Collaborative Bayesian Optimization via Wasserstein Barycenters

Donglin Zhan, Haoting Zhang, Rhonda Righter,
Zeyu Zheng, and James Anderson

Columbia University & UC Berkeley

64th IEEE Conference on Decision and Control, December 2025

Motivation: Collaborative Optimization with Privacy

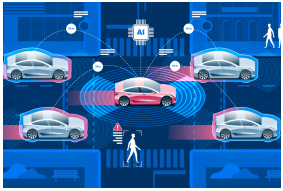
Context:

- **Black-box Optimization:** Maximizing $f(x)$ where evaluations are costly/opaque (e.g., hyperparameter tuning, materials design).
- **Collaborative Setting:** N agents solve the same problem to accelerate learning.

Applications:

- **Autonomous Vehicle Platooning:** Vehicles optimize control parameters but cannot share raw sensor logs.
- **Grid Optimization:** Units optimize operations without exchanging sensitive consumption data.

The Challenge: How to enable collaboration when agents can share *models* (GPs) but **not** raw data (x, y)?



Why Bayesian Optimization

1. Objective f is "expensive" to evaluate
   - time to compute $f(x)$
   - monetary cost associated with each evaluation
   - safety may limit the number of evaluations
   - humans may only tolerate so many questions

Why Bayesian Optimization

1. Objective f is "expensive" to evaluate
   - time to compute $f(x)$
   - monetary cost associated with each evaluation
   - safety may limit the number of evaluations
   - humans may only tolerate so many questions
2. Structure of f
   - f should be continuous so GP can be used for approximation
   - "black box" optimization – will not take advantage of structure in f

Why Bayesian Optimization

1. Objective f is "expensive" to evaluate
   - time to compute $f(x)$
   - monetary cost associated with each evaluation
   - safety may limit the number of evaluations
   - humans may only tolerate so many questions
2. Structure of f
   - f should be continuous so GP can be used for approximation
   - "black box" optimization – will not take advantage of structure in f
3. Derivative-free
   - observe $f(x)$, no access to $\nabla_x f(x)$, $\nabla_x^2 f(x)$, ...

Problem Formulation

Bayesian Optimization:

$$\max_{x \in \mathcal{X}} f(x)$$

where $\mathcal{X} \subset \mathbb{R}^d$ is compact.

Observation Model:

- Noisy observations: $y_t = f(x_t) + \epsilon_t$, with $\epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2)$.
- **Surrogate Model:** Gaussian Process (GP) $f(x) \sim \mathcal{GP}(0, K(x, x'))$.
- Common Kernel: Radial Basis Function (RBF) or similar.

Problem Formulation

Bayesian Optimization:

$$\max_{x \in \mathcal{X}} f(x)$$

where $\mathcal{X} \subset \mathbb{R}^d$ is compact.

Observation Model:

- Noisy observations: $y_t = f(x_t) + \epsilon_t$, with $\epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2)$.
- **Surrogate Model:** Gaussian Process (GP) $f(x) \sim \mathcal{GP}(0, K(x, x'))$.
- Common Kernel: Radial Basis Function (RBF) or similar.

The Privacy Gap in Literature:

- *Standard BO:* Centralized data aggregation.
- **Our Solution:** Aggregate GPs via **Wasserstein Barycenters** to form a central prior without data access.

Bayesian Optimization: The Surrogate Model

BO uses a probabilistic surrogate (Gaussian Process) to model the objective $f(x)$. Given data $\mathcal{D}_t = \{(x_i, y_i)\}_{i=1}^t$, we assume $y_i = f(x_i) + \epsilon_i$, where $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$.

Posterior Mean

$$\mu_t(x) = K_t(x)^\top (K_{tt} + \sigma_\epsilon^2 I)^{-1} y_t$$

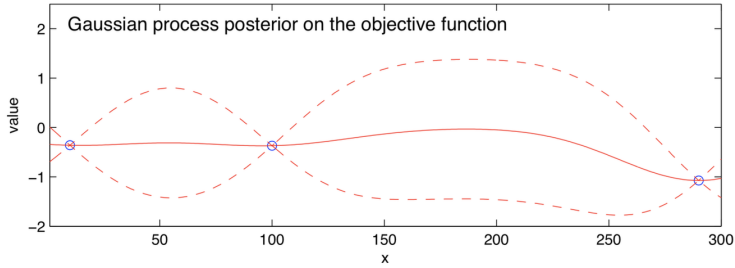
where $K_t(x) = [k(x, x_1), \ldots, k(x, x_t)]^\top$, and K_{tt} is the kernel matrix with $(\tau, \tau\prime)$-th entry $K(\tau, \tau\prime)$.

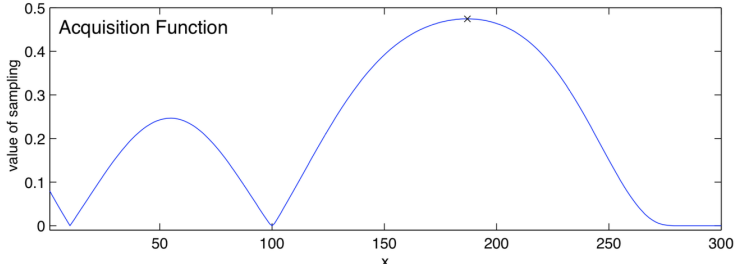
- The "best guess" of the function value.
- Used for **Exploitation** (sampling where we think the value is high).

Posterior Variance

$$k_t(x, x') = k(x, x') - K_t(x)^\top (K_{tt} + \sigma_\epsilon^2 I)^{-1} K_t(x')$$

- Measures uncertainty.
- Used for **Exploration** (sampling where we know little).





[**A Tutorial on Bayesian Optimization**, Frazier, 2018]

Acquisition Functions

Selecting the next query point x_{t+1}

maximize an Acquisition Function $\alpha(x)$ that balances exploration and exploitation

Common Examples:

- **Upper Confidence Bound (UCB):**

$$\alpha_{UCB}(x) = \mu_t(x) + \beta\sigma_t(x)$$

- **Expected Improvement (EI):** Measures expected gain over the current best observation f^*.

$$\alpha_{EI}(x) = \mathbb{E}[\max\{f(x) - f^*, 0\}]$$

Limitation: EI is often myopic (one-step lookahead) and struggles with noisy data.

Knowledge Gradient (KG)

We choose x to maximize the expected increase in the maximum of the posterior mean *after* observing y at x.

$$\alpha_{KG}(x) = \mathbb{E}_y \left[\max_{x' \in \mathcal{X}} \mu_{t+1}(x') \,|\, x \right] - \max_{x' \in \mathcal{X}} \mu_t(x')$$

- KG measures improvement of the final best decision

Why KG for Collaboration?

- **Noise Robustness:** Handles noisy observations better than EI.
- **Batch Selection (q-KG):** Naturally extends to selecting multiple points (agents) simultaneously.

Collaborative BO Framework

Protocol per Iteration t:

1. **Local Update:** Each agent n updates local posterior based on private history $\tilde{\mathcal{S}}_n$:

$$\tilde{f}_n \sim \mathcal{GP}(\tilde{\mu}_n, \tilde{K}_n)$$

2. **Communication:** Agents send posterior parameters ($\tilde{\mu}_n, \tilde{K}_n$ on discretized grid) to server.
3. **Aggregation:** Server computes Central GP f^c via Wasserstein Barycenter.
4. **Action Selection:** Server maximizes **Collaborative Acquisition Function** to assign batch $x_{1:N}$ to agents.

**Note: Discretization prevents reverse-engineering of raw data points.*

Wasserstein Distance

2-Wasserstein distance (informal):

Let μ and ν be Borel probability measures on $\mathcal{X}$ with finite second moments, then

$$W_2(\mu,\nu) \triangleq \left(\inf_{\gamma \in \Gamma[\mu,\nu]} \int_{(x,x') \in \mathcal{X}\times\mathcal{X}} \|x - x'\|^2 \, \mathrm{d}\gamma\,(x,x') \right)^{\frac{1}{2}},$$

where $\Gamma[\mu,\nu]$ denotes the set of probability measures on $\mathcal{X} \times \mathcal{X}$, with marginal distributions μ, ν.

- $W_2(\mu,\nu)$ is the minimal root-mean-square transport cost for moving mass from μ to ν

Aggregating GPs via Wasserstein Barycenters

We treat GPs as probability measures on the function space.

Definition: The Central Model f^c minimizes the sum of squared 2-Wasserstein distances to local posteriors:

$$f^c = \inf_{f' \in \mathcal{P}(\mathcal{X})} \sum_{n=1}^{N} [W_2(f', \tilde{f}_n)]^2$$

where $\mathcal{P}(\mathcal{X})$ denotes the set of all probability measures on domain $\mathcal{X}$.

Aggregating GPs via Wasserstein Barycenters

We treat GPs as probability measures on the function space.

Definition: The Central Model f^c minimizes the sum of squared 2-Wasserstein distances to local posteriors:

$$f^c = \inf_{f' \in \mathcal{P}(\mathcal{X})} \sum_{n=1}^{N} [W_2(f', \tilde{f}_n)]^2$$

where $\mathcal{P}(\mathcal{X})$ denotes the set of all probability measures on domain $\mathcal{X}$.

Proposition 1 (Mallasto et al. 2017): The barycenter $f^c \sim \mathcal{GP}(\mu^c, K^c)$ is unique.

- **Mean:** Average of local means: $\mu^c(x) = \frac{1}{N} \sum_{n=1}^{N} \tilde{\mu}_n(x)$.
- **Kernel:** Fixed-point of the operator equation:

$$K^c = \frac{1}{N} \sum_{n=1}^{N} ((K^c)^{1/2} \tilde{K}_n (K^c)^{1/2})^{1/2}$$

Intuition: Preserves geometric structure of uncertainty better than linear averaging.

The Co-KG Acquisition Function

We propose a hybrid objective balancing **Collaboration** vs. **Self-Reliance**:

$$\alpha(x) = \underbrace{\alpha^c(x)}_{\text{Central (Collaboration)}} + \beta_t \sum_{n=1}^{N} \underbrace{\alpha_n(x_n)}_{\text{Local (Self-Reliance)}}$$

Structure:

- $\alpha^c(x)$: Batch Knowledge Gradient (q-KG) on central model f^c.
- $\alpha_n(x_n)$: Standard KG on local model $\tilde{f}_n$.
- β_t: Trade-off parameter.

Design of β_t:

- $\beta_t \to 0$: Trust Central (Early stage, sparse data $\to$ aggregation helps).
- $\beta_t \to \infty$: Trust Local (Late stage, sufficient local data $\to$ reduce bias from approximation).

Co-KG: Mathematical Definition

$$\alpha_{\text{Co-KG}}(x) \triangleq \mathbb{E}_{\tilde{\mathcal{S}}}\left[\max_{x'} \mathbb{E}[f^c(x')|\tilde{S}_x^*] + \beta_t \sum_{n=1}^{N} \max_{x'} \mathbb{E}[\tilde{f}_n(x')|\tilde{\mathcal{S}}_{x_n}^*]\right]$$

Implementation Details:

- **No Data Access:** The conditional expectation $\mathbb{E}[f^c|\tilde{S}_x^*]$ is computed using only μ^c, K^c.
- **Monte Carlo Approximation (Balandat et al. 2019):** We simulate future observations $y(x)$ using the "reparameterization trick":

$$\mathbb{E}\left[f\left(x'\right) \mid \tilde{\mathcal{S}}_{x_n}^*\right] = \tilde{\mu}_n\left(x'\right) + \tilde{\sigma}_n\left(x_n, x'\right)\xi, \quad \xi \sim \mathcal{N}(0, I)$$

where $\tilde{\sigma}_n\left(x_n, x'\right) = \tilde{K}_n\left(x_n, x'\right) / \sqrt{\tilde{K}_n\left(x_n, x_n\right) + \sigma_\epsilon^2}$.

Algorithm 1: Collaborative BO with Co-KG

Input: Prior kernel K, Agents $1..N$.
Warm-up: Agents collect initial observations independently.
for iteration $t = 1, \ldots, T$ **do**
    Communication: Agents send discretized posterior parameters to Server.
    Server: Computes Barycenter mean μ^c and covariance K^c.
    Server: Solves $x^*_{1:N} = \arg\max_x \hat{\alpha}_{Co\text{-}KG}(x)$ (via MC).
    Agents: Collect observation at assigned $x_{n;t}$.
    Agents: Update local GP posteriors.
end for
Output: Server aggregates local optima: $\hat{x}^* = \arg\max\{\mu^*_1, \ldots, \mu^*_N\}$.

Theoretical Guarantee: Consistency

Assumption 1 (Standard BO Assumptions):

- $\mathcal{X}$ is compact.
- Kernel $K(x, x') = \tau^2 \rho(x - x')$ satisfies specific decay rates (e.g., RBF).

Theorem 2: Consistency of Co-KG

Under Assumption 1, as the iterations $T \to \infty$, the collaborative BO procedure is consistent:

$$\lim_{T \to \infty} f(\hat{x}^*) = \max_{x \in \mathcal{X}} f(x)$$

Consistency of Approximation

Since Co-KG cannot be computed analytically, we use Monte Carlo (MC) sampling.

Theorem 3: Convergence of MC Approximation

Let $\hat{x}_M^*$ be the maximizer of the MC approximation with M samples. Then:

$$\lim_{M \to \infty} \text{dist}(\hat{x}_M^*, \mathcal{X}^*) = 0$$

where $\mathcal{X}^*$ is the set of true optimizers of Co-KG.

Implication: The numerical implementation faithfully represents the theoretical policy.

Experimental Setup

Baselines:

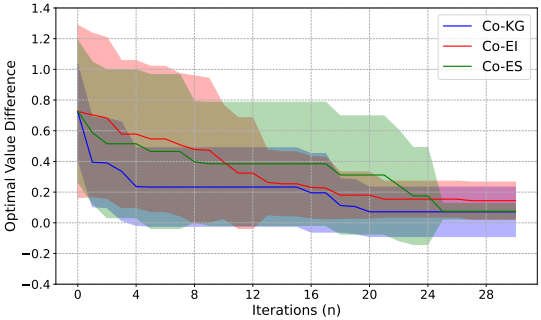
- **Data Communication (Oracle):** Parallel BO sharing all raw data (Upper Bound).
- **No Collaboration:** Independent agents (Lower Bound).
- **Barycenter-qKG:** Only central model term (No local differentiation).
- **Co-EI / Co-ES:** Collaborative Expected Improvement / Entropy Search.

Metrics: Optimal Value Difference (Regret) & Validation Loss.

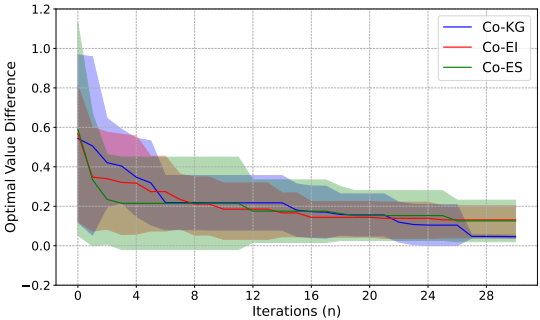
Tasks:

1. Synthetic:
   - $f_1(x) = x_1^2 + x_2^2 + \sin(2\pi x_1) + \cos(2\pi x_2)$
   - $f_2(x) = (1 - x_1)^2 + 100(x_2 - x_1^2)^2$
2. Neural network tuning (Breast Cancer & California Housing data sets)

Acquisition Function Comparison



Synthetic Function f_1

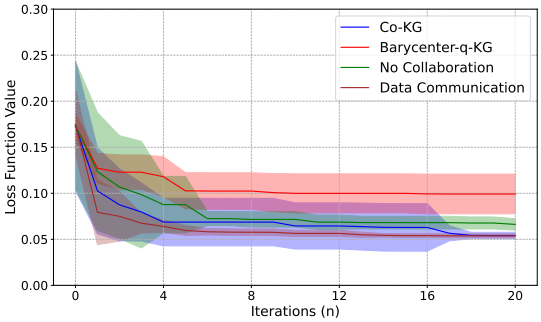


Rosenbrock Function f_2

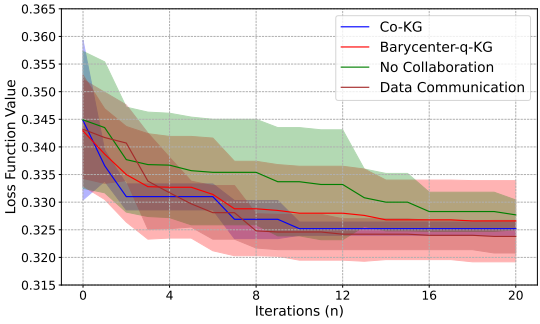
Findings:

- **Co-KG** (Blue) consistently outperforms Co-EI and Co-ES.
- *Observation on f_2:* Co-KG is slightly overconfident initially but surpasses others as data accumulates, demonstrating better long-term convergence.

NN Hyperparameter Tuning



Breast Cancer Dataset

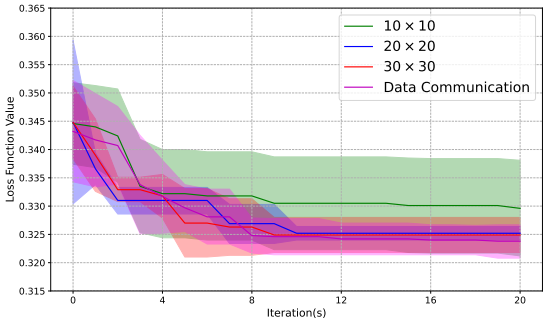


California Housing Dataset

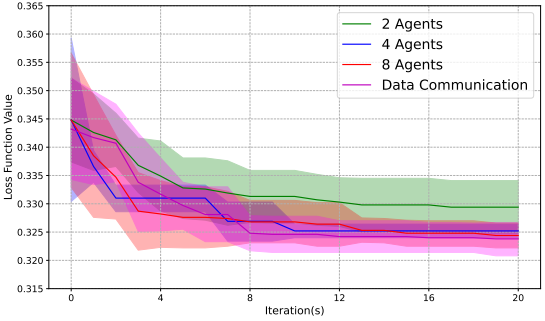
Conclusion:

- **Co-KG** (Blue) is competitive with "Data Communication" (Oracle).
- Outperforms "No Collaboration" and "Barycenter-only".
- Successfully mitigates the cost of privacy constraints.

Scalability: Discretization & Agents



Discretization Grid Scale



Number of Agents

Scalability:

- Discretization Grid: Finer grid (30×30 vs 10×10) improves accuracy.
- Number of Agents: Tested $N = 2, 4, 8$. $N = 8$ outperforms $N = 4$ and $N = 2$ initially.

Summary & Future Work

Contributions:

1. **Framework:** Privacy-preserving Collaborative BO using Wasserstein Barycenters for GPs.
2. **Co-KG:** A novel acquisition function balancing Central vs. Local exploration.
3. **Theory:** Proven consistency for both the algorithm and its MC approximation.
4. **Results:** Competitive with non-private baselines on real tasks.

Future Directions:

- Adaptive agent weighting (robustness to poor initializations).
- Adaptive discretization grids to manage computational cost.
- High-dimensional extensions beyond grid-based discretization.