VII. Proof of Theoretical Results

We prove Theorem 2 in the main text here. The result requires Assumption 1 as stated in the main text and repeated below:

Assumption 1.

1) The feasible set $\mathcal{X}$ is a compact set.

2) Regarding the kernel function $K(x, x')$, there exists a constant $\tau > 0$ and a continuous function $\rho : \mathbb{R}^d \mapsto \mathbb{R}_+$ such that $K(x, x') = \tau^2 \rho(x - x')$. Moreover,

a) $\rho(|\delta|) = \rho(\delta)$, where $|\cdot|$ means taking the absolute value component-wise;

b) $\rho(\delta)$ is decreasing in δ component-wise for $\delta \geqslant \mathbf{0}$;

c) $\rho(\mathbf{0}) = 1, \rho(\delta) \to 0$ as $\|\delta\| \to \infty$, where $\|\cdot\|$ denotes the Euclidean norm;

d) there exist some $0 < C < \infty$ and $\varepsilon, u > 0$ such that

$$1 - \rho(\delta) \leqslant \frac{C}{|\log(\|\delta\|)|^{1+\varepsilon}},$$

for all δ such that $\|\delta\| < u$.

We first focus on the local GP model. Here we hide the index n of each local GP. Recall that the objective function to optimize $f(x)$ is a GP model with the prior kernel function $K(x, x')$. Specifically, we have the following proposition on the convergence of posterior kernel functions of $f(x)$.

Proposition 4 (Proposition 1 of [38])**.** *If the kernel function $K(x, x')$ satisfies the condition in Assumption 1, then*

$$\lim_{t \to \infty} \tilde{K}(x, x') \xrightarrow{a.s.} K^{\infty}(x, x'),$$

and the convergence is uniform. Here $\tilde{K}(x, x')$ is the posterior kernel function after collecting t observations, as defined in the main text, and $K^{\infty}(x, x')$ is a function that does not depend on t.

Next we provide a corollary regarding the posterior variance

$$\mathrm{Var}\left[f(x) \mid \tilde{\mathcal{S}}\right] = \tilde{K}(x, x).$$

We note that, under Assumption 1, there would be an accumulative point $x^{acc} \in \mathcal{X}$ for each local GP model. We here provide an asymptotic upper bound of $\mathrm{Var}\left[f(x) \mid \tilde{\mathcal{S}}\right]$ within an area centered at this accumulative point.

Lemma 5 (Lemma 6 of [38])**.** *Under Assumption 1, $\forall \epsilon > 0$, we have*

$$\limsup_{t \to \infty} \max_{x \in \mathcal{B}(x^{acc}, \epsilon)} \mathrm{Var}\left[f(x) \mid \tilde{\mathcal{S}}\right] \leqslant \tau^2 \left[1 - \rho^2(2\epsilon \mathbf{1})\right],$$

where $\mathbf{1}$ is the vector of all ones with size $d \times 1$, $\mathcal{B}(x^{acc}, \epsilon)$ is the ball centered at x^{acc} with radius ϵ.

Recall that our Co-KG function is composed of one q-KG function with multiple decision variables as the input and multiple (regular) KG functions with one decision variable as the input. In the main text, we subtract the maximum posterior mean for simplification. That is, an equivalent definition of the KG function is

$$\alpha_{\mathrm{KG}}(x) = \mathbb{E}_{\tilde{\mathcal{S}}}\left[\max_{x' \in \mathcal{X}} \mathbb{E}\left[f(x') \mid \tilde{\mathcal{S}}_x^*\right]\right] - \max_{x' \in \mathcal{X}} \mathbb{E}\left[f(x') \mid \tilde{\mathcal{S}}\right]. \tag{10}$$

Here $\tilde{\mathcal{S}}$ represents the historical observations, and $\tilde{\mathcal{S}}_x^* = \tilde{\mathcal{S}} \cup \{(x, y(x))\}$ is the updated dataset if an additional observation $y(x)$ is decided to be collected at the decision variable x. In this manner, regarding the posterior mean as the approximated objective function, KG represents the increment of the optimal value if an additional sample is collected at x. Since the term

$$\max_{x' \in \mathcal{X}} \mathbb{E}\left[f(x') \mid \tilde{\mathcal{S}}\right]$$

does not involve the decision variable x to be optimized, we do not include it in the main text considering the limited length. A similar definition of the q-KG is in [9]. When we prove the consistency of the collaborative BO procedure with Co-KG in this section, these terms are included. Furthermore, we note that the KG function is non-negative, to see this, we use the Jensen inequality:

$$\begin{aligned}
\mathbb{E}\left[\max_{x' \in \mathcal{X}} \mathbb{E}\left[f(x') \mid \tilde{\mathcal{S}}_x^*\right]\right] &= \mathbb{E}\left[\max_{x' \in \mathcal{X}} \mathbb{E}\left[\tilde{\mu}(x') + \tilde{\sigma}(x, x')\,\xi\right]\right] \\
&\geqslant \max_{x' \in \mathcal{X}} \tilde{\mu}(x') + \mathbb{E}\left[\tilde{\sigma}(x, x')\,\xi\right] \\
&= \max_{x' \in \mathcal{X}} \mathbb{E}\left[f(x') \mid \tilde{\mathcal{S}}\right],
\end{aligned}$$

where $\tilde{\sigma}(x,x') = \tilde{K}(x,x')/\sqrt{\tilde{K}(x,x)+\sigma_\epsilon^2}$.The non-negativity based on the Jensen inequality also holds for the q-KG function with a similar argument [9]. Since our Co-KG function is a weighted summation of a q-KG function and multiple regular KG functions, it is non-negative as well.

Regarding the KG function associated with each local GP model (10), we have

$$\begin{aligned}
\alpha_{\mathrm{KG}}(x) =& \mathbb{E}\left[\max_{x'\in\mathcal{X}} \mathbb{E}\left[f(x') \mid \tilde{\mathcal{S}}_x^*\right]\right] - \max_{x'\in\mathcal{X}} \mathbb{E}\left[f(x') \mid \tilde{\mathcal{S}}\right] \\
=& \mathbb{E}\left[\max_{x'\in\mathcal{X}} \mathbb{E}\left[\tilde{\mu}(x') + \tilde{\sigma}(x,x')\,\xi\right]\right] - \max_{x'\in\mathcal{X}} \mathbb{E}\left[f(x') \mid \tilde{\mathcal{S}}\right] \\
\leqslant& \max_{x'\in\mathcal{X}} \tilde{\mu}(x') + \mathbb{E}\left[\max_{x'\in\mathcal{X}} \tilde{\sigma}(x,x')\,\xi\right] - \max_{x'\in\mathcal{X}} \mathbb{E}\left[f(x') \mid \tilde{\mathcal{S}}\right] \\
=& \mathbb{E}\left[\max_{x'\in\mathcal{X}} \tilde{\sigma}(x,x')\,\xi\right] \\
\leqslant& \mathbb{E}\left[|\xi|\right] \max_{x'\in\mathcal{X}} \tilde{\sigma}(x,x') \\
=& \sqrt{\frac{2}{\pi}} \max_{x'\in\mathcal{X}} \tilde{\sigma}(x,x')
\end{aligned} \tag{11}$$

Furthermore, we have that

$$\begin{aligned}
\tilde{\sigma}(x,x') &= \frac{\tilde{K}(x,x')}{\sqrt{\tilde{K}(x,x)+\sigma_\epsilon^2}} \\
&\leqslant \sqrt{\frac{\tilde{K}(x,x)\,\tilde{K}(x',x')}{\tilde{K}(x,x)+\sigma_\epsilon^2}} \\
&\leqslant \sqrt{\frac{\tau^2\tilde{K}(x,x)}{\sigma_\epsilon^2}},
\end{aligned} \tag{12}$$

where the last inequality comes from the fact that $\tilde{K}(x',x')$ is a non-increasing sequence regarding t and the conditions in Assumption 1. Thus, from (11) and (12), we have

$$\alpha_{\mathrm{KG}}(x) \leqslant \sqrt{\frac{2\tau^2\tilde{K}(x,x)}{\pi\sigma_\epsilon^2}} \tag{13}$$

Regarding the q-KG function, we have

$$\mathbb{E}\left[f^c(x') \mid \tilde{\mathcal{S}}_{\mathbf{x}}^*\right] = \mu^c(x') + \boldsymbol{\sigma}^c(\mathbf{x},x')\,\boldsymbol{\xi},$$

where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_N)$. Additionally,

$$\boldsymbol{\sigma}^c(\mathbf{x},x') = \boldsymbol{K}^c(\mathbf{x},x')^\top \left(\boldsymbol{D}(\mathbf{x})^\top\right)^{-1},$$

where $\boldsymbol{D}(\mathbf{x})$ is the Cholesky factor of the matrix

$$\boldsymbol{\Sigma}(\mathbf{x}) = \begin{pmatrix} K^c(x_1,x_1) & \dots & K^c(x_1,x_N) \\ \vdots & \ddots & \vdots \\ K^c(x_N,x_1) & \dots & K^c(x_N,x_N) \end{pmatrix} + \sigma_\epsilon^2 \boldsymbol{I}_N.$$

Note that,

$$\boldsymbol{\sigma}^c(\mathbf{x},x')\,\boldsymbol{\xi} \sim \mathcal{N}\left(0, \|\boldsymbol{\sigma}^c(\mathbf{x},x')\|\right).$$

With a similar argument as in (11), the q-KG function is bounded by

$$\alpha_{q\text{-KG}}(\mathbf{x}) \leqslant \sqrt{\frac{2}{\pi}} \max_{x'\in\mathcal{X}} \|\boldsymbol{\sigma}^c(\mathbf{x},x')\|.$$

Furthermore,

$$\begin{aligned}
&\left\|\boldsymbol{\sigma}^c\left(\mathbf{x}, x'\right)\right\|^2 \\
=&\boldsymbol{K}^c\left(\mathbf{x}, x'\right)^\top \boldsymbol{\Sigma}\left(\mathbf{x}\right) \boldsymbol{K}^c\left(\mathbf{x}, x'\right) \\
\leqslant& K^c\left(x', x'\right) \\
=&\left(\frac{1}{N}\sum_{n=1}^N \tilde{K}_n\left(x', x'\right)\right) + \frac{1}{N}\sum_{n=1}^N \left(\tilde{\mu}_n\left(x'\right) - \mu^c\left(x'\right)\right)^2,
\end{aligned} \tag{14}$$

where the last equality comes from the fact that the central GP is a 2-Wasserstein barycenter of local GPs. In this manner, we have connected the terms of the Co-KG function to the variances of local models. We here present a proposition regarding the conditional mean function $\tilde{\mu}\left(x\right)$.

Proposition 6 (Proposition 2.9 in [43])**.** *Under Assumption 1, the conditional mean function converges to* $\mu^\infty(x) \doteq \mathbb{E}\left[f(x) \mid \mathcal{F}_\infty\right]$ *uniformly in* $x \in \mathcal{X}$ *almost surely (a.s.). That is,*

$$\mathbb{P}\left\{\sup_{x\in\mathcal{X}} |\tilde{\mu}(x) - \mu^\infty(x)| \to 0\right\} = 1$$

as $t \to \infty$.

Here $\mathcal{F}_\infty$ *denotes the filtration of the dataset collection when the number of iterations approaches infinity.*

Next we consider a rescaled Co-KG function:

$$\alpha_{\text{Co-KG}}\left(\mathbf{x}\right) \doteq \mathbb{E}_{\tilde{\mathcal{S}}}\left\{\frac{\alpha^c\left(\mathbf{x}\right)}{\beta_t} + \sum_{n=1}^N \alpha_n\left(x\right)\right\},$$

where $\alpha^c(\mathbf{x})$ is the q-KG function defined on the central GP and $\alpha_n(x)$ is the regular KG function defined on the n-th local GP model. We provide an asymptotic property of Co-KG:

Lemma 7. *Under Assumption 1, the limit inferior of Co-KG is 0. That is,*

$$\liminf_{t\to\infty} \alpha_{\textit{Co-KG}}(\mathbf{x}) = 0 \quad \forall \mathbf{x} \in \mathcal{X}^N.$$

Proof. We first provide the notation here. We denote the sequence of decision variables selected by maximizing Co-KG as

$$\mathbf{x}(t) = \left(x_1\left(t\right), x_2\left(t\right), \ldots, x_N\left(t\right)\right),$$

where $x_n(t)$ denotes the decision variables for the n-th agent in the t-th iteration. In the main text, we did not emphasize the dependence on t for notational simplicity. For each local model, note that the sequence of the selected $x_n(t)$'s for this local GP has an accumulative point x^{acc}. We denote the subsequence as $z_n(1), z_n(2), \ldots, z_n\left(t'\right)$ such that $z_n\left(t'\right) \to x^{acc}$. Based on Lemma 5, we have

$$\limsup_{t\to\infty} \operatorname{Var}\left[f\left(z_n\left(t'\right)\right) \mid \tilde{\mathcal{S}}_n\right] \leqslant \tau^2\left[1 - \rho^2(2\epsilon\mathbf{1})\right].$$

Furthermore, from (13), we have

$$\limsup_{t\to\infty} \alpha(z_n\left(t'\right)) \leqslant \sqrt{\frac{2\tau^2\tilde{K}\left(z_n\left(t'\right), z_n\left(t'\right)\right)}{\pi\sigma_\epsilon^2}} \leqslant \limsup_{t\to\infty} \sqrt{\frac{2\tau^4}{\pi\sigma_\epsilon^2}}\left[1 - \rho^2(2\epsilon\mathbf{1})\right].$$

Let $\epsilon \to 0$, we have that

$$\liminf_{t\to\infty} \alpha_{\text{KG}}\left(x_n(t)\right) \leqslant \liminf_{t\to\infty} \alpha_{\text{KG}}\left(z_n\left(t'\right)\right) \leqslant \limsup_{t\to\infty} \alpha_{\text{KG}}(z_n\left(t'\right)) = 0 \quad \forall n. \tag{15}$$

The last equality holds because the KG function $\alpha(x)$ is non-negative.

Then we look at the central GP and the associated q-KG function. Recall that, from (14), we have

$$\alpha_{q\text{-KG}}\left(\mathbf{x}\right) \leqslant \sqrt{\frac{2}{\pi}\left(\left(\frac{1}{N}\sum_{n=1}^N \tilde{K}_n\left(x', x'\right)\right) + \frac{1}{N}\sum_{n=1}^N \left(\tilde{\mu}_n\left(x'\right) - \mu^c\left(x'\right)\right)^2\right)} \quad \forall \mathbf{x},$$

where x' denotes some maximizer. Note that $\tilde{K}_n\left(x', x'\right) \leqslant K\left(x', x'\right)$ is bounded, and each $\left(\tilde{\mu}_n\left(x'\right) - \mu^c\left(x'\right)\right)^2$ is bounded as well because of Proposition 6. Thus, as $t \to \infty$,

$$\lim_{t\to\infty} \frac{\alpha_{q\text{-KG}}\left(\mathbf{x}\left(t\right)\right)}{\beta_t}$$

since $\beta_t \to \infty$ from the definition in the main text. This further leads to

$$\liminf_{t\to\infty} \alpha_{\text{Co-KG}}\left(\mathbf{x}\left(t\right)\right) = 0$$

because of (15). Recall that

$$\mathbf{x}\left(t\right) = \arg\max_{\mathbf{x}\in\mathcal{X}^N} \alpha_{\text{Co-KG}}\left(\mathbf{x}\right).$$

Thus, $\forall \mathbf{x}$, we have

$$0 \leqslant \liminf_{t\to 0} \alpha_{\text{Co-KG}}(\mathbf{x}) \leqslant \liminf_{t\to 0} \alpha_{\text{Co-KG}}(\mathbf{x}(t)) = 0$$

□

Now we prove that for all local GP models, the uncertainty at each decision variable (represented by the conditional variance) shrinks to zero. Specifically, we have the following lemma.

Lemma 8. *Under Assumption 1, $\forall n$ and $\forall x \in \mathcal{X}$, we have*

$$\lim_{t\to\infty} \operatorname{Var}\left[f\left(x\right) \mid \tilde{\mathcal{S}}_n\right] = 0 \quad \forall x \in \mathcal{X}, \forall n. \tag{16}$$

Proof. Without loss of generality, we assume that

$$\lim_{t\to\infty} \operatorname{Var}\left[f\left(\tilde{x}\right) \mid \tilde{\mathcal{S}}_1\right] = c > 0. \tag{17}$$

This limit exists because of Lemma 5. Also, since $\tilde{K}\left(x, x'\right)$ is continuous and the convergence in Lemma 5 is uniform, $\operatorname{Var}\left[f\left(x\right) \mid \tilde{\mathcal{S}}_1\right]$ is continuous as well. Regarding this agent, we denote the current posterior function as $\tilde{\mu}\left(x\right) = \mathbb{E}\left[f\left(x\right) \mid \tilde{\mathcal{S}}_1\right]$ and the posterior function with additional x as $\tilde{\mu}^*\left(x'\right) = \mathbb{E}\left[f\left(x'\right) \mid \tilde{\mathcal{S}}_x^*\right]$. Furthermore, we denote $\tilde{x}^* = \arg\max_{x\in\mathcal{X}} \tilde{\mu}(x)$. Also, let

$$\begin{aligned}
a_1 =& \tilde{\mu}\left(\tilde{x}^*\right)\\
b_1 =& \tilde{\sigma}\left(x, \tilde{x}^*\right)\\
a_2 =& \tilde{\mu}\left(x\right)\\
b_2 =& \tilde{\sigma}\left(x, x\right).
\end{aligned}$$

We consider the KG function associated with this agent:

$$\begin{aligned}
&\alpha_{\text{KG}}(x)\\
&= \mathbb{E}\left[\max_{x'\in\mathcal{X}} \mathbb{E}\left[f\left(x'\right) \mid \tilde{\mathcal{S}}_x^*\right]\right] - \max_{x'\in\mathcal{X}} \mathbb{E}\left[f\left(x'\right) \mid \tilde{\mathcal{S}}\right]\\
&= \mathbb{E}\left[\max_{x'\in\mathcal{X}} \tilde{\mu}^*\left(x'\right)\right] - \max\left(\tilde{\mu}\left(\tilde{x}^*\right), \tilde{\mu}(x)\right)\\
&\geqslant \mathbb{E}\left[\max\left(\tilde{\mu}^*\left(\tilde{x}^*\right), \tilde{\mu}^*(x)\right)\right] - \max\left(\tilde{\mu}\left(\tilde{x}^*\right), \tilde{\mu}(x)\right)\\
&= \mathbb{E}\left[\max\left(\tilde{\mu}\left(\tilde{x}^*\right) + \tilde{\sigma}\left(x, \tilde{x}^*\right)\xi, \tilde{\mu}(x) + \tilde{\sigma}\left(x, x\right)\xi\right)\right] - \max\left(\tilde{\mu}\left(\tilde{x}^*\right), \tilde{\mu}(x)\right)\\
&= \mathbb{E}\left[\max\left(a_1 + b_1\xi, a_2 + b_2\xi\right)\right] - \max\left(a_1, a_2\right)\\
&= \begin{cases}
\int_{-\infty}^{\frac{a_2-a_1}{b_1-b_2}} \left(a_2 + b_2\xi\right)\phi(\xi)d\xi + \int_{\frac{a_2-a_1}{b_1-b_2}}^{\infty} \left(a_1 + b_1\xi\right)\phi(\xi)d\xi - \max\left(a_1, a_2\right), & \text{if } b_2 \leqslant b_1\\
\int_{-\infty}^{\frac{a_2-a_1}{b_1-b_2}} \left(a_1 + b_1\xi\right)\phi(\xi)d\xi + \int_{\frac{a_2-a_1}{b_1-b_2}}^{\infty} \left(a_2 + b_2\xi\right)\phi(\xi)d\xi - \max\left(a_1, a_2\right), & \text{if } b_1 < b_2
\end{cases}\\
&= \begin{cases}
a_2\Phi\left(\frac{a_2-a_1}{b_1-b_2}\right) - b_2\phi\left(\frac{a_2-a_1}{b_1-b_2}\right) + a_1\left(1 - \Phi\left(\frac{a_2-a_1}{b_1-b_2}\right)\right) + b_1\phi\left(\frac{a_2-a_1}{b_1-b_2}\right) - \max\left(a_1, a_2\right), & \text{if } b_2 \leqslant b_1\\
a_1\Phi\left(\frac{a_2-a_1}{b_1-b_2}\right) - b_1\phi\left(\frac{a_2-a_1}{b_1-b_2}\right) + a_2\left(1 - \Phi\left(\frac{a_2-a_1}{b_1-b_2}\right)\right) + b_2\phi\left(\frac{a_2-a_1}{b_1-b_2}\right) - \max\left(a_1, a_2\right), & \text{if } b_1 < b_2
\end{cases}\\
&= a_2\Phi\left(\frac{a_2-a_1}{|b_1-b_2|}\right) + a_1\left(1 - \Phi\left(\frac{a_2-a_1}{|b_1-b_2|}\right)\right) + |b_1-b_2|\,\phi\left(\frac{a_2-a_1}{|b_1-b_2|}\right) - \max\left(a_1, a_2\right)\\
&= -|a_2-a_1|\,\Phi\left(\frac{-|a_2-a_1|}{|b_1-b_2|}\right) + |b_1-b_2|\,\phi\left(\frac{|a_2-a_1|}{|b_1-b_2|}\right),
\end{aligned}$$

where Φ is the standard normal distribution function and ϕ is its density function.

Let $g(s,t) := t\phi(s/t) - s\Phi(-s/t)$. Then 1) $g(s,t) > 0$ for all $s \geqslant 0$ and $t > 0$; 2) $g(s,t)$ is strictly decreasing in $s \in [0,\infty)$ and strictly increasing in $t \in (0,\infty)$; and 3) $g(s,t) \to 0$ as $s \to \infty$ or as $t \to 0$. See more details in [38], [44]. By letting

$x = \tilde{x}$, which is defined in (17), we have that $\liminf_{t\to\infty} |b_1 - b_2| \geqslant c'$ for some constant $c' > 0$. Meanwhile, based on Proposition 6, we have $\limsup_{t\to\infty} |a_2 - a_1| \leqslant r'$ for some constant $r' < \infty$. Thus,

$$\liminf_{t\to\infty} \alpha_{\text{KG}}(\tilde{x}) \geqslant g(r', c') > 0.$$

A similar argument is in Theorem 5.6 in [44]. On the other hand, since q-KG and KG are non-negative, we have

$$\liminf_{t\to\infty} \alpha_{\text{Co-KG}}(\mathbf{x}) \geqslant \liminf_{t\to\infty} \alpha_{\text{KG}}(x) > 0,$$

where $\tilde{\mathbf{x}}$ includes $\tilde{x}$ as one component. This provides a contradiction with Lemma 7, which proves the result in (16).

□

Based on Lemma 8, we prove Theorem 2 in the main text.

Proof. For each local GP, we have

$$\mathbb{E}\left[\tilde{\mu}_n(x) - f(x)\right]^2 = \text{Var}\left[f(x) \mid \tilde{\mathcal{S}}_n\right] \to 0$$

from Lemma 8. In addition, based on Proposition 6, we have that

$$\mathbb{P}\left\{\sup_{x\in\mathcal{X}} |\tilde{\mu}_n(x) - \mu_n^{\infty}(x)| \to 0\right\} = 1 \quad \forall n$$

as $t \to \infty$. Thus, $\mu_n^{\infty}(x) \overset{a.s.}{=} f(x)$ and

$$\mathbb{P}\left\{\sup_{x\in\mathcal{X}} |\tilde{\mu}_n(x) - f(x)| \to 0\right\} = 1 \quad \forall n$$

as $t \to \infty$. Since kernel function $K(x, x')$ is continuous, the conditional mean function $\tilde{\mu}_n(x)$, as well as $f(x)$, is continuous as well. Therefore, for each local model, the optimizer submitted satisfies that

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

We refer to [45] for a detailed proof. In this manner, we have the consistency of the collaborative BO procedure with Co-KG (as summarized in Algorithm 1 in the main text):

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

□

VIII. Additional Experiments

We present additional experiments here, including 1) the effects of different selections of β_t in the Co-KG function, 2) the effects of the discretization of the feasible set $\mathcal{X}$, 3) the comparison between the Co-KG procedure with different numbers of agents. Our experiments were conducted with Botorch [39] and Python 3.9 on a computer equipped with two AMD Ryzen Threadripper 3970X 32-Core Processors, 128 GB memory, and a Nvidia GeForce RTX A6000 GPU with 48GB of RAM. The implementation will be released once accepted.

A. Feasible Set Discretization

We discuss the impacts of feasible set discretization here. Regarding the hyperparameter, we set $\beta_t = \log(2t + 1)$. We have $N = 4$ agents. We normalize the feasible set to $\mathcal{X} = [0, 1]^2$ and discretize the feasible set using 1) 10×10, 2) 20×20, and 3) 30×30 uniform mesh grids. We also include the results associated with the parallel BO approach without data privacy concerns (q-KG) for comparison, and the procedure is indicated by "Data Communication".

Regarding the black-box optimization problem, we consider minimizing the validation loss of training a neural network. Specifically, the decision variable is the learning rate and the hidden layer node size of the neural networks. The unknown objective function is the validation loss we would minimize. The dataset is about California housing[3], where the neural network is learned to predict median value of houses in different districts given demographic attributes. We do not impose noise on the observations in the real dataset.

The experimental results are included in Figure 7, and provide the following insights: First, increasing the granularity of mesh grid enhances the performance of the collaborative BO approach with Co-KG, with more accurate central model construction and more flexible decision variable selections. Second, when the grid is precise enough (20×20 and 30×30), our approach is comparable with that without data privacy concerns, which is also supported by the experiments in the main text. Lastly,

[3]https://www.dcc.fc.up.pt/~ltorgo/Regression/cal_housing.html

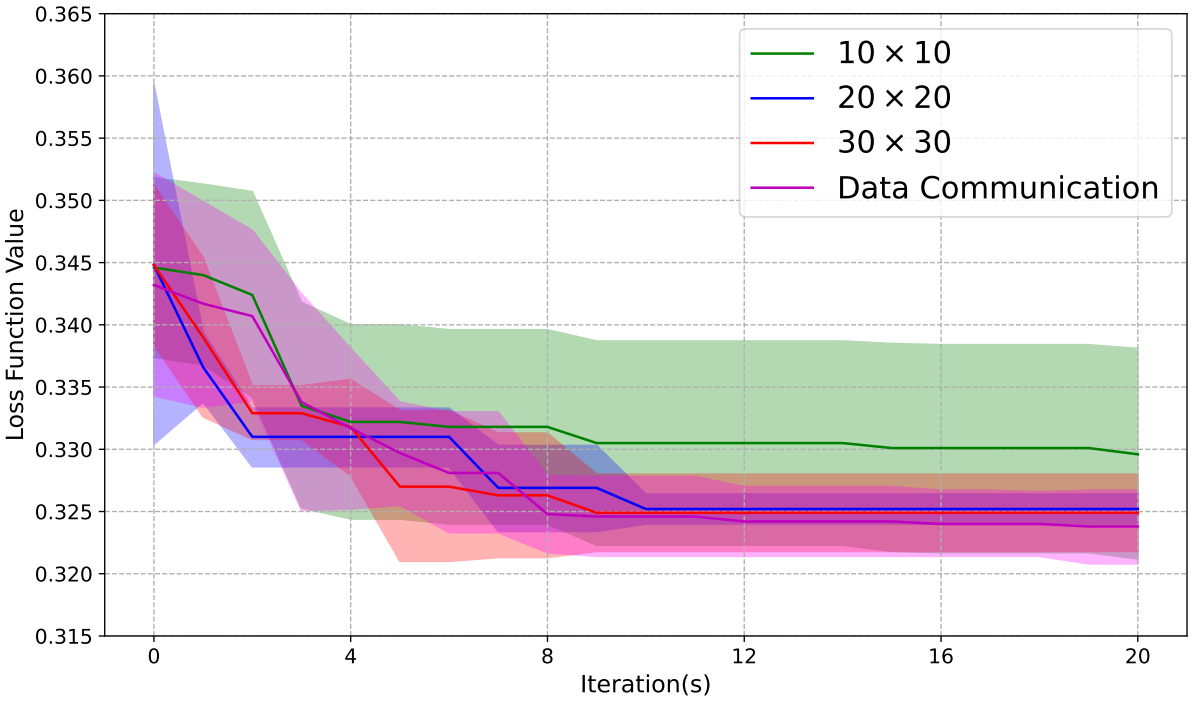

Fig. 7. Validation loss in iterations with different discretization sizes on California Housing Dataset.

although increasing the granularity from 20×20 to 30×30 enhances the performance of Co-KG, the enhancement is not significant. On the other hand, increasing the granularity significantly increases the computational burdens, we include the average running time per iteration in Table I. From the table, we observe that the time complexity for discretization is nearly $O(N^2)$, leading to heavy computational cost if we set a dense discretization for feasible set. Furthermore, we admit that the discretization could be time-varying regarding different iterations and adaptive to the collaborative optimization procedure, while the detailed discussions are left for future work.

Discretization mesh grid	Computational time (s)
10×10	2.02
20×20	6.68
30×30	16.20

TABLE I
COMPUTATIONAL TIME WITH DIFFERENT DISCRETIZATION STRATEGIES.

B. Agent Number Comparison

We compare the effects of the number of agents. The experimental settings are the same in Section VIII-A with the mesh grid fixed to be 20×20. We consider $N = 2, 4, 8$ in our experiments.

The experimental results in Figure 8 reveal the following insights. First, when the number of agents is low ($N = 2$), the performance of Co-KG is suboptimal. Second, comparing the results between $N = 4$ and $N = 8$, we observe that having more agents does not necessarily lead to better performance, especially during the initial iterations. This is because, with more agents, some may be initialized in less promising regions, which negatively impacts collaboration. The Co-KG function currently assigns equal weights to all agents, which is sensitive to suboptimal observations collected by some agents. Future work could explore assigning different weights in the collaborative acquisition function to enhance algorithm robustness. Finally, as the number of iterations increases, the procedure with $N = 8$ agents slightly outperforms that with $N = 4$, but the difference is not substantial. Both procedures tend to stabilize without significant improvements, due to the effects of discretization. This suggests that the optimal number of agents may also depend on the level of discretization, a topic that is left for future discussion.

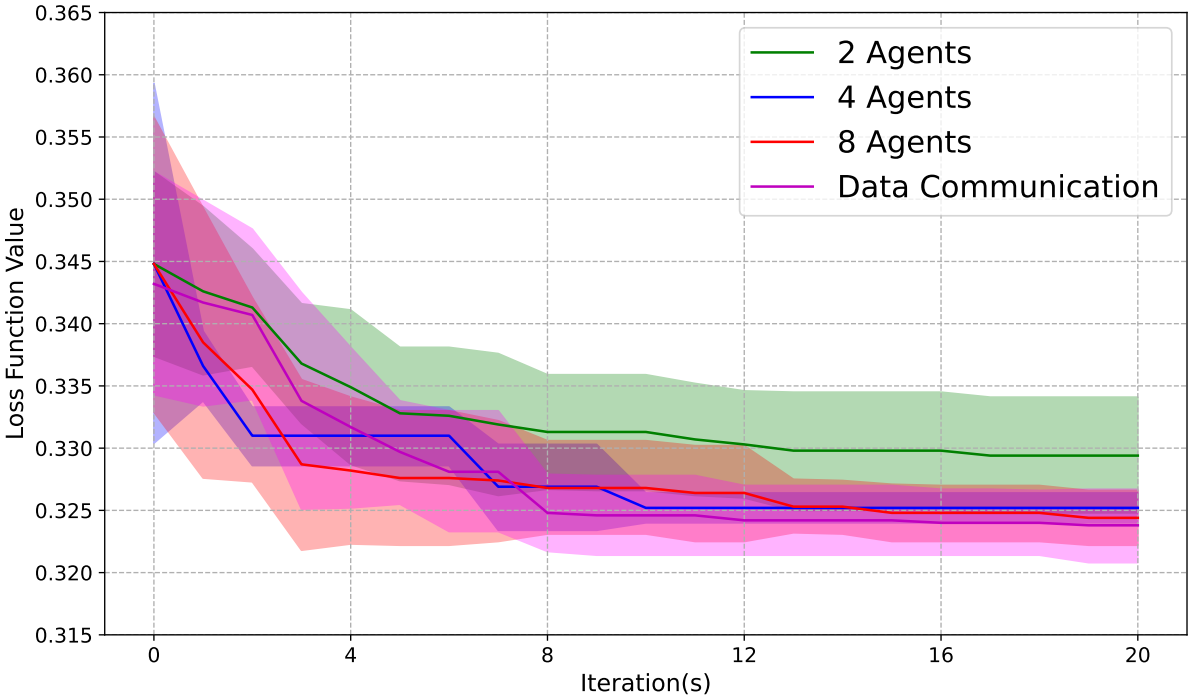

Fig. 8. Validation loss in iterations with different number of agents on California Housing Dataset.