Functional Causal Bayesian Optimization (Supplementary material)

Limor Gultchin[1,*] **Virginia Aglietti**[2,*] **Alexis Bellot**[2] **Silvia Chiappa**[2]

[1]University of Oxford, The Alan Turing Institute, Work done at DeepMind, London, UK
[2]DeepMind, London, UK

1 PROOFS

Proposition 3.2. *Let $\mathcal{G}$ be a causal graph such that (i) $\exists C \in pa_{\mathcal{G}}(Y)$ with $C \notin \boldsymbol{I}$; or (ii) $\exists C \in sp_{\mathcal{G}}(Y)$. If $\exists X \in an_{\mathcal{G}}(Y) \cap \boldsymbol{I}$ such that $\{\langle X, C\rangle\}$ is an* MPS*, then there exists at least one* SCM *compatible with $\mathcal{G}$ for which* $\min_{\mathcal{S}\in\Sigma_{hard},\pi_{\mathcal{S}}\in\Pi_{\mathcal{S}}} \mu^Y_{\pi_{\mathcal{S}}} > \min_{\mathcal{S}\in\Sigma,\pi_{\mathcal{S}}\in\Pi_{\mathcal{S}}} \mu^Y_{\pi_{\mathcal{S}}}$.

Proof. Case (i): Assume that there exists $C \in \text{pa}_{\mathcal{G}}(Y)$ with $C \notin \boldsymbol{I}$ and $X \in \text{an}_{\mathcal{G}}(Y) \cap \boldsymbol{I}$ such that $\{\langle X, C\rangle\}$ is an MPS. As $X \in \text{an}_{\mathcal{G}}(Y)$, there exists a directed path from X to Y, say $X \to X_i \to X_{i-1} \to \cdots \to X_1 \to Y$ without loss of generality. Let $\mathcal{M} = \langle \boldsymbol{V}, \boldsymbol{U}, \mathcal{F}, p(\boldsymbol{U})\rangle$ be an SCM such that

$$\begin{aligned} &C = U_C,\, U_C \sim \mathcal{N}(0,1),\\ &X_i = X,\, X_{i-1} = X_i,\, \ldots,\, X_1 = X_2,\\ &Y = X_1 C U_Y,\, U_Y \sim \mathcal{N}(1,1). \end{aligned}$$

$\mathcal{M}$ is compatible with $\mathcal{G}$. In this SCM, any DMP $\pi_{\mathcal{S}}$ with $\mathcal{S} \in \Sigma_{\text{hard}}$ would give $\mu^Y_{\pi_{\mathcal{S}}} = \mathbb{E}_{\pi_{\mathcal{S}}}[Y] = 0$. In contrast, a DMP $\pi_{\mathcal{S}}$ including the functional intervention $\pi_{X|C}(C) = -1/C$ would result in $Y = -U_Y$ and therefore $\mu^Y_{\pi_{\mathcal{S}}} = -1$, giving $\min_{\mathcal{S}\in\Sigma_{\text{hard}},\pi_{\mathcal{S}}\in\Pi_{\mathcal{S}}} \mu^Y_{\pi_{\mathcal{S}}} = 0 > -1 \geq \min_{\mathcal{S}\in\Sigma,\pi_{\mathcal{S}}\in\Pi_{\mathcal{S}}} \mu^Y_{\pi_{\mathcal{S}}}$.

Case (ii): Assume that there exists $C \in \text{sp}_{\mathcal{G}}(Y)$ and $X \in \text{an}_{\mathcal{G}}(Y) \cap \boldsymbol{I}$ such that $\{\langle X, C\rangle\}$ is an MPS. As $X \in \text{an}_{\mathcal{G}}(Y)$, there exists a directed path from X to Y, say $X \to X_i \to X_{i-1} \to \cdots \to X_1 \to Y$ without loss of generality. Let $\mathcal{M} = \langle \boldsymbol{V}, \boldsymbol{U}, \mathcal{F}, p(\boldsymbol{U})\rangle$ be an SCM such that

$$\begin{aligned} &C = U_{CY},\, U_{CY} \sim \mathcal{N}(0,1),\\ &X_i = X,\, X_{i-1} = X_i,\, \ldots,\, X_1 = X_2,\\ &Y = X_1 U_{CY} U_Y,\, U_Y \sim \mathcal{N}(1,1). \end{aligned}$$

$\mathcal{M}$ is compatible with $\mathcal{G}$. In this SCM, any DMP $\pi_{\mathcal{S}}$ with $\mathcal{S} \in \Sigma_{\text{hard}}$ would give $\mu^Y_{\pi_{\mathcal{S}}} = \mathbb{E}_{\pi_{\mathcal{S}}}[Y] = 0$. In contrast, a DMP $\pi_{\mathcal{S}}$ containing the functional intervention $\pi_{X|C}(C) = -1/C$, would result in $Y = -U_Y$ and therefore $\mu^Y_{\pi_{\mathcal{S}}} = -1$, giving $\min_{\mathcal{S}\in\Sigma_{\text{hard}},\pi_{\mathcal{S}}\in\Pi_{\mathcal{S}}} \mu^Y_{\pi_{\mathcal{S}}} = 0 > -1 \geq \min_{\mathcal{S}\in\Sigma,\pi_{\mathcal{S}}\in\Pi_{\mathcal{S}}} \mu^Y_{\pi_{\mathcal{S}}}$. □

In the following proposition we use the notation $\mathcal{G}_{\underline{\boldsymbol{X}}}$ to indicate the modification of $\mathcal{G}$ obtained by removing the outgoing edges from $\boldsymbol{X}$.

Proposition 3.3. *In a casual graph $\mathcal{G}$, if $pa_{\mathcal{G}}(Y) \subseteq \boldsymbol{I}$ and $sp_{\mathcal{G}}(Y) = \emptyset$ there exists a* DMP *compatible with* MPS *$\mathcal{S} = \{\langle X, \emptyset\rangle : X \in pa_{\mathcal{G}}(Y)\}$ that solves the f*CGO *problem.*

[*]Equal contribution.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

Proof. Consider MPS $\mathcal{S} \in \Sigma$ for $\mathcal{G}$ and DMP $\pi_{\mathcal{S}}$ compatible with $\mathcal{S}$. Let $\boldsymbol{Z} = \text{pa}_{\mathcal{G}}(Y) \backslash ((\boldsymbol{X}_{\mathcal{S}} \cup \boldsymbol{C}_{\mathcal{S}}) \cap \text{pa}_{\mathcal{G}}(Y))$. As $\text{pa}_{\mathcal{G}}(Y) \subseteq \boldsymbol{I}$, we can define the MPS $\mathcal{S}_{\text{pa}} = \{\langle X, \emptyset\rangle : \forall X \in \text{pa}_{\mathcal{G}}(Y)\}$. Denote by $p_{\pi^*_{\mathcal{S}_{\text{pa}}}}(Y)$ the distribution of Y induced by an optimal DMP $\pi^*_{\mathcal{S}_{\text{pa}}}$ compatible with $\mathcal{S}_{\text{pa}}$, i.e. such that $\int_{\mathcal{R}_Y} Y p_{\pi^*_{\mathcal{S}_{\text{pa}}}}(Y) dY \leq \int_{\mathcal{R}_Y} Y p_{\pi_{\mathcal{S}_{\text{pa}}}}(Y) dY$, for every DMP $\pi_{\mathcal{S}_{\text{pa}}}$ compatible with $\mathcal{S}_{\text{pa}}$, and let $\mathcal{R} = \mathcal{R}_Y \times \mathcal{R}_{\boldsymbol{X}_{\mathcal{S}} \cup \boldsymbol{C}_{\mathcal{S}}} \times \mathcal{R}_{\boldsymbol{Z}}$. Exploiting the rules of do-calculus [Pearl, 2000] and σ-calculus [Correa and Bareinboim, 2020a] we obtain

$$
\begin{aligned}
\mu^Y_{\pi_{\mathcal{S}}} &= \int_{\mathcal{R}} Y p_{\pi_{\mathcal{S}}}(Y \mid \boldsymbol{X}_{\mathcal{S}} \cup \boldsymbol{C}_{\mathcal{S}} \cup \boldsymbol{Z}) \underbrace{p_{\pi_{\mathcal{S}}}(\boldsymbol{X}_{\mathcal{S}} \cup \boldsymbol{C}_{\mathcal{S}} \cup \boldsymbol{Z}) d\boldsymbol{X}_{\mathcal{S}} \cup \boldsymbol{C}_{\mathcal{S}} d\boldsymbol{Z} dY}_{\mathcal{A}} \\
&= \int_{\mathcal{R}} Y p_{\pi_{\mathcal{S}}}(Y \mid \text{pa}_{\mathcal{G}}(Y)) \mathcal{A} && \text{(rule 1 } \sigma\text{-calculus) } Y \perp\!\!\!\perp_{\mathcal{G}_{\mathcal{S}}} (\boldsymbol{X}_{\mathcal{S}} \cup \boldsymbol{C}_{\mathcal{S}} \cup \boldsymbol{Z}) \backslash \text{pa}_{\mathcal{G}}(Y) \mid \text{pa}_{\mathcal{G}}(Y) \\
&= \int_{\mathcal{R}} Y p(Y \mid \text{pa}_{\mathcal{G}}(Y)) \mathcal{A} && \text{(rule 2 } \sigma\text{-calculus) } Y \perp\!\!\!\perp_{\mathcal{G}_{\mathcal{S}, \underline{\boldsymbol{X}_{\mathcal{S}}}}, \mathcal{G}_{\underline{\boldsymbol{X}_{\mathcal{S}}}}} \boldsymbol{X}_{\mathcal{S}} \mid (\text{pa}_{\mathcal{G}}(Y) \backslash (\text{pa}_{\mathcal{G}}(Y) \cap \boldsymbol{X}_{\mathcal{S}})) \\
&= \int_{\mathcal{R}} Y p(Y \mid \text{do}(\text{pa}_{\mathcal{G}}(Y))) \mathcal{A} && \text{(rule 2 do-calculus) } Y \perp\!\!\!\perp_{\mathcal{G}_{\underline{\text{pa}_{\mathcal{G}}(Y)}}} \text{pa}_{\mathcal{G}}(Y) \\
&= \int_{\mathcal{R}} Y p_{\pi_{\mathcal{S}_{\text{pa}}}}(Y) \mathcal{A} \geq \int_{\mathcal{R}} Y p_{\pi^*_{\mathcal{S}_{\text{pa}}}}(Y) \mathcal{A} = \mu^Y_{\pi^*_{\mathcal{S}_{\text{pa}}}},
\end{aligned}
$$

where $\perp\!\!\!\perp_{\mathcal{G}_{\mathcal{S}, \underline{\boldsymbol{X}_{\mathcal{S}}}}, \mathcal{G}_{\underline{\boldsymbol{X}_{\mathcal{S}}}}}$ denotes d-separation in both $\mathcal{G}_{\mathcal{S}, \underline{\boldsymbol{X}_{\mathcal{S}}}}$ and $\mathcal{G}_{\underline{\boldsymbol{X}_{\mathcal{S}}}}$. □

Proposition 3.4. *If* $\mathcal{S}^*, \pi^*_{\mathcal{S}^*} = \arg\min_{\mathcal{S} \in \Sigma, \pi_{\mathcal{S}} \in \Pi_{\mathcal{S}}} \mu^Y_{\pi_{\mathcal{S}}}$*, then* $\mathcal{S}^*, \pi^*_{\mathcal{S}^*} = \arg\min_{\mathcal{S} \in \Sigma^{\boldsymbol{C}}, \pi_{\mathcal{S}} \in \Pi_{\mathcal{S}}} \mu^Y_{\pi_{\mathcal{S}}, \boldsymbol{C}=\boldsymbol{c}}$ $\forall \boldsymbol{C} \subset \boldsymbol{V} \backslash Y$ *such that* $\boldsymbol{C} \cap de_{\mathcal{G}}(\boldsymbol{I}) = \emptyset$ *and* $\forall \boldsymbol{c} \in \mathcal{R}_{\boldsymbol{C}}$ *with* $\Sigma^{\boldsymbol{C}} = \{\mathcal{S} \in \Sigma : \boldsymbol{X}_{\mathcal{S}} = \boldsymbol{X}_{\mathcal{S}^*}$ *and* $\{\langle X, \boldsymbol{C}^{\mathcal{S}^*}_X \cup \boldsymbol{C}^{\mathcal{S}}_X \cup \boldsymbol{C}\rangle : X \in \boldsymbol{X}_{\mathcal{S}^*}\}$ *is an* MPS$\}$.

Proof. Assume, by contradiction, that $(\mathcal{S}^*, \pi^*_{\mathcal{S}^*})$, with $\pi^*_{\mathcal{S}^*} = \left\{\pi^{\mathcal{S}^*}_{X|\boldsymbol{C}^{\mathcal{S}^*}_X}\right\}_{X \in \boldsymbol{C}^{\mathcal{S}^*}_X}$, is a solution to the fCGO problem but there exist $\boldsymbol{C} \subset \boldsymbol{V} \backslash Y$ and a value $\boldsymbol{c} \in \mathcal{R}_{\boldsymbol{C}}$ such that the tuple $(\mathcal{S}^1, \pi_{\mathcal{S}^1})$ with $\mathcal{S}^1 \in \Sigma^{\boldsymbol{C}}$ and $\pi_{\mathcal{S}^1} = \left\{\pi^{\mathcal{S}^1}_{X|\boldsymbol{C}^{\mathcal{S}^1}_X}\right\}_{X \in \boldsymbol{C}^{\mathcal{S}^1}_X} \in \Pi_{\mathcal{S}}$ satisfies $\mu^Y_{\pi_{\mathcal{S}^1}, \boldsymbol{C}=\boldsymbol{c}} < \mu^Y_{\pi^*_{\mathcal{S}^*}, \boldsymbol{C}=\boldsymbol{c}}$. As $\mathcal{S}^1 \in \Sigma^{\boldsymbol{C}}$, we can construct MPS $\mathcal{S}^2 = \{\langle X, \boldsymbol{C}^{\mathcal{S}^*}_X \cup \boldsymbol{C}^{\mathcal{S}^1}_X \cup \boldsymbol{C}\rangle : X \in \boldsymbol{X}_{\mathcal{S}^*}\}$ and the compatible $\pi_{\mathcal{S}^2} = \left\{\pi^{\mathcal{S}^2}_{X|\boldsymbol{C}^{\mathcal{S}^*}_X \cup \boldsymbol{C}^{\mathcal{S}^1}_X \cup \boldsymbol{C}}\right\}_{X \in \boldsymbol{X}_{\mathcal{S}^*}}$ with

$$
\pi^{\mathcal{S}^2}_{X|\boldsymbol{C}^{\mathcal{S}^*}_X \cup \boldsymbol{C}^{\mathcal{S}^1}_X \cup \boldsymbol{C}} = \begin{cases} \pi^{\mathcal{S}^1}_{X|\boldsymbol{C}^{\mathcal{S}^1}_X} \text{ if } \boldsymbol{C} \in [\boldsymbol{c} - \delta, \boldsymbol{c} + \delta] \\ \pi^{\mathcal{S}^*}_{X|\boldsymbol{C}^{\mathcal{S}^*}_X} \text{ otherwise,} \end{cases}
$$

for a small enough $\delta > 0$. As $\boldsymbol{C} \cap \text{de}_{\mathcal{G}}(\boldsymbol{I}) = \emptyset$, variables in $\boldsymbol{C}$ are not affected by interventions on variables in $\boldsymbol{X}_{\mathcal{S}^*}$, and therefore $p_{\pi^*_{\mathcal{S}^*}}(\boldsymbol{C}) = p_{\pi_{\mathcal{S}^1}}(\boldsymbol{C}) = p(\boldsymbol{C})$. Thus we obtain:

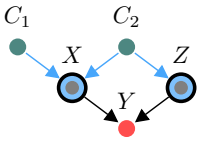
$$
\begin{aligned}
\mu^Y_{\pi_{\mathcal{S}^2}} &= \int_{\mathcal{R}_{\boldsymbol{C}}} \mu^Y_{\pi_{\mathcal{S}^2}, \boldsymbol{C}=\boldsymbol{c}'} \, p_{\pi_{\mathcal{S}^2}}(\boldsymbol{C} = \boldsymbol{c}') d\boldsymbol{c}' \\
&= \int_{[\boldsymbol{c}-\delta, \boldsymbol{c}+\delta]} \mu^Y_{\pi_{\mathcal{S}^2}, \boldsymbol{C}=\boldsymbol{c}'} \, p_{\pi_{\mathcal{S}^2}}(\boldsymbol{C} = \boldsymbol{c}') d\boldsymbol{c}' + \int_{\mathcal{R}_{\boldsymbol{C}} \backslash [\boldsymbol{c}-\delta, \boldsymbol{c}+\delta]} \mu^Y_{\pi_{\mathcal{S}^2}, \boldsymbol{C}=\boldsymbol{c}'} \, p_{\pi_{\mathcal{S}^2}}(\boldsymbol{C} = \boldsymbol{c}') d\boldsymbol{c}' \\
&= \int_{[\boldsymbol{c}-\delta, \boldsymbol{c}+\delta]} \mu^Y_{\pi_{\mathcal{S}^1}, \boldsymbol{C}=\boldsymbol{c}'} \, p_{\pi_{\mathcal{S}^1}}(\boldsymbol{C} = \boldsymbol{c}') d\boldsymbol{c}' + \int_{\mathcal{R}_{\boldsymbol{C}} \backslash [\boldsymbol{c}-\delta, \boldsymbol{c}+\delta]} \mu^Y_{\pi^*_{\mathcal{S}^*}, \boldsymbol{C}=\boldsymbol{c}'} \, p_{\pi^*_{\mathcal{S}^*}}(\boldsymbol{C} = \boldsymbol{c}') d\boldsymbol{c}' \\
&< \int_{[\boldsymbol{c}-\delta, \boldsymbol{c}+\delta]} \mu^Y_{\pi^*_{\mathcal{S}^*}, \boldsymbol{C}=\boldsymbol{c}'} \, p_{\pi^*_{\mathcal{S}^*}}(\boldsymbol{C} = \boldsymbol{c}') d\boldsymbol{c}' + \int_{\mathcal{R}_{\boldsymbol{C}} \backslash [\boldsymbol{c}-\delta, \boldsymbol{c}+\delta]} \mu^Y_{\pi^*_{\mathcal{S}^*}, \boldsymbol{C}=\boldsymbol{c}'} \, p_{\pi^*_{\mathcal{S}^*}}(\boldsymbol{C} = \boldsymbol{c}') d\boldsymbol{c}' \\
&= \mu^Y_{\pi^*_{\mathcal{S}^*}},
\end{aligned}
$$

with contradicts the assumption that $(\mathcal{S}^*, \pi^*_{\mathcal{S}^*})$ is a solution to the fCGO problem. □

2 ALTERNATIVE KERNEL CONSTRUCTION

The kernel function $\kappa^{\xi}_{\mathcal{S}}$ introduced in Section 4.2 sets the covariance between the elements in the vector π_{func} associated to a DMP $\pi_{\mathcal{S}}$ to 0, thus restricting the type of functions that can be selected during optimization[1].

[1]Notice that, for hard interventions, this corresponds to limiting the range of values that can be set when intervening.



For instance, consider the graph on the right with $\mathcal{S} = \{\langle X, (C_1, C_2)\rangle, \langle Z, C_2\rangle\}$ and $\pi_{\mathcal{S}} = \{\pi_{X|\{C_1,C_2\}}, \pi_{Z|C_2}\}$. The proposed kernel function would set $\text{Cov}(\pi_{X|\{C_1,C_2\}}, \pi_{Z|C_2}) = 0$. While a study of the effect of choosing different covariance structures on the optimal target effect goes beyond the scope of this paper, in this section we provide alternative kernel constructions that relax this constraint.

Given a DMP $\pi_{\mathcal{S}}$, one can define the correlation between elements in π_{func} by introducing a $|\boldsymbol{C}_{\mathcal{S}}|$-dimensional vector $\boldsymbol{\omega}$ of parameters for each function $\pi_{X|\boldsymbol{C}_X}$ in π_{func} such that the j-th term $\omega_j = 1$ if the j-th term in $\boldsymbol{C}_{\mathcal{S}}$ is in $\boldsymbol{C}_X$ and $\omega_j = 0$ otherwise. For instance, for $\pi_{\mathcal{S}} = \{\pi_{X|\{C_1,C_2\}}, \pi_{Z|C_2}\} = \pi_{\text{func}}$, we have $\omega_1 = \omega_2 = 1$ for $\pi_{X|\{C_1,C_2\}}$ as both variables in $\boldsymbol{C}_{\mathcal{S}} = \{C_1, C_2\}$ are in $\boldsymbol{C}_X$, while $\omega_1 = 0$ and $\omega_2 = 1$ for $\pi_{Z|C_2}$ as only C_2 is in $\boldsymbol{C}_Z$.

We can then redefine $\kappa^{\xi}_{\mathcal{S}}$ to be an RBF kernel on an input space given by product between the the context variables and the $\boldsymbol{\omega}$ parameters. Denote by $\boldsymbol{\omega}^i, \boldsymbol{\omega}^j$ two possible values for the $\boldsymbol{\omega}$ vector, for instance we could have $\boldsymbol{\omega}^i = [1, 1]^\top$ and $\boldsymbol{\omega}^j = [0, 1]^\top$ in the example above; and by $\boldsymbol{c}^i = [c^i_1, \dots, c^i_{|\boldsymbol{C}_{\mathcal{S}}|}]^\top$ and $\boldsymbol{c}^j = [c^j_1, \dots, c^j_{|\boldsymbol{C}_{\mathcal{S}}|}]^\top$ two vector of values for $\boldsymbol{C}_{\mathcal{S}}$. We can define $\kappa^{\xi}_{\mathcal{S}} : (\mathcal{R}_{\boldsymbol{C}_{\mathcal{S}}} \times \Omega) \times (\mathcal{R}_{\boldsymbol{C}_{\mathcal{S}}} \times \Omega) \to \mathbb{R}^{|\mathcal{S}_{\text{func}}| \times |\mathcal{S}_{\text{func}}|}$ where Ω is the space of values for each vector $\boldsymbol{\omega}$ and $\kappa^{\xi}_{\mathcal{S}}((\boldsymbol{c}, \boldsymbol{\omega})^i, (\boldsymbol{c}, \boldsymbol{\omega})^j) = \kappa^{\xi}_{\mathcal{S}}((\boldsymbol{c}^i)^\top \boldsymbol{\omega}^i, (\boldsymbol{c}^j)^\top \boldsymbol{\omega}^j) = \gamma \exp(-0.5/l^2 \sum_{n=1}^{|\boldsymbol{C}_{\mathcal{S}}|} (c^i_n \omega^i_n - c^j_n \omega^j_n)^2)$ where $\xi = \{\gamma, l\}$. For the example above, we can write $\kappa^{\xi}_{\mathcal{S}}((\boldsymbol{c}^i)^\top \boldsymbol{\omega}^i, (\boldsymbol{c}^j)^\top \boldsymbol{\omega}^j) = \gamma \exp(-0.5/l^2[(c^i_1 \omega^i_1 - c^j_1 \omega^j_1)^2 + (c^i_2 \omega^i_2 - c^j_2 \omega^j_2)^2])$. When $\gamma \neq 0$, $\boldsymbol{\omega}^i = [1, 1]^\top$ and $\boldsymbol{\omega}^j = [0, 1]^\top$, this kernel would return a covariance between $\pi_{X|C_1,C_2}$ and $\pi_{Z|C_2}$ equal to $\kappa^{\xi}_{\mathcal{S}}((\boldsymbol{c}^i)^\top \boldsymbol{\omega}^i, (\boldsymbol{c}^j)^\top \boldsymbol{\omega}^j) = \gamma \exp(-0.5/l^2[(c^i_1)^2 + (c^i_2 - c^j_2)^2])$. The covariance would thus depend on the context values in the overlapping part of the context variables space and a correction term $(c^i_1)^2$. Instead of fixing the values in $\boldsymbol{\omega}$ to either zero or one based on the graph structure, one could think about optimizing the values that are different from zero so as to achieve a higher flexibility in terms of allowed covariance while still imposing structure via the zero values.

As a more general kernel construction, given a DMP $\mathcal{S}$, a vector of parameter values $\boldsymbol{\omega}^i$ and a vector of context values $\boldsymbol{c}^i = [c^i_1, \dots, c^i_{|\boldsymbol{C}_{\mathcal{S}}|}]^\top$, one could define the augmented input vector $\boldsymbol{c}^i_{\text{aug}} = [(\boldsymbol{c}^i)^\top \boldsymbol{\omega}^i, (\boldsymbol{c}^i)\boldsymbol{\omega}^i, t]^\top$ (and similarly for two alternative vector of values $\boldsymbol{c}^j$ and $\boldsymbol{\omega}^j$) given by the concatenation of two $|\boldsymbol{C}_{\mathcal{S}}|$-dimensional vector obtained by $(\boldsymbol{c}^i)^\top \boldsymbol{\omega}^i$ and a task index t that gives the index of the function in $\pi_{\mathcal{S}_{\text{func}}}$, similarly to what was introduced in Section 4.2.

For an augmented vector of hyper-parameters $\xi = [\gamma, l, \tilde{\gamma}, \tilde{l}]$, one could then define the following kernel:

$$\begin{aligned}
\kappa^{\xi}_{\mathcal{S}}(\boldsymbol{c}^i_{\text{aug}}, \boldsymbol{c}^j_{\text{aug}}) &= \mathbb{I}_{t=t'} \gamma^2 \exp\left(-\frac{0.5}{l^2} \sum_{n=1}^{|\boldsymbol{C}_{\mathcal{S}}|} (\boldsymbol{c}^i_{\text{aug},n} - \boldsymbol{c}^j_{\text{aug},n})^2\right) + \mathbb{I}_{t\neq t'} \tilde{\gamma}^2 \exp\left(-\frac{0.5}{\tilde{l}^2} \sum_{n=|\boldsymbol{C}_{\mathcal{S}}|+1}^{2|\boldsymbol{C}_{\mathcal{S}}|} (\boldsymbol{c}^i_{\text{aug},n} - \boldsymbol{c}^j_{\text{aug},n})^2\right) \\
&= \mathbb{I}_{t=t'} \gamma^2 \exp\left(-\frac{0.5}{l^2} \sum_{n=1}^{|\boldsymbol{C}_{\mathcal{S}}|} (c^i_n \omega_n - c^j_n \omega'_n)^2\right) + \mathbb{I}_{t\neq t'} \tilde{\gamma}^2 \exp\left(-\frac{0.5}{\tilde{l}^2} \sum_{n=|\boldsymbol{C}_{\mathcal{S}}|+1}^{2|\boldsymbol{C}_{\mathcal{S}}|} (c^i_n \omega_n - c^j_n \omega'_n)^2\right), \quad (1)
\end{aligned}$$

where c^i_n is the n-th term of the $\boldsymbol{c}^i$ vector (similarly for $\boldsymbol{c}^j$ and $\boldsymbol{\omega}^i$), and $\mathbb{I}_{t=t'}$ is an indicator function equal to one if $t = t'$ and zero otherwise. The first term in Eq. (1) represents an RBF kernel capturing the covariance structure *within* the t-th function in π_{func} while the second term is again an RBF kernel that captures the covariance *across* functions in π_{func}. Differently from the kernel described above we now have two sets of hyper-parameters: γ, l for the first RBF kernel and $\tilde{\gamma}, \tilde{l}$ for the second. This gives higher flexibility in terms of the functional interventions we can learn and thus the target effect values we can achieve. As in the previous kernel we can let the parameters in $\boldsymbol{\omega}$, as well as in ξ, change to capture different level of correlations or set them equal to one and zero depending on the structure of the graph. In the latter case and for the example introduced above, we would have $\omega_1 = \omega_2 = 1$ for $\pi_{X|C_1,C_2}$ which would lead to a standard RBF kernel for the first term in Eq. (1). We could then set $\tilde{\gamma} = 0$ to have a zero covariance across functions or finally vary ω_3 and ω_4 for both $\pi_{X|C_1,C_2}$ and $\pi_{Z|C_2}$ to allow for increasing level of correlation.

3 CHAIN EXPERIMENTS

For the CHAIN experiments we use the following SCM:

$$X = U_X,\ W = U_W,\ Z = -0.5X + U_Z,\ Y = -W - 3ZX + U_Y,\ \text{with } U_X, U_W, U_Z, U_Y \sim \mathcal{N}(0, 1).$$

We set the range for hard interventions on both Z and W to $[-1, 1]$. The set of non-redundant MPSs is $\mathbb{M}_\Sigma = \{\{\langle Z, \emptyset\rangle\}, \{\langle W, \emptyset\rangle\}, \{\langle Z, \emptyset\rangle, \langle W, \emptyset\rangle\}, \{\langle Z, \{X\}\rangle\}, \{\langle Z, \{X\}\rangle, \langle W, \emptyset\rangle\}\}$.

We set $\texttt{GridSize} = 10$ and represent each functional intervention with $N_\alpha = N_\beta = 10$ samples for the context variables. We sample the coefficients $\boldsymbol{\alpha}_i$ (for $i = 1, \ldots, N_\alpha$) and $\boldsymbol{\beta}_j$ (for $j = 1, \ldots, N_\beta$) uniformly in the interval $[-0.27, 0.27]$, in order to keep the range of values obtained for the intervened variables following a functional intervention similar to the ranges set for the hard interventions. For each $\mathcal{S} \in \mathbb{M}_\Sigma$, we initialize the linear kernel $\kappa^\xi_\mathcal{S}$ with $\xi = 1$. Exploration is hard to achieve when the GP models for $\mathcal{S}$ including functional interventions are initialized with RBF $K^\theta_\mathcal{S}$ and hyper-parameters $\theta = (\ell, \sigma^2_f) = (1, 1)$. We thus perform hyper-parameters search exploring continuous values $\sigma^2_f \in [1, 10000]$ and $\ell \in [1, 30]$, which results in selecting $\sigma^2_f = 7000$, and $\ell = 20$ for both fCBO and BFO. For CBO and BO, which consider only hard interventions and thus do not suffer from exploration issues, we initialize $K^\theta_\mathcal{S}$ with $\theta = (1, 1)$. For MCBO we use the default setting (Matérn $5/2$ kernel), as it is not possible to tune the kernel and corresponding hyper-parameters. In order to run MCBO with contextual interventions, we use the augmented SCM with action variables $X = U_X$, $W = U_W + A_W$, $Z = -0.5X + U_Z + A_Z$, $Y = -W - 3ZX + U_Y$. In this setting, the average CPU execution time for a single fCBO run is $\sim$ 6 minutes, while for a single MCBO run is $\sim$ 14 minutes.

4 HEALTH EXPERIMENTS

For the HEALTH experiments, we use the SCM from Ferro et al. [2015]:

$$
\begin{aligned}
&\text{Age} = U_{\text{Age}}, \text{CI} = U_{\text{CI}}, \text{BMR} = 1500 + 10 \times U_{\text{BMR}},\\
&\text{Height} = 175 + 10 \times U_{\text{Height}},\\
&\text{Weight} = \frac{\text{BMR} + 6.8 \times \text{Age} - 5 \times \text{Height}}{13.7 + \text{CI} \times 150/7716},\\
&\text{BMI} = \text{Weight}/(\text{Height}/100)^2,\\
&\text{Aspirin} = \sigma(-8 + 0.1 \times \text{Age} + 0.03 \times \text{BMI}),\\
&\text{Statin} = \sigma(-13 + 0.1 \times \text{Age} + 0.2 \times \text{BMI}),\\
&\text{PSA} = 6.8 + 0.04 \times \text{Age} - 0.15 \times \text{BMI} - 0.6 \times \text{Statin} + 0.55 \times \text{Aspirin}\\
&\qquad + \sigma(2.2 - 0.05 \times \text{Age} + 0.01 \times \text{BMI} - 0.04 \times \text{Statin} + 0.02 \times \text{Aspirin}) + U_{\text{PSA}},
\end{aligned}
$$

with $U_{\text{Age}} \sim \mathcal{U}(55, 75)$, $U_{\text{CI}} \sim \mathcal{U}(-100, 100)$, $U_{\text{BMR}} \sim t\mathcal{N}(-1, 2)$, $U_{\text{Height}} \sim t\mathcal{N}(-0.5, 0.5)$, $U_{\text{PSA}} \sim \mathcal{N}(0, 0.4)$, where $\mathcal{U}(\cdot, \cdot)$ denotes a uniform distribution, $t\mathcal{N}(a, b)$ a standard Gaussian distribution truncated between a and b, and $\sigma(\cdot)$ the sigmoidal transformation defined as $\sigma(x) = \frac{1}{1+\exp(-x)}$.

We set the ranges for hard interventions on Aspirin, Statin, and CI to $[0.1, 1]$. The set of non-redundant MPSs is $\mathbb{M}_\Sigma = \{\{\langle \text{Aspirin}, \emptyset\rangle\}, \{\langle \text{Statin}, \emptyset\rangle\}, \{\langle \text{CI}, \emptyset\rangle\}, \{\langle \text{Aspirin}, \emptyset\rangle, \langle \text{Statin}, \emptyset\rangle\}, \{\langle \text{Aspirin}, \emptyset\rangle, \langle \text{CI}, \emptyset\rangle\}, \{\langle \text{Statin}, \emptyset\rangle, \langle \text{CI}, \emptyset\rangle\}, \{\langle \text{Aspirin}, \emptyset\rangle, \langle \text{Statin}, \emptyset\rangle, \langle \text{CI}, \emptyset\rangle\}, \{\langle \text{Aspirin}, \{\text{Age}, \text{BMI}\}\rangle\}, \{\langle \text{Statin}, \{\text{Age}, \text{BMI}\}\rangle\}, \{\langle \text{Aspirin}, \{\text{Age}, \text{BMI}\}\rangle, \langle \text{Statin}, \{\text{Age}, \text{BMI}\}\rangle\}, \{\langle \text{Aspirin}, \{\text{Age}, \text{BMI}\}\rangle, \langle \text{Statin}, \emptyset\rangle\}, \{\langle \text{Aspirin}, \emptyset\rangle, \langle \text{Statin}, \{\text{Age}, \text{BMI}\}\rangle\}, \{\langle \text{Aspirin}, \{\text{Age}, \text{BMI}\}\rangle, \langle \text{CI}, \emptyset\rangle\}, \{\langle \text{Statin}, \{\text{Age}, \text{BMI}\}\rangle, \langle \text{CI}, \emptyset\rangle\}, \{\langle \text{Aspirin}, \{\text{Age}, \text{BMI}\}\rangle, \langle \text{Statin}, \{\text{Age}, \text{BMI}\}\rangle, \langle \text{CI}, \emptyset\rangle\}, \{\langle \text{Aspirin}, \emptyset\rangle, \langle \text{Statin}, \{\text{Age}, \text{BMI}\}\rangle, \langle \text{CI}, \emptyset\rangle\}, \{\langle \text{Aspirin}, \{\text{Age}, \text{BMI}\}\rangle, \langle \text{Statin}, \emptyset\rangle, \langle \text{CI}, \emptyset\rangle\}\}$.

We represent each functional intervention with $N_\alpha = N_\beta = 10$ samples for the context variables. We sample the coefficients $\boldsymbol{\alpha}_i$ (for $i = 1, \ldots, N_\alpha$) and $\boldsymbol{\beta}_j$ (for $j = 1, \ldots, N_\beta$) uniformly in the interval $[0, 3.3]$, in order to keep the total cost of functional interventions and hard interventions comparable. The RBF kernels $K^\theta_\mathcal{S}$ and $\kappa^\xi_\mathcal{S}$ are initialized with $\theta = (1, 1)$ and $\xi = (1, 1)$ for each $\mathcal{S} \in \mathbb{M}_\Sigma$. In this setting, the average CPU execution time for a single fCBO run is $\sim$ 3 hours and 20 minutes, while for a single MCBO run is $\sim$ 10 hours.