Bayesian Optimization

- Bayesian optimization (BO) is an effective and popular approach for global optimization of black-box functions [2].
- Using BO we want to find an input \( x \in \mathcal{X} \) that maximizes the true-valued black-box function \( f : \mathcal{X} \to \mathbb{R} \) defined on a compact domain \( \mathcal{X} \subseteq \mathbb{R}^D \).
- The factorization of the log-marginal likelihood degenerates the full covariance matrix
- The Generalized Product of Expert (gPoE) model combines each individual GP expert prediction into a final aggregate model

\[ f(k_{x, x}) = \text{argmax}_{k} f(x) \]

given noisy observations \( y = N(f(x), \sigma_f^2) \) with noise variance \( \sigma_f^2 \).
- Build probabilistic surrogate model based on observations \( D_n = \{(x_i, y_i)\}_{i=1}^{n} \).
- Find the next candidate point \( x_{n+1} \) which maximizes the acquisition function \( \phi, X \to \mathbb{R} \)

\[ x_{n+1} = \text{argmax}_x \phi(x) \in \{(x_i)\}_{i=1}^{n} \]

Probabilistic Surrogate Models

Gaussian process

- A Gaussian process \( GP(\mu, \sigma^2) \) is fully specified by a mean function \( \mu(x) \) and a covariance function \( k(x, .) \) [3].
- The objective is to infer the latent function \( f \) from a training set \( (x_i, y_i) \) where \( X = \{x_i\}_{i=1}^{m}, y = \{y_i\}_{i=1}^{m} \).
- GP posterior predictive distribution at a test point \( \mu, y \sim N(\mu(x), \sigma_f^2) \) is Gaussian with the mean and variance given by

\[ \mu_f = k_{x, x}^T k_x^{-1} y, \quad \sigma_f^2 = k_{x, x} - k_{x, x} k_x^{-1} k_{x, x}^T \]

The main challenge of GP is that training requires the inversion and the determinant of \( K_n = \sigma_f^2 I \), which is frequently realised via the Cholesky decomposition with computational cost of \( O(n^3) \). For this reason, training GP on large datasets is computationally intractable.

Generalized Product Of Experts

- Partitions the data into \( M \) subsets \( D_i = \{(x_i, y_i)\}_{i=1}^{n_i} \), where \( 1 \leq i \leq M \), and train GP on \( D_i \) as an expert GP model [1].
- Predictive distribution of GP expert \( i \) conditioned on the related subset of the data \( D_i \) and test input \( x \), \( X^D_i \) is Gaussian \( \mu_f(x_i, \sigma_f^2(x_i)) \) with mean and covariance

\[ \mu_f(x) = k_{x, x} k_x^{-1} y, \quad \sigma_f^2(x) = k_{x, x} - k_{x, x} k_x^{-1} k_{x, x}^T \]

- The Generalized Product Of Expert (gPoE) model combines each individual GP expert prediction into the final aggregate model

\[ \mu_f(x, D) = \sum_{i=1}^{M} p(D_i)(x) \mu_f(x, D_i), \]

which is again Gaussian \( \mu_f(\mu_f, \sigma_f^2) \) with mean and covariance given by

\[ \mu_f^2(x) = \sum_{i=1}^{M} p(D_i)(x) \mu_f^2(x, D_i), \quad \sigma_f^2(x) = \sum_{i=1}^{M} p(D_i)(x) \sigma_f^2(x, D_i) \]

The weight \( p(D_i)(x) \) is a measure of reliability and controls the contribution of each expert \( f_i \) at test point \( x \), where \( \sum_{i=1}^{M} p(D_i)(x) = 1 \).
- The factorization of the log-marginal likelihood degenerates the full covariance matrix \( K_n = k(X, X) \) into block-diagonal matrix:

\[ K_n = \text{diag}(K_{11}, . . . , K_{MM}) \]

Trust region Bayesian optimization with Generalized PoE

Algorithm 1: Generalized PoE based trust region Bayesian optimization (gPoETRBO)

Input: Number of initializing points \( N \), iterations \( T \), points per expert \( \alpha \), initial TR parameters.
Output: The best recommendation \( x^* \in \mathcal{X} \)

1. Randomly select and evaluate \( N \) points in the search space \( D_n = \{(x_i, f(x_i))\}_{i=1}^{n} \).
2. for \( i = 1 \) to \( T \) do
   3. Randomly partition \( D_n \) into \( M \) subsets.
   4. Train \( M \) local GP experts on \( (D_i) \) subsets.
   5. Construct TR of length \( d \) around the best point \( x^* = \text{argmax}_{x \in D} f(x) \).
   6. Generate \( \alpha \) candidate points \( X = \{x_i\}_{i=1}^{\alpha} \) from TR(\( x^* \)).
   7. Evaluate \( \alpha \) local GP expert posterior mean \( \mu_f(x) \) and variance \( \sigma_f^2(x) \) on \( X \) points.
   8. Aggregate \( \alpha \) and \( \sigma_f^2 \) using (6) and (7).
   9. Maximize UCB acquisition function \( \tilde{\phi}(x) = \text{argmax}_{x \in X} \mu_f(x) + \sqrt{2 \sigma_f^2(x)} \)
   10. Evaluate the objective function \( g = f(x) \).
   11. Add a new data point to the dataset \( D_{n+1} = D_n \cup \{x, g\} \).
   12. Update the TR parameters and check whether to restart.

Numerical experiments

Ablation study

References