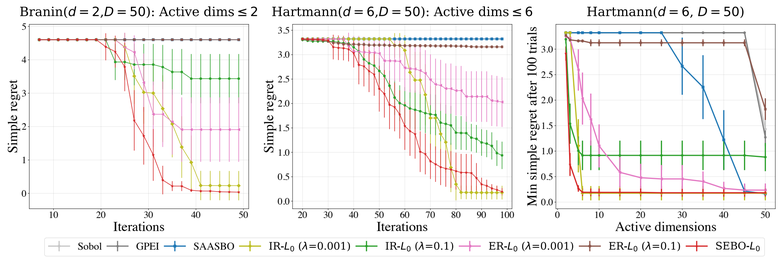
$$\varphi_a(\mathbf{x}) := D - \sum_{i=1}^{D} \exp\left(-0.5\,(\mathbf{x}_i/a)^2\right) \approx \|\mathbf{x}\|_0$$



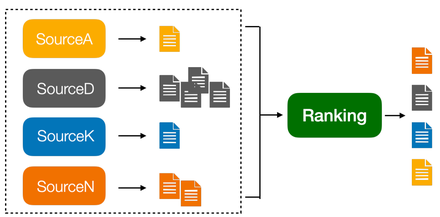
Optimizing a toy problem:
black box function $f(x) = -x^2$ and $\xi(\mathbf{x}) = \|x - 0.5\|_0$
observations $X^{\text{obs}} = \{0, 0.25, 0.75, 1.0\}$

Experiments

- **Synthetic experiments**
  - Low-dimensional synthetic functions embedded in a 50D space.
  - SAASBO: a high-dimensional BO method that uses a GP to learn to ignore unimportant parameters, but without optimizing sparsity in solutions
  - SEBO is the most efficient in identifying optimal sparse solutions



- **Real-world experiments**
  - Sourcing component of a recommendation system
    - Retrieve items using a weighted combination of multiple sources



  - A sparse policy can drop redundant sources or low-quality sources, hence increasing interpretability and reducing "tech debt"