Preference-Guided Bayesian Optimization for Control Policy Learning: Application to Personalized Plasma Medicine

Ketong Shao KETONG_SHAO@BERKELEY.EDU
Department of Chemical & Biomolecular Engineering
University of California, Berkeley, CA 94720, USA

Diego Romeres ROMERES@MERL.COM
Mitsubishi Electric Research Laboratories (MERL)
Cambridge, MA 02139, USA

Ankush Chakrabarty ACHAKRABARTY@IEEE.ORG
Mitsubishi Electric Research Laboratories (MERL)
Cambridge, MA 02139, USA

Ali Mesbah MESBAH@BERKELEY.EDU
Department of Chemical & Biomolecular Engineering
University of California, Berkeley, CA 94720, USA

Abstract

This paper investigates the adaptation of control policies for personalized dose delivery in plasma medicine using preference-learning based Bayesian optimization. Preference learning empowers users to incorporate their preferences or domain expertise during the exploration of optimal control policies, which often results in fast attainment of personalized treatment outcomes. We establish that, compared to multi-objective Bayesian optimization (BO), preference-guided BO offers statistically faster convergence and computes solutions that better reflect user preferences. Moreover, it enables users to actively provide feedback during the policy search procedure, which helps to focus the search in sub-regions of the search space likely to contain preferred local optima. Our findings highlight the suitability of preference-learning-based BO for adapting control policies in plasma treatments, where both user preferences and swift convergence are of paramount importance.

Keywords: Bayesian optimization; Preference learning; Personalized plasma medicine.

1. Introduction

A real-world challenge associated with learning optimal control policies stems from its "evaluate-to-know" nature, i.e., the true performance of a control policy becomes evident only after it has been applied. Thus, control policy learning can be naturally cast as a black-box optimization problem. Policy gradient methods (Silver et al., 2014) have emerged as a popular approach for control policy learning within the realm of continuous control-input spaces, especially for model predictive control (MPC) policies (e.g., (Zanon and Gros, 2020)). One of the ways to view policy search is to optimize the parameters of the policy, which can be approximated by deep neural networks (Levine and Koltun, 2013; Sehnke et al., 2010). While policy-gradient reinforcement learning (RL) offers scalability, it often lags in data efficiency, especially with poor initialization. In contrast, Bayesian opti-

mization (BO) provides a powerful framework for data-efficient policy search, particularly when dealing with limited performance data or interactions with real-world environments (Paulson et al., 2023b). BO, a derivative-free and probabilistic method for "global" optimization (Shahriari et al., 2015), is well-suited to handle a combination of continuous, discrete, and categorical decision variables, and BO algorithms have proven to be highly customizable for controller optimization in real-world applications such as energy systems, robotics, and manufacturing (Chakrabarty et al., 2021; Paulson et al., 2023a; Marco et al., 2017; Koenig et al., 2023; Rothfuss et al., 2023; Hoang et al., 2023).

Control policy search often requires users to assess the quality of learned policies, especially when multiple conflicting objectives are at play. BO seamlessly accommodates the multi-objective nature of policy search when the goal is to discover a set of optimal policies with conflicting objectives (Makrygiorgos et al., 2022; Turchetta et al., 2020). Multi-objective BO (MOBO) methods may expend unnecessary efforts in search regions with no preferred optimal solutions. Moreover, user expertise is typically leveraged only in the final stage of policy selection. To overcome these challenges and maximize the utilization of a user's knowledge and preferences regarding closed-loop performance objectives, preference-learning-based methods have been proposed. An approach is to construct the utility function directly in terms of control policy, known as preferential BO (Eric et al., 2007; Brochu, 2010; González et al., 2017; Siivola et al., 2021). While these methods avoid the need for performance outcome information, they may pose challenges for non-expert users in distinguishing between two control policies. Lin et al. (2022) proposed preference exploration BO to address scenarios with multiple outcomes. This approach uses two Gaussian processes to learn input-to-outcome and outcome-to-utility mappings, with alternating knowledge enhancement in the different stages.

This paper presents a preferential BO strategy for adaptive deep learning-based approximate MPC for preference-based and personalized plasma medicine. We demonstrate that the proposed preference-guided BO outperforms classical MOBO strategies, and enables users to efficiently adjust control policies based on their preferences. This is a critical step towards expedited and effective plasma treatments in the context of biomedical applications.

2. Dose Delivery in Plasma Medicine

2.1 Atmospheric Pressure Plasma Jet (APPJ)

We use a kHz-excited atmospheric pressure plasma jet (APPJ) in helium (He), with prototypical applications for treatment of heat- and pressure-sensitive (bio)materials. Schematic of the APPJ is shown in Appendix A; see (Gidon et al., 2019) for a detailed description. The manipulated inputs include the applied power P and He flow rate q, while the measured outputs are the maximum surface temperature T and the total optical intensity I of plasma at its incident point with the surface. The APPJ dynamics are described by a linear time-invariant state-space model identified from input-output data (Chan et al., 2023). Here, we look to control the delivery of thermal effects of plasma to a target surface (e.g., patient's skin). To quantify the delivered thermal effects to a surface, we use the so-called cumulative equivalent minutes (CEM) metric defined as

$$\mathrm{CEM}(k+1) = \mathrm{CEM}(k) + K^{(T_{\mathrm{ref}} - T(k))}\delta t, \tag{1}$$

where constant K is governed by physical properties of surface, T_{ref} is a reference temperature (43°C), and δt is the sampling time (Sapareto and Dewey, 1984; Gidon et al., 2017). As such, the overall system dynamics are described by a nonlinear discrete-time model

$$x(k+1) = f(x(k), u(k), w(k)), \tag{2}$$

where $x = [T, I, \text{CEM}]^\top$ are the overall system states, $u = [P, q]^\top \in \mathbb{R}^2$ are the manipulated inputs, and w is a stochastic variable encapsulating system uncertainties.

In pursuit of the goal of delivering a specific thermal dose (CEM) within the shortest possible time frame, all while adhering to a crucial safety constraint concerning surface temperature (T), which is vital for an individual's comfort and well-being, we employ a robust MPC strategy that takes into consideration uncertainties within the system. Subsequently, the optimal state-input data are harnessed to train a deep neural network (DNN) policy. It is because the computational demands and memory requirements associated with the control policy generated by the robust MPC present a significant challenge when embedding the controller on cost-effective, resource-constrained hardware (Bonzanini et al., 2021). Embedded control is essential for the operation of point-of-use plasma biomedical devices like APPJs; see Appendix B for further details on the DNN control policy.

2.2 Preference-Guided Control Policy Learning for Personalized Dose Delivery

Adapting the treatment protocol for individual subjects is essential for personalized plasma medicine in order to enhance the therapeutic efficacy of treatment without compromising the safety and comfort of patients. In particular, the protocol adaptation will enable accounting for the variability among different target surfaces (i.e., patients), as well as the time-varying nature of the plasma and surface characteristics during successive treatments. However, two main challenges arise in adapting the control policy (8): (i) a limited number of treatments/trials can be performed in a biomedical context, which makes data efficiency a prerequisite for control policy adaptation; and (ii) there do not exist closed-form expressions for the mappings between control policy parameters, θ, and user-defined "performance measures" that quantify the efficacy and safety of a plasma treatment. In addition, the treatment efficacy is typically defined in terms of multiple, possibly conflicting, performance measures that are often observed only at the end of a treatment.

In this work, the patient-to-patient variability stems from the value of constant K in the CEM dose (1). Generally, K is estimated for a population of subjects and, thus, follows a probability distribution. We look to adapt the control policy parameters θ during successive plasma treatments to tailor the treatment to an individual subject with a fixed, but unknown value of K. We seek to realize two objectives pertaining to patient comfort and safety via adapting the policy (8), as quantified by the following performance objectives. We aim to minimize the treatment time

$$\psi_1 = \tau_p, \tag{3}$$

while concurrently minimizing the cumulative surface temperature constraint violation cost

$$\psi_2 = \sum_{k=0}^{N} ([T(k) - T_{\text{tol}}]^+)^2, \tag{4}$$

where T_{tol} is the nominal tolerated temperature constraint, $[\cdot]^+$ denotes the positive part of the function, and N is the total number of surface temperature measurements over the treatment. We note that minimizing the treatment time, essential for improved patient experience and comfort, would involve aggressive thermal dose delivery to the surface, which would in turn lead to significant violations of the surface temperature constraint that is unsafe. *Hence, the control policy adaptation problem must naturally be cast as a multi-objective optimization problem to trade-off between these two conflicting objectives.*

In practice, the degree of trade-off between these two objectives can be informed by a user's preferences (i.e., physician's expertise). This leads to the notion of preference learning (Houlsby et al., 2011; Obeng and Bakshy, 2020), where a user's preferences guide the search for decisions that result in the most desirable outcomes. In the context of the above plasma treatment problem with multiple objectives, preference learning can be formulated as a single-objective optimization problem to maximize a user-defined score. Formally, the preference learning problem with multiple outcomes can be defined as

$$\max_{\theta \in \Theta} u(\boldsymbol{f}(\theta)), \tag{5}$$

where $\Theta \subset \mathbb{R}^{n_\theta}$ represents the space for policy parameters θ, $\boldsymbol{f} : \Theta \to \Psi \subset \mathbb{R}^{n_\psi}$ is a multi-outcome function, and $u : \Psi \to \mathbb{U} \subset \mathbb{R}$ is the utility function.

Here, we formulate the utility function based on the performance metrics described in equations (3) and (4). Our premise is grounded in the assumption that an ideal plasma treatment, from the user's standpoint, should ideally have a duration of 30 seconds while maintaining zero cumulative temperature violations. Consequently, the utility function is:

$$u = -\alpha|\psi_1 - 30| - \beta|\psi_2| = -\alpha|t_p - 30| - \beta \sum_{k=0}^{N} ([T(k) - T_{\text{tol}}]^+)^2. \tag{6}$$

In this expression, α and β serve as weight parameters for ψ_1 and ψ_2, respectively. It's essential to clarify that this utility function's primary role is to facilitate the comparison of outcomes from the user's perspective. Notably, the black-box optimization technique detailed below operates with no prior knowledge of the utility function, reflecting the fact that the preference should be learned during the algorithm.

3. Co-Active Preference Learning Bayesian Optimization

We now introduce a preference learning BO strategy that can accommodate user preferences in control policy adaptation. The proposed strategy iteratively loops over two main stages called *preference exploration* (PE) stage and *experimentation* (EXP) stage to learn the utility function u and the multi-outcome function $\boldsymbol{f}$, respectively, until the optimal parameters θ^* of the control policy are obtained. The main purpose is to provide the patient or the physician an intuitive way to adapt the treatment accordingly to their standard of comfort. The outcomes time (3) and temperature (4) are easily interpretable by humans and our method allows the user to propose ideal outcomes, accordingly to their preference, that are used by the optimization algorithm to guide the search of the optimal parameters.

The method is summarized in Appendix C in Algorithm 1. During each iteration of the PE stage M pairs of possible outcomes $\boldsymbol{\psi}_{1,m}$ and $\boldsymbol{\psi}_{2,m}$ are presented to the user to

obtain a pairwise preference $r(\boldsymbol{\psi}_{1,m}, \boldsymbol{\psi}_{2,m})$. Moreover, the user is asked to provide a *desired outcome*, $\boldsymbol{\psi}_{3,m}$, accordingly to their own preference. We can therefore assume that $u(\boldsymbol{\psi}_{3,m}) > u(\boldsymbol{\psi}_{1,m}), u(\boldsymbol{\psi}_{2,m})$ since the user provides the feedback based on their best interest. This step is what we define the *co-active feedback* because both the user and the algorithm are actively suggesting new outcomes to achieve reachable solutions that satisfy the user's preference. The outcome pairs are obtained by optimizing the acquisition function Expected Utility of Best Option (EUBO) that was proposed in (Lin et al., 2022). The model of the utility function, $\hat{u}$, is a pairwise Gaussian process trained on the comparisons $\{(\boldsymbol{\psi}_{i,m}, \boldsymbol{\psi}_{j,m}), r(\boldsymbol{\psi}_{i,m}, \boldsymbol{\psi}_{j,m})\}$ with $i, j \in \{1, 2, 3\} \vee i \neq j$ using a probit likelihood as described in Chu and Ghahramani (2005).

Next, at each b^{th} iteration of the EXP stage N sets of optimal parameters $\boldsymbol{\theta}_{1:N,b}$ are obtained based on the current utility and multi-outcome models by optimizing the batch expected improvement under utility uncertainty (qEIUU) proposed in (Astudillo and Frazier, 2020; Lin et al., 2022). The optimal parameters are then tested on the plasma treatment scenario (2) to obtain the real outcomes $\boldsymbol{\psi}_{1:N,b}$, and make the user decide if the optimization has produced the parameters θ^* that satisfy their preferences. The model of the outcome function $\hat{\boldsymbol{f}}$ is a multi-output Gaussian process trained based on the outcomes data $(\theta_{i,b}, \boldsymbol{\psi}_{i,b})$.

The proposed method draws inspiration from (Lin et al., 2022), which however does not incorporate the co-active feedback.

4. Case Study and Results

The DNN control policy in Section 2, a fully-connected feedforward DNN with $L = 4$ number of hidden layers, $H = 7$ number of nodes for each hidden layer and ReLU activation functions, is trained using $n_s = 5000$ samples of state-to-optimal-input mappings using PyTorch (Paszke et al., 2019). The nominal control policy before adaptation is obtained with parameters $K_{\text{pop}} = 0.5$ and $T_{\text{tol,pop}} = 45°\text{C}$ in (1) assuming these are the population mean requirements. In Marelli and Sudret (2014) is shown that the customization to a new patient can be minimizing the final layer of the DNN policy, which consists of $|\theta| = 14$ parameters including weights and biases. Notice that the DNN-policy nearly matches the MPC control law's performance while being approximately 1,000 times faster on a standard CPU (2.4 GHz quad-core Intel i5 processor).

Assume that the comfort of a new patient is expressed by the ideal parameters $K = 0.55$ in (1), $T_{\text{tol}} = 45.5°\text{C}$ in (4) and $\alpha = 1$ and $\beta = 1000$ in (6) which are unknown to the optimization algorithm. To avoid infinite-time treatment due to consistent low temperature, the treatment will be terminated if the setpoint CEM_{sp} cannot be reached after 120 seconds. Therefore, $\tau_{p,\max} = 120$. In Algorithm 1, the number of pairwise comparisions in the PE stage is $M = 1$ and the number of experiments in the EXP stage is $N = 3$. Selecting large values for either N or M carries the risk of making decisions during the preference exploration or exploration stages when lacking of accurate surrogate models. In the context of plasma treatment, our goal is to ensure that the information for both $\boldsymbol{f}$ and u grows equitably. The models $\hat{\boldsymbol{f}}$ and $\hat{u}$ are initially trained with 5 initial datapoints for $\mathcal{D}_0 = \{(\boldsymbol{\psi}_i, \theta_i)\}_{i=1}^5$ and 3 pairwise comparisons for $\mathcal{P}_0 = \{(\boldsymbol{\psi}_{1,j}, \boldsymbol{\psi}_{2,j}, r(\boldsymbol{\psi}_{1,j}, \boldsymbol{\psi}_{2,j}))\}_{j=1}^3$ constructed via uniform random sampling from Θ and Ψ. The co-active feedback from the user is

simulated to give an outcome that is closer than $\boldsymbol{\psi}_1, \boldsymbol{\psi}_1$ to the ideal $\boldsymbol{\psi}^*$ of the patient:

$$\boldsymbol{\psi}_3 = \boldsymbol{\psi}^* + \epsilon \boldsymbol{\psi}^+, \tag{7a}$$

$$\boldsymbol{\psi}^+ = [\boldsymbol{\psi}^* - \boldsymbol{\psi}_1, \boldsymbol{\psi}^* - \boldsymbol{\psi}_2]_{\mathcal{I}_\psi}, \tag{7b}$$

$$\mathcal{I}_\psi = \text{argmin}\{\text{abs}(\boldsymbol{\psi}^* - \boldsymbol{\psi}_1), \text{abs}(\boldsymbol{\psi}^* - \boldsymbol{\psi}_2)\}, \tag{7c}$$

where $\mathcal{I}_\psi$ represents the index vector resulting from element-wise arg-minimization, with argmin and abs applied element-wise. $\boldsymbol{\psi}^+$ is the improvement vector gathered from two vectors according to $\mathcal{I}_\psi$. The parameter ϵ determines the extent to which the new recommendation $\boldsymbol{\psi}_3$ is close to $\boldsymbol{\psi}^*$ or to the suggested $\boldsymbol{\psi}_{1,2}$. We set ϵ to 0.3. In addition, we applied 10% probability that a user will make a wrong preference feedback on a given outcome pair $\{\boldsymbol{\psi}_1, \boldsymbol{\psi}_2\}$. This reflects the fact that human beings are usually unreliable and a correct judgement based on preference is often noisy in plasma medicine.

For comparison, we consider the following four methods: (i) random sampling, (ii) multi-objective BO using qNEHVI (MOBO), the batch of which is also 3 for consistency, (iii) the proposed method without co-active feedback (Preference Learning BO), and (iv) the proposed method in Section 3 (Co-active Preference Learning BO). Each method is repeated 256 times, with different $\mathcal{D}_0$ and $\mathcal{P}_0$ to initialize the outcome and utility surrogate model, and $Q = 20$ loops of the PE and EXP stages. As can be seen on the left of

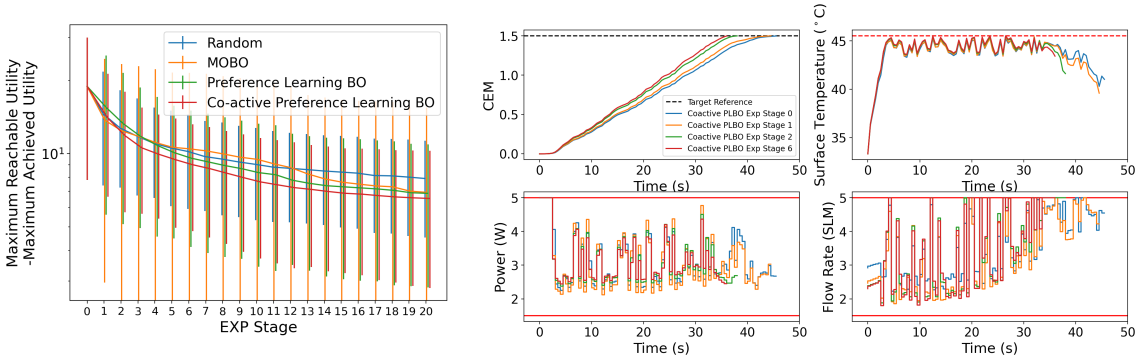

Figure 1: Left: Comparison of the best achieved utility performance at the experimentation stages. Right: State and input profiles of closed-loop experiments at various experimentation stages of co-active preference-learning BO.

Fig. 1, preference-based BO algorithm outperforms both random and MOBO methods. In addition, when considering the co-active feedback higher utility (lower distance to best reachable utility) can be achieved in a lower number of iterations. This means that under the proper guidance of a user via co-active feedback, preference-learning BO can leverage the user's knowledge to enable faster identification of high preference region. We further look into the detailed performance of the adapted closed-loop treatment. Fig. 1 right shows how the inputs and states profiles evolve as more experimentation stages are executed in co-active preference-learning BO. It can be seen that this method can quickly reduce the required treatment time while still satisfying the temperature constraint. We validated the proposed method also in benchmark problems as shown in Appendix D.

References

Raul Astudillo and Peter Frazier. Multi-attribute bayesian optimization with interactive preference learning. In *International Conference on Artificial Intelligence and Statistics*, pages 4496–4507. PMLR, 2020.

Angelo D Bonzanini, Joel A Paulson, David B Graves, and Ali Mesbah. Toward safe dose delivery in plasma medicine using projected neural network-based fast approximate NMPC. *IFAC-PapersOnLine*, 53(2):5279–5285, 2020.

Angelo D Bonzanini, Ketong Shao, Augusto Stancampiano, David B Graves, and Ali Mesbah. Perspectives on machine learning-assisted plasma medicine: Toward automated plasma treatment. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 6(1): 16–32, 2021.

Eric Brochu. *Interactive Bayesian optimization: learning user preferences for graphics and animation*. PhD thesis, University of British Columbia, 2010.

Ankush Chakrabarty, Claus Danielson, Scott A Bortoff, and Christopher R Laughman. Accelerating self-optimization control of refrigerant cycles with bayesian optimization and adaptive moment estimation. *Applied Thermal Engineering*, 197:117335, 2021.

Kimberly J Chan, Georgios Makrygiorgos, and Ali Mesbah. Towards personalized plasma medicine via data-efficient adaptation of fast deep learning-based MPC policies. In *Proceedings of the American Control Conference*, pages 2769–2775, 2023.

Wei Chu and Zoubin Ghahramani. Preference learning with gaussian processes. In *Proceedings of the 22nd international conference on Machine learning*, pages 137–144, 2005.

Brochu Eric, Nando Freitas, and Abhijeet Ghosh. Active preference learning with discrete choice data. *Advances in neural information processing systems*, 20, 2007.

Dogan Gidon, David B Graves, and Ali Mesbah. Effective dose delivery in atmospheric pressure plasma jets for plasma medicine: A model predictive control approach. *Plasma Sources Science and Technology*, 26(8):085005, 2017.

Dogan Gidon, David B Graves, and Ali Mesbah. Predictive control of 2D spatial thermal dose delivery in atmospheric pressure plasma jets. *Plasma Sources Science and Technology*, 28(8):085001, 2019.

Javier González, Zhenwen Dai, Andreas Damianou, and Neil D Lawrence. Preferential bayesian optimization. In *International Conference on Machine Learning*, pages 1282–1291. PMLR, 2017.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

K Tuan Hoang, Sjoerd Boersma, Ali Mesbah, and Lars Imsland. Heteroscedastic bayesian optimisation for active power control of wind farms. *IFAC-PapersOnLine*, 2023.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Christopher Koenig, Miks Ozols, Anastasia Makarova, Efe C Balta, Andreas Krause, and Alisa Rupenyan. Safe risk-averse bayesian optimization for controller tuning. *arXiv preprint arXiv:2306.13479*, 2023.

Sergey Levine and Vladlen Koltun. Guided policy search. In *International conference on machine learning*, pages 1–9. PMLR, 2013.

Zhiyuan Jerry Lin, Raul Astudillo, Peter Frazier, and Eytan Bakshy. Preference exploration for efficient bayesian optimization with multiple outcomes. In *International Conference on Artificial Intelligence and Statistics*, pages 4235–4258. PMLR, 2022.

Georgios Makrygiorgos, Angelo D Bonzanini, Victor Miller, and Ali Mesbah. Performance-oriented model learning for control via multi-objective bayesian optimization. *Computers & Chemical Engineering*, 162:107770, 2022.

Alonso Marco, Felix Berkenkamp, Philipp Hennig, Angela P Schoellig, Andreas Krause, Stefan Schaal, and Sebastian Trimpe. Virtual vs. real: Trading off simulations and physical experiments in reinforcement learning with bayesian optimization. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1557–1563. IEEE, 2017.

Stefano Marelli and Bruno Sudret. Uqlab: A framework for uncertainty quantification in matlab. In *Vulnerability, uncertainty, and risk: quantification, mitigation, and management*, pages 2554–2563. 2014.

Adam Obeng and Eytan Bakshy. Preference learning for real-world multi-objective decision making. In *ICML 2020 Workshop on Real World Experiment Design and Active Learning*, 2020.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Joel A Paulson, Farshud Sorouifar, Christopher R Laughman, and Ankush Chakrabarty. LSR-BO: Local search region constrained Bayesian optimization for performance optimization of vapor compression systems. In *Proceeding of the American Control Conference*, pages 576–582, 2023a.

Joel A Paulson, Farshud Sorourifar, and Ali Mesbah. A tutorial on derivative-free policy learning methods for interpretable controller representations. In *Proceedings of the American Control Conference*, pages 1295–1306, 2023b.

Jonas Rothfuss, Christopher Koenig, Alisa Rupenyan, and Andreas Krause. Meta-learning priors for safe bayesian optimization. In *Conference on Robot Learning*, pages 237–265. PMLR, 2023.

Stephen A Sapareto and William C Dewey. Thermal dose determination in cancer therapy. *International Journal of Radiation Oncology* Biology* Physics*, 10(6):787–800, 1984.

Frank Sehnke, Christian Osendorfer, Thomas Rückstieß, Alex Graves, Jan Peters, and Jürgen Schmidhuber. Parameter-exploring policy gradients. *Neural Networks*, 23(4): 551–559, 2010.

Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2015.

Eero Siivola, Akash Kumar Dhaka, Michael Riis Andersen, Javier González, Pablo García Moreno, and Aki Vehtari. Preferential batch bayesian optimization. In *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2021.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International conference on machine learning*, pages 387–395. Pmlr, 2014.

Matteo Turchetta, Andreas Krause, and Sebastian Trimpe. Robust model-free reinforcement learning with multi-objective bayesian optimization. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10702–10708. IEEE, 2020.

Mario Zanon and Sébastien Gros. Safe reinforcement learning using robust mpc. *IEEE Transactions on Automatic Control*, 66(8):3638–3652, 2020.

Appendix A: Experimental Setup

In this work, a prototypical atmospheric pressure plasma jet is employed to simulate the cold plasma treatment process in plasma medicine. Briefly, plasma is generated in a quartz tube by applying high-frequency AC voltage to a copper electrode, wrapped around the tube, as He flows through. The plasma is then directed onto a grounded metal plate covered with a substrate (here, glass), situated 3 mm below the APPJ's tip. The manipulated inputs include the applied power P and He flow rate q, while the measured outputs are the maximum surface temperature T and the total optical intensity I of plasma at its incident point with the surface. Measurements are taken every 0.5 seconds. Control of the plasma is achieved via Arduino manipulating the flowrate of Helium and the applied power. The data for training the DNN policy are also collected using this setup.

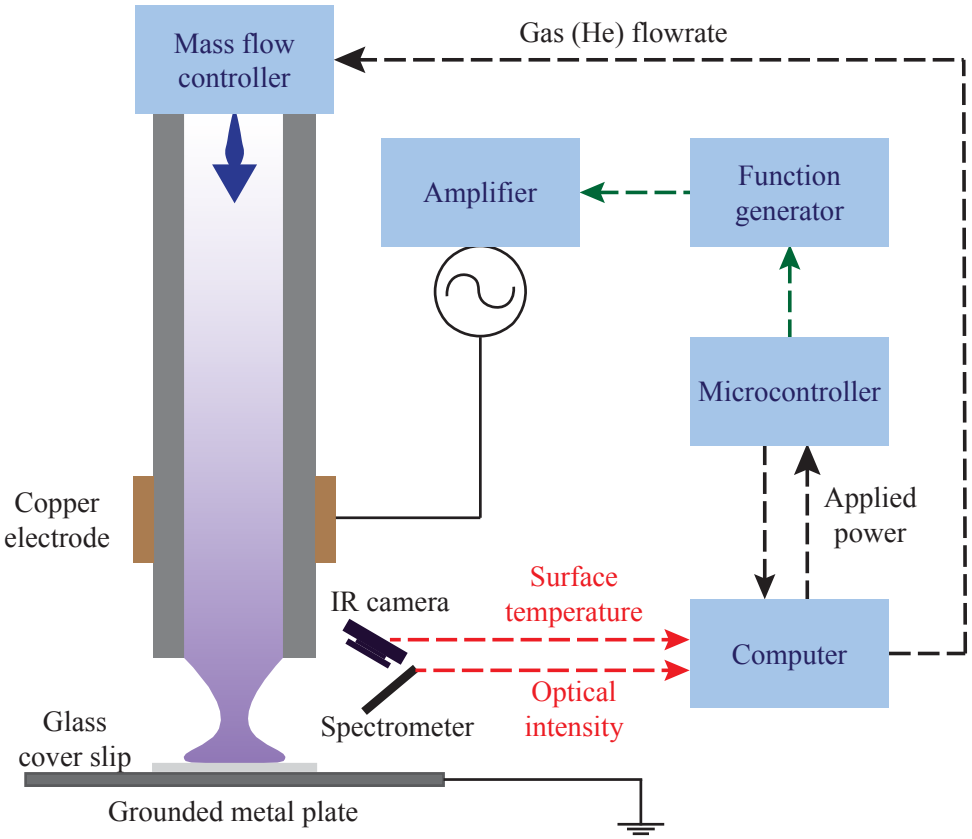

Figure 2: Schematic of the kHz-excited atmospheric pressure plasma jet in helium. The manipulated inputs are denoted in black dashed lines and the measured outputs are denoted in red dashed lines.

Appendix B: Safe and Fast Control of Thermal Dose Delivery

Our objective is to deliver a desired amount of thermal dose CEM in minimal time without violating a safety-critical constraint on surface temperature T that corresponds to a person's comfort and safety. Additionally, in order to ensure a safe plasma treatment, we look to systematically account for system uncertainties, modeled by w in (2), in the control problem formulation. To this end, the control problem is cast as a robust MPC problem based on the system model (2) with the terminal control objective $V(\tau_p) = \big(\mathrm{CEM}_{\mathrm{sp}} - \mathrm{CEM}(\tau_p)\big)^2$, where $\mathrm{CEM}_{\mathrm{sp}}$ is a user-specified thermal dose setpoint, τ_p is a given treatment time, and the surface temperature constraint is $T \leq 45°\mathrm{C}$; c.f. Chan et al. (2023) for more details.

The computational cost and memory footprint of the control policy obtained via the above-described robust MPC pose a key challenge to controller implementation on low-cost, resource-limited hardware required for point-of-use plasma biomedical devices such as APPJs. Thus, the robust MPC policy is approximated by a deep neural network (DNN) using *in-silico* data collected by solving the controller in closed-loop with the model (2) (Chan et al., 2023). The basic structure of the DNN-based control policy consists of L hidden layers with H nodes, creating a nonlinear mapping that transfers information from input (i.e., measured states) to output (i.e., control inputs) via feedforward propagation (Goodfellow et al., 2016). As such, given a measured system state z, the control policy used in this work takes the form of

$$\Pi\left(z;\theta_0,\mathcal{C}\right) = \boldsymbol{W}_{L+1} \circ \left(\sigma_L \circ \boldsymbol{W}_L\right) \circ \cdots \circ \left(\sigma_1 \circ \boldsymbol{W}_1\right)(z), \tag{8}$$

where $\theta_0 = \{\boldsymbol{W}_i\}_{i=1}^{L+1}$ denotes the DNN parameters obtained by minimizing a mean-squared-error loss function using the *in-silico* closed-loop data. In (8), $\circ$ symbolizes composition; $\mathcal{C}$ represents the hyperparameters; and $\boldsymbol{W}_i$ includes the weight matrix and bias for the i^{th} and $(i+1)^{\mathrm{th}}$ layers, with σ_i denoting the activation function. The DNN-based control policy (8) enables fast control computations as a function of measured system states at kHz sampling rates (Bonzanini et al., 2020). However, (8) is trained offline based on a plasma dose model established for a population of subjects; that is, the constant K in (1) is estimated for a population, which can hinder therapeutically effective plasma treatment of individual subjects. Thus, it is imperative to adapt the control policy (8), namely parameters θ_0, in a run-to-run manner to enable tailoring the plasma treatment to each individual subject.

Appendix C: Pseudocode of the Co-active Preference Learning BO

Algorithm 1: Co-Active Preference Exploration Bayesian Optimization

Input: $\mathcal{D}_0, \mathcal{P}_0, \Theta$
Output: θ^*
Parameter: N, M, Q
0. Initialization
Train $\hat{\boldsymbol{f}}(\cdot|\mathcal{D}_0)$ the multi-outcome model;
Train $\hat{u}(\cdot|\mathcal{P}_0)$ the utility model;
Set $\mathcal{D} = \mathcal{D}_0, \mathcal{P} = \mathcal{P}_0$;
for *b=1:Q* **do**
  1. PE Stage
  for *m=1:M* **do**
    $\hat{\boldsymbol{f}}_{\text{rff}} \leftarrow$ sample a function from $\hat{\boldsymbol{f}}$ from via random Fourier features;
    $\boldsymbol{\theta}_{1,m}, \boldsymbol{\theta}_{2,m} = \underset{\boldsymbol{\psi}_1, \boldsymbol{\psi}_2}{\text{argmax}}\ \text{EUBO}(\hat{u}(\hat{\boldsymbol{f}}_{\text{rff}}(\boldsymbol{\theta}_1)), \hat{u}(\hat{\boldsymbol{f}}_{\text{rff}}(\boldsymbol{\theta}_1)))$ optimize outcomes for the user;
    $\boldsymbol{\psi}_{1,m}, \boldsymbol{\psi}_{2,m} = \hat{\boldsymbol{f}}_{\text{rff}}(\boldsymbol{\theta}_{1,m}), \hat{\boldsymbol{f}}_{\text{rff}}(\boldsymbol{\theta}_{2,m})$;
    $r(\boldsymbol{\psi}_{1,m}, \boldsymbol{\psi}_{2,m}) \leftarrow$ user provides a comparison;
    $\boldsymbol{\psi}_{3,m} \leftarrow$ user provides a desired outcome based on $\boldsymbol{\psi}_{1,m}, \boldsymbol{\psi}_{2,m}$;
    Update $\mathcal{P} = \mathcal{P} \cup \{(\boldsymbol{\psi}_{1,m}, \boldsymbol{\psi}_{2,m}, r(\boldsymbol{\psi}_{1,m}, \boldsymbol{\psi}_{2,m})), (\boldsymbol{\psi}_{3,m}, \boldsymbol{\psi}_{1,m}, r = 1), (\boldsymbol{\psi}_{3,m}, \boldsymbol{\psi}_{2,m}, r = 1)\}$;
    Train $\hat{u}(\cdot|\mathcal{P})$ the utility model;
  end
  2. EXP Stage
  $\boldsymbol{\theta}_{1:N,b} = \text{argmax}\, q\text{EIUU}(\boldsymbol{\theta}_{1:N,b})$ optimize N set of policy parameters;
  $\boldsymbol{\psi}_{1:N,b} = \boldsymbol{f}(\boldsymbol{\theta}_{1:N,b})$ experiment the parameters in the real multi-outcome function;
  Update $\mathcal{D} = \mathcal{D} \cup \{(\theta_{i,b}, \boldsymbol{\psi}_{i,b})\}_{i=1}^{N}$;
  Train $\hat{\boldsymbol{f}}(\cdot|\mathcal{D})$ the multi-outcome model;
  $\theta^* = \text{argmax}\{u(\boldsymbol{\psi}_{1:N,b}), u(\boldsymbol{f}(\theta^*))\}$ optimal parameters;
end

Appendix D: Performance on Benchmarks

In this section we report the effect of considering the co-active feedback in three benchmark problems taken from (Lin et al., 2022). Here the Co-Active Preference learning BO method is called EUBO+Y3 and the Preference learning BO method is called EUBO.

Vehicle safety (5 inputs 3 outcomes) with Product of Kumaraswamy CDFs utility

DTLZ2 (8 inputs 5 outcomes) with negative ℓ_1 distance utility

OSY (6 inputs 8 outcomes) with piece-wise linear utility

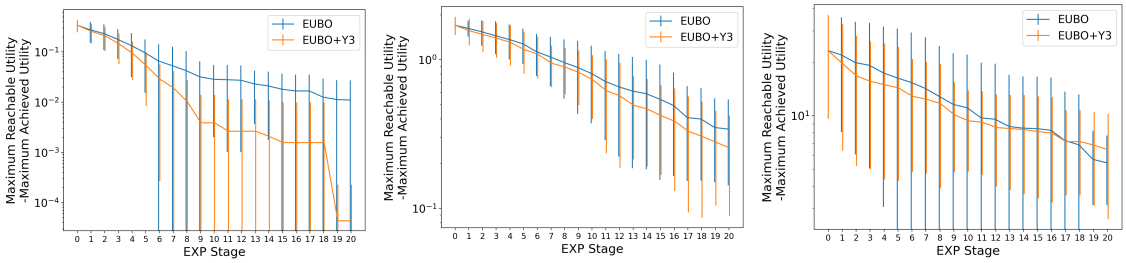

Figure 3: The performance of co-active feedback on benchmark problems. The existence of co-active feedback makes the algorithm outperform preference-learning BO without co-active feedback.

The Co-Active Preference learning BO method significantly outperform or perform as well as the Preference learning BO.