Bayesian Optimization with Informative Covariance

Afonso Eduardo
afonso.eduardo@ed.ac.uk

School of Informatics
University of Edinburgh

June 2022

Background on Bayesian Optimization

Goal: Find global minimizer of $f : \mathbb{X} \to \mathbb{Y} \subseteq \mathbb{R}$,
(unknown, expensive to evaluate)

$$\boldsymbol{x}^\star = \arg\min_{\boldsymbol{x} \in \mathbb{X}} f(\boldsymbol{x})$$

Algorithm 1 Bayesian Optimization (BO)

Input: objective f and acquisition α functions, surrogate model $\mathcal{M}$, initial evidence set $\mathcal{D}^{(n_0)}$

repeat

$\boldsymbol{x}_{n+1} = \arg\max \alpha(\boldsymbol{x} \mid \mathcal{D}_n, \mathcal{M})$ ▷ Find best candidate

$y_{n+1} = f(\boldsymbol{x}_{n+1})$ ▷ Evaluate candidate

$\mathcal{D}_{n+1} = \mathcal{D}_n \cup \{(\boldsymbol{x}_{n+1}, y_{n+1})\}$ ▷ Update evidence set

until stopping condition is met

Background on Bayesian Optimization

Surrogate Model: Gaussian Process (GP) Regression
Prior on functions $f \sim GP(m_{\boldsymbol{\theta}}, C_{\boldsymbol{\theta}})$

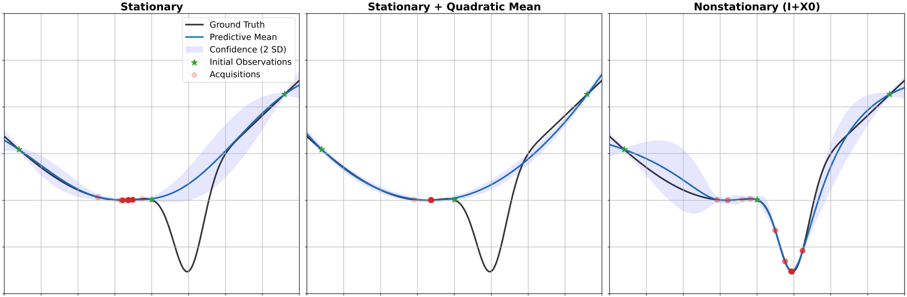
- Mean function m, Covariance function C, (Hyper)parameters $\boldsymbol{\theta}$
- Train on $\mathcal{D}_n = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$
- (Univariate) Posterior predictive distribution $\mathcal{N}(m_n(\boldsymbol{x}), v_n(\boldsymbol{x}))$

Background on Bayesian Optimization

Surrogate Model: Gaussian Process (GP) Regression

Prior on functions $f \sim GP(m_{\boldsymbol{\theta}}, C_{\boldsymbol{\theta}})$

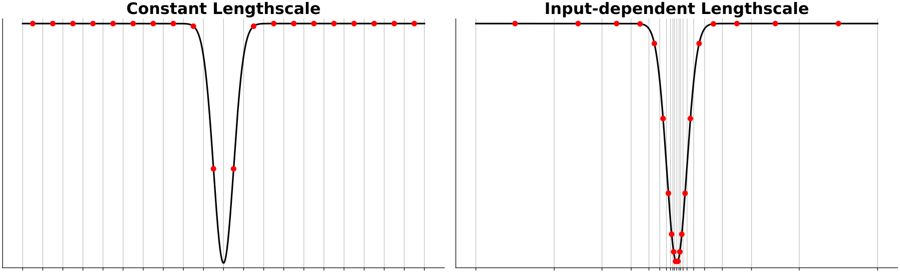
- Mean function m, Covariance function C, (Hyper)parameters $\boldsymbol{\theta}$
- Train on $\mathcal{D}_n = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$
- (Univariate) Posterior predictive distribution $\mathcal{N}(m_n(\boldsymbol{x}), v_n(\boldsymbol{x}))$



Benefits of Nonstationarity

More efficient representations via spatially-varying lengthscales

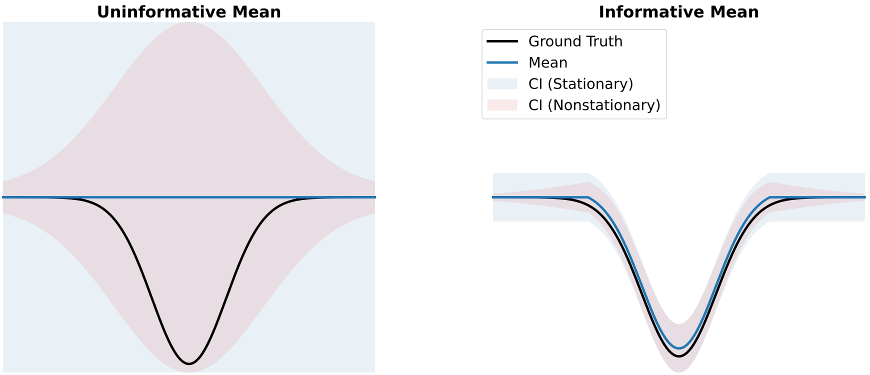
- How to partition the search space? Shorter lengthscales where objective varies rapidly, but longer lengthscales elsewhere.
  $\rightarrow$ Heterogeneous exploration



Benefits of Nonstationarity

Better worst-case optimization performance via spatially-varying prior variance

- Instantaneous regret $r_{n+1} = f(\boldsymbol{x}_{n+1}) - f(\boldsymbol{x}^{\star})$
- For popular acq functions (LCB, EI), $\max r_{n+1} \propto \sqrt{v_{n+1}(\boldsymbol{x}_{n+1})}$
- Tighter bounds lead to lower worst-case regret



Benefits of Nonstationarity

No practical finite-time convergence guarantees with stationary covar functions

- Impossibility of exploring entire high-dimensional spaces.
- Unless budget increases exponentially, no global reduction in uncertainty $\rightarrow$ constant worst-case regret.

Benefits of Nonstationarity

No practical finite-time convergence guarantees with stationary covar functions

- Impossibility of exploring entire high-dimensional spaces.
- Unless budget increases exponentially, no global reduction in uncertainty $\rightarrow$ constant worst-case regret.
  - Predictive uncertainty = prior uncertainty − uncertainty explained by observations, $v_n(\boldsymbol{x}) = C_{\boldsymbol{\theta}^\star}(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{c}_n(\boldsymbol{x})^\intercal \mathbf{C}_n^{-1} \boldsymbol{c}_n(\boldsymbol{x})$
  - Stationary covar $v_n(\boldsymbol{x}) = \sigma_0^2 - \boldsymbol{c}_n(\boldsymbol{x})^\intercal \mathbf{C}_n^{-1} \boldsymbol{c}_n(\boldsymbol{x})$
  - Observations not close enough to $\boldsymbol{x} \rightarrow \boldsymbol{c}_n(\boldsymbol{x}) \approx 0$, $v_n(\boldsymbol{x}) \approx \sigma_0^2$

Benefits of Nonstationarity

No practical finite-time convergence guarantees with stationary covar functions

- Impossibility of exploring entire high-dimensional spaces.
- Unless budget increases exponentially, no global reduction in uncertainty $\rightarrow$ constant worst-case regret.
  - Predictive uncertainty = prior uncertainty − uncertainty explained by observations, $v_n(\boldsymbol{x}) = C_{\boldsymbol{\theta}^\star}(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{c}_n(\boldsymbol{x})^\intercal \mathbf{C}_n^{-1} \boldsymbol{c}_n(\boldsymbol{x})$
  - Stationary covar $v_n(\boldsymbol{x}) = \sigma_0^2 - \boldsymbol{c}_n(\boldsymbol{x})^\intercal \mathbf{C}_n^{-1} \boldsymbol{c}_n(\boldsymbol{x})$
  - Observations not close enough to $\boldsymbol{x} \rightarrow \boldsymbol{c}_n(\boldsymbol{x}) \approx 0$, $v_n(\boldsymbol{x}) \approx \sigma_0^2$

Nonstationary covar functions are spatially informative

- Predictive variance $v_{n-1}(\boldsymbol{x}_n)$ depends on $\boldsymbol{x}_n$ even when $\boldsymbol{c}_{n-1}(\boldsymbol{x}_n) \approx 0$
- **Informative models**: some regions more informative $\rightarrow$ increased efficiency if beliefs correct to some degree.

Informative Covariance Functions

Promote exploration of regions deemed more promising according to beliefs where the optimum might be, $\boldsymbol{x}_0 \sim p(\boldsymbol{x}^\star) \propto \phi(\boldsymbol{x}^\star)$,

$$\phi(\boldsymbol{x}^\star) = 1 + \frac{1}{L} \sum_{l \leq L} (w_l - 1)\, k_l \left(d_l(\boldsymbol{x}^\star, \boldsymbol{x}_0^{(l)}) \right)$$

- Set of anchor points $\{\boldsymbol{x}_0^{(l)}\}$.
- Positive weights w_l.
- Distance functions d_l and kernels k_l characterize neighborhoods.
- Uninformative slab ensures optimum is included in the support (bounded search space).

Use ϕ to induce spatially-varying prior (co)variance and lengthscales.

Informative Covariance Functions

Spatially-varying prior covariance

$$C_{\mathrm{NS}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma_0^2(\boldsymbol{x}_i, \boldsymbol{x}_j) C_{\mathrm{S}}(\boldsymbol{x}_i, \boldsymbol{x}_j), \qquad \sigma_0^2(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma_p^2 \sqrt{\phi(\boldsymbol{x}_i)} \sqrt{\phi(\boldsymbol{x}_j)},$$

- $\sigma_0^2(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is symmetric and separable $\rightarrow$ valid covariance function, i.e., symmetric positive-definite function.
- Product of two covariance functions is a covariance function.

Informative Covariance Functions

Spatially-varying prior covariance

$$C_{\mathrm{NS}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma_0^2(\boldsymbol{x}_i, \boldsymbol{x}_j) C_{\mathrm{S}}(\boldsymbol{x}_i, \boldsymbol{x}_j), \qquad \sigma_0^2(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma_p^2 \sqrt{\phi(\boldsymbol{x}_i)} \sqrt{\phi(\boldsymbol{x}_j)},$$

- $\sigma_0^2(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is symmetric and separable $\rightarrow$ valid covariance function, i.e., symmetric positive-definite function.
- Product of two covariance functions is a covariance function.
- **Intuition**
  - Covariance functions compute $\mathrm{cov}(f(\boldsymbol{x}_i), f(\boldsymbol{x}_j))$.

Informative Covariance Functions

Spatially-varying prior covariance

$$C_{\mathrm{NS}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma_0^2(\boldsymbol{x}_i, \boldsymbol{x}_j) C_{\mathrm{S}}(\boldsymbol{x}_i, \boldsymbol{x}_j), \qquad \sigma_0^2(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma_p^2 \sqrt{\phi(\boldsymbol{x}_i)} \sqrt{\phi(\boldsymbol{x}_j)},$$

- $\sigma_0^2(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is symmetric and separable $\rightarrow$ valid covariance function, i.e., symmetric positive-definite function.
- Product of two covariance functions is a covariance function.
- **Intuition**
  - Covariance functions compute $\mathrm{cov}(f(\boldsymbol{x}_i), f(\boldsymbol{x}_j))$.
  - Higher probability under $p(\boldsymbol{x}^\star) \rightarrow$ Larger $\sigma_0^2(\boldsymbol{x}, \boldsymbol{x}) \rightarrow +$ Informative
  - For 2 points with high probability, both values should be small and highly correlated.

Informative Covariance Functions

Spatially-varying prior covariance

$$C_{\mathrm{NS}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma_0^2(\boldsymbol{x}_i, \boldsymbol{x}_j) C_{\mathrm{S}}(\boldsymbol{x}_i, \boldsymbol{x}_j), \qquad \sigma_0^2(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma_p^2 \sqrt{\phi(\boldsymbol{x}_i)} \sqrt{\phi(\boldsymbol{x}_j)},$$

- $\sigma_0^2(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is symmetric and separable $\rightarrow$ valid covariance function, i.e., symmetric positive-definite function.
- Product of two covariance functions is a covariance function.
- **Intuition**
  - Covariance functions compute $\mathrm{cov}(f(\boldsymbol{x}_i), f(\boldsymbol{x}_j))$.
  - Higher probability under $p(\boldsymbol{x}^\star)$ $\rightarrow$ Larger $\sigma_0^2(\boldsymbol{x}, \boldsymbol{x})$ $\rightarrow$ + Informative
  - For 2 points with high probability, both values should be small and highly correlated.
  - As probability decreases for one point $\boldsymbol{x}_j$, we believe $f(\boldsymbol{x}_j)$ to be less constrained, and less correlated with a small $f(\boldsymbol{x}_i)$.

Informative Covariance Functions

Spatially-varying lengthscales

Without loss of generality, possible to rewrite as

$$C_{\mathrm{NS}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma_0^2(\boldsymbol{x}_i, \boldsymbol{x}_j) C_{\mathrm{S}}(h_\lambda(\boldsymbol{x}_i), h_\lambda(\boldsymbol{x}_j)),$$

- h_λ is an input-warping function.

Informative Covariance Functions

Spatially-varying lengthscales

Without loss of generality, possible to rewrite as

$$C_{\mathrm{NS}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma_0^2(\boldsymbol{x}_i, \boldsymbol{x}_j) C_{\mathrm{S}}(h_\lambda(\boldsymbol{x}_i), h_\lambda(\boldsymbol{x}_j)),$$

- h_λ is an input-warping function.
- Set h_λ to a nonlinear transformation that shrinks the lengthscales locally around anchors.
- **Intuition:** Finer detail in promising regions (expansion), coarser scale (contraction) otherwise.

Experiments: Main Methods

Baselines:

- **S**: BO with GP model specified by an uninformative constant prior mean and a stationary covariance function.
- **S+QM**: S with an axis-aligned quadratic prior mean function.

Experiments: Main Methods

Baselines:

- **S**: BO with GP model specified by an uninformative constant prior mean and a stationary covariance function.
- **S+QM**: S with an axis-aligned quadratic prior mean function.
- **C**: BO with a GP model specified by a constant prior mean and a cylindrical covariance function.

Experiments: Main Methods

Baselines:

- **S**: BO with GP model specified by an uninformative constant prior mean and a stationary covariance function.
- **S+QM**: S with an axis-aligned quadratic prior mean function.
- **C**: BO with a GP model specified by a constant prior mean and a cylindrical covariance function.
  - Transformation maps balls of radius R onto the surface of a cylinder of height R.
  - Center expansion, boundary contraction (Euclidean space).
  - Belief that optimal values are near the center.

Experiments: Main Methods

Baselines:

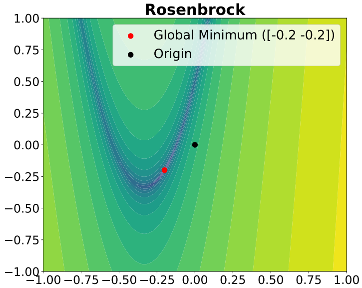
- **S**: BO with GP model specified by an uninformative constant prior mean and a stationary covariance function.
- **S+QM**: `S` with an axis-aligned quadratic prior mean function.
- **C**: BO with a GP model specified by a constant prior mean and a cylindrical covariance function.
  - Transformation maps balls of radius R onto the surface of a cylinder of height R.
  - Center expansion, boundary contraction (Euclidean space).
  - Belief that optimal values are near the center.

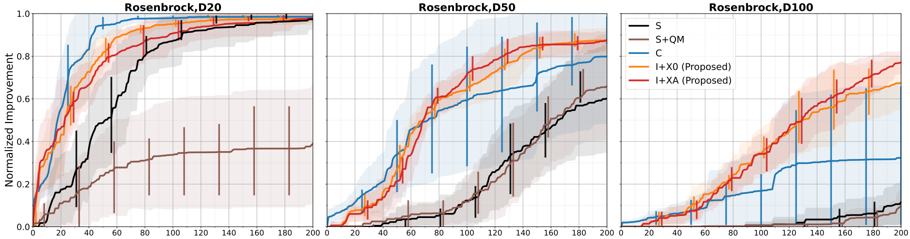
Proposed:

- **I+X0**: BO with a GP model specified by a constant prior mean and informative covariance. Single fixed anchor at the center.
- **I+XA**: Anchor in `I+X0` set to incumbent solution (adaptive greedy).

Characterization:

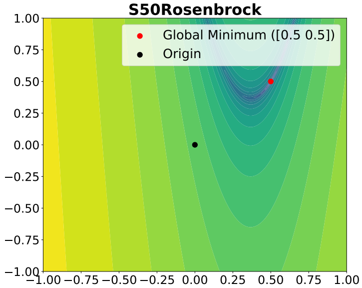
- Bowl-shaped objective.
- Narrow banana-shaped valleys.
- Optimum relatively close to center.

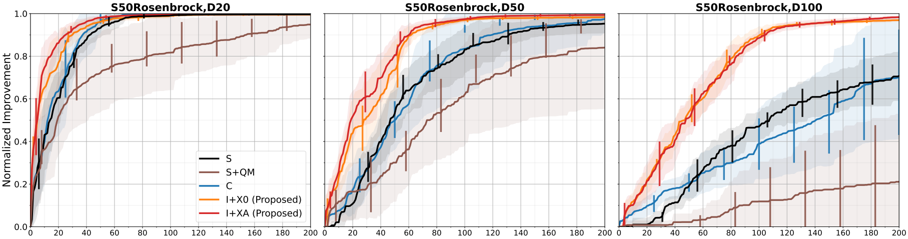




Characterization:

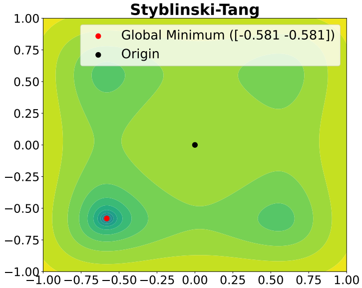
- Bowl-shaped objective.
- Narrow banana-shaped valleys.
- Optimum further away from the center.

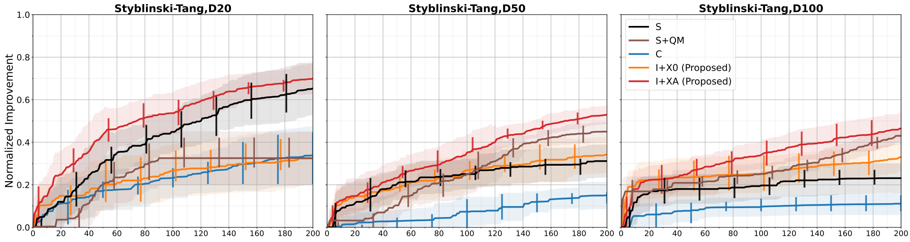




Characterization:

- Roughly bowl-shaped objective.
- Center is a local maximum.
- Exponentially many local modes.
- Optimum relatively far from center.

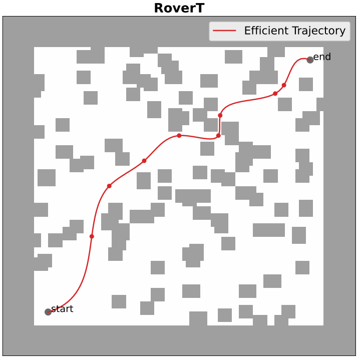


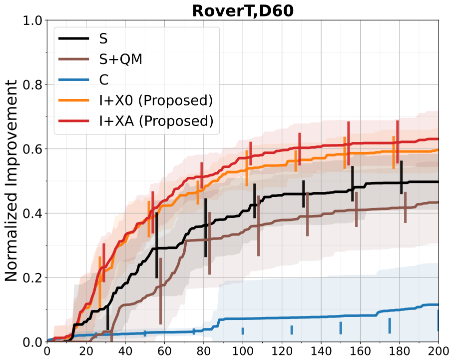


Experiments: Rover Trajectory

Goal: Optimize 2D trajectory of a rover.

- Trajectory given by a spline, fitted to 30 2-dimensional points (60D).





Conclusion

- Analysis of the benefits of nonstationarity for BO.
- Informative covariance functions for GP-based BO, leveraging nonstationarity to express input-dependent information.
  - Information about the optimum induces spatially-varying prior covariance and lengthscales $\rightarrow$ promote exploration of promising regions.

Conclusion

- Analysis of the benefits of nonstationarity for BO.
- Informative covariance functions for GP-based BO, leveraging nonstationarity to express input-dependent information.
  - Information about the optimum induces spatially-varying prior covariance and lengthscales $\rightarrow$ promote exploration of promising regions.
- High-dimensional Experiments
  - Challenge the use of stationarity and informative mean functions.
  - Proposed methodology can lead to significant increase in performance, even under weak prior information (`I+XA`).

Experiments: Rover Trajectory

Objective does not penalize distance (less efficient trajectories)

- Rover is free to roam anywhere, as long as it satisfies target endpoints and avoids collisions.

Example trajectories