Bayesian Optimization over Discrete and Mixed Spaces via Probabilistic Reparameterization

Samuel Daulton[*†] SAMUEL.DAULTON@ENG.OX.AC.UK
Xingchen Wan[*†] XWAN@ROBOTS.OX.AC.UK
David Eriksson[†] DERIKSSON@FB.COM
Maximilian Balandat[†] BALANDAT@FB.COM
Michael A. Osborne[*] MOSB@ROBOTS.OX.AC.UK
Eytan Bakshy[†] EBAKSHY@FB.COM

Abstract

Optimizing expensive-to-evaluate black-box functions of discrete (and potentially continuous) design parameters is a ubiquitous problem in scientific and engineering applications. Bayesian optimization (BO) is a popular sample-efficient method that selects promising designs to evaluate by optimizing an acquisition function (AF) over some domain based on a probabilistic surrogate model. However, maximizing the AF over mixed or high-cardinality discrete search spaces is challenging as we cannot use standard gradient-based methods or evaluate the AF at every point in the search space. To address this issue, we propose using probabilistic reparameterization (PR). Instead of directly optimizing the AF over the search space containing discrete parameters, we instead maximize the expectation of the AF over a probability distribution defined by continuous parameters. We prove that under suitable reparameterizations, the BO policy that maximizes the probabilistic objective is the same as that which maximizes the AF, and therefore, PR enjoys the same regret bounds as the underlying AF. Moreover, our approach admits provably converges to a stationary point of the probabilistic objective under gradient ascent using scalable, unbiased estimators of both the probabilistic objective and its gradient, and therefore, as the numbers of starting points and gradient steps increase our approach will recover of a maximizer of the AF (an often neglected requisite for commonly-used BO regret bounds). We validate our approach empirically and demonstrate state-of-the-art optimization performance on many real-world applications. PR is complementary to (and benefits) recent work and naturally generalizes to settings with multiple objectives and black-box constraints.

Keywords: Bayesian Optimization

1. Introduction

Many scientific and engineering problems involve tuning discrete and/or continuous parameters to optimize an objective function. Often, the objective function is "black-box", meaning it has no known closed-form expression. For example, optimizing the design of an electrospun oil sorbent—a material that can be used to absorb oil in the case of a marine oil spill to mitigate ecological harm—can involve tuning both discrete ordinal experimental conditions

[*]University of Oxford
[†]Meta

and continuous parameters controlling the composition of the material to maximize the oil absorption capacity, mechanical strength, and water contact angle (Wang et al., 2020a). We consider the scenario where querying the objective function is expensive, in which case sample-efficiency is crucial. In the case of designing the oil sorbent, evaluating the objective function requires manufacturing the material and measuring its properties in a laboratory, requiring significant time and resources.

Bayesian optimization (BO) is a popular technique for sample-efficient black-box optimization, due to its proven performance guarantees in many settings (Srinivas et al., 2010; Berkenkamp et al., 2019) and its strong empirical performance (Frazier, 2018). BO leverages a probabilistic surrogate model of the unknown objective(s) and an acquisition function (AF) that provides utility values for evaluating a new design to balance exploration and exploitation. Typically, the maximizer of the AF is selected as the next design to evaluate. However, maximizing the AF over mixed search spaces (i.e., those consisting of discrete and continuous parameters) or large discrete search spaces is challenging[1] because continuous (or gradient-based) optimization routines cannot be directly applied. Theoretical performance guarantees of BO policies require that the maximizer of the acquisition function is found and selected as the next design to evaluate on the black-box objective function (Srinivas et al., 2010). When the maximizer is not found, regret properties are not guaranteed, and the performance of the BO policy may degrade.

To tackle these challenges, we propose to use probabilistic reparameterization (PR) to improve AF optimization. Our main contributions are:

1. We use PR to optimize a probabilistic objective (PO): the expectation of the AF over an induced distribution over the discrete parameters.
2. We prove that there is an equivalence between the maximizers of the acquisition function and the the maximizers of the PO and hence, the policy that chooses designs that are best with respect to the PO enjoys the same performance guarantees as the standard BO policy.
3. We derive scalable, unbiased Monte Carlo estimators of the PO and its gradient with respect to the parameters of the induced distribution. We show that stochastic gradient ascent using our gradient estimator is guaranteed to converge to a stationary point on the PO surface and will recover a global maximum of the underlying AF as the number of starting points and gradient steps increase. This is important because many BO regret bounds require maximizing the AF (Srinivas et al., 2010). Although the AF is often non-convex and maximization is hard, empirically, with a modest number of starting points, PR leads to better AF optimization than alternative methods.
4. We show that PR yields state-of-the-art optimization performance on a wide variety of real-world design problems with discrete and mixed search spaces. Importantly, PR is *complementary* to many existing approaches such as popular multi-objective, constrained, and trust region-based approaches; in particular, PR is agnostic to the underlying probabilistic model over discrete parameters—which is not the case for many alternative methods.

[1] If the discrete search space has low enough cardinality that the AF can be evaluated at every discrete element, then acquisition optimization can be solved trivially.

2. Preliminaries

Bayesian Optimization We consider the problem of optimizing a black-box function $f : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ over a compact search space $\mathcal{X} \times \mathcal{Z}$, where $\mathcal{X} = \mathcal{X}^{(1)} \times \cdots \times \mathcal{X}^{(d)}$ is the domain of the $d \geq 0$ continuous parameters ($x^{(i)} \in \mathcal{X}^{(i)}$ for $i = 1, ..., d$) and $\mathcal{Z} = \mathcal{Z}^{(1)} \times \cdots \times \mathcal{Z}^{(d_z)}$ is the domain of the $d_z \geq 1$ discrete parameters ($z^{(i)} \in \mathcal{Z}^{(i)}$ for $i = 1, ..., d_z$).[2]

When the black-box function is expensive-to-evaluate, Bayesian optimization (BO) is popular owing to its sample efficiency. BO leverages (i) a probabilistic surrogate model—typically a Gaussian process (GP) (Rasmussen, 2004)—fit to a data set $\mathcal{D}_n = \{\boldsymbol{x}_i, \boldsymbol{z}_i, y_i\}_{i=1}^n$ of designs and corresponding (potentially noisy) observations $y_i = f(\boldsymbol{x}_i, \boldsymbol{z}_i) + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, \sigma^2)$, and (ii) an acquisition function $\alpha(\boldsymbol{x}, \boldsymbol{z})$ that uses the surrogate model's posterior distribution to quantify the value of evaluating a new design. Common AFs include expected improvement (EI) (Jones et al., 1998) and upper confidence bound (UCB) (Srinivas et al., 2010)—the latter of which enjoys no-regret guarantees in certain settings. The next design to evaluate is chosen by maximizing the AF $\alpha(\boldsymbol{x}, \boldsymbol{z})$ over $\mathcal{X} \times \mathcal{Z}$. Although the black-box objective f is expensive-to-evaluate, the AF is relatively cheap-to-query, and therefore, can be optimized numerically (gradient-based optimization routines are often used to maximize the AF over continuous domains).

Discrete Parameters Typically, BO assumes that the inputs are continuous. However, discrete parameters such as binary parameters, discrete ordinal parameters, and non-ordered categorical parameters, which are summarized in Table 1, are ubiquitous in many applications. In the presence of such parameters, optimizing the AF is more difficult, as standard gradient-based approaches cannot be directly applied. Recent works have proposed various approaches including multi-armed bandits (Nguyen et al., 2020; Ru et al., 2020) and local search (Oh et al., 2019) for discrete domains and interleaved discrete/continuous optimization procedures for mixed domains (Deshwal et al., 2021a; Wan et al., 2021). A simple and widely-used approach across many popular BO packages (Balandat et al., 2020; The GPyOpt authors, 2016) is to one-hot encode the categorical parameters, apply a continuous relaxation when solving the optimization, and discretize (round) the resulting continuous candidates. Examples of continuous relaxations and discretization functions are listed in Table 1.

Table 1: Different parameter types, their continuous relaxations, and discretization functions.

Type	Domain	Cont. Relaxation	discretize(·) Function
Binary	$z \in \{0, 1\}$	$z' \in [0, 1]$	$\text{round}(z')$
Ordinal	$z \in \{0, ..., C-1\}$	$z' \in [0, C-1]$	$\text{round}(z')$
Categorical	$z \in \{0, ..., C-1\}$	$z' \in [0, 1]^C$	$\arg\max_c z'^{(c)}$

Although using a continuous relaxation allows for efficient optimization using standard optimization routines in an alternate continuous domain $\mathcal{Z}' \subset \mathbb{R}^m$, the acquisition value for an infeasible continuous value (i.e., $z' \notin \mathcal{Z}$) does not account for the discretization that must occur before the black-box function is evaluated. Moreover, the acquisition value for an infeasible continuous value can be larger than the acquisition value after discretization. For

[2] Throughout this paper, we use a mixed search space $\mathcal{X} \times \mathcal{Z}$ in our derivations, theorems, and proofs, without loss of generality with respect to the case of a purely discrete search space. If $d = 0$, then the objective function $f : \mathcal{Z} \to \mathbb{R}$ is defined over the discrete space $\mathcal{Z}$ and the continuous parameters in this exposition can simply be ignored.

an illustration of this, see Fig. 1 (middle/right). In the worst case, BO will repeatedly select the same infeasible continuous design due to its high acquisition value, but discretization will result in a design that has already been evaluated and has zero acquisition value. To mitigate this degenerate behavior, Garrido-Merchán and Hernández-Lobato (2020) propose to discretize the candidate before evaluating the AF. While this improves performance on small search spaces, the resulting AF and surrogate have large flat regions, which makes it difficult to optimize the AF. The authors of (Garrido-Merchán and Hernández-Lobato, 2020) propose to approximate the gradients using finite differences, but, empirically, we find that this approach to be suboptimal for optimizing the AF.

3. Probabilistic Reparameterization

We propose an alternative approach based on probabilistic reparameterization, a relaxation of the original optimization problem involving discrete parameters. Rather than directly optimizing the AF via a continuous relaxation $\boldsymbol{z}'$ of the design $\boldsymbol{z}$, we instead reparameterize the optimization problem by introducing a discrete probability distribution $p(\boldsymbol{Z}|\boldsymbol{\theta})$ over a random variable $\boldsymbol{Z}$ with support exclusively over $\mathcal{Z}$. This distribution is parameterized by a vector of continuous parameters $\boldsymbol{\theta}$.

Table 2: Examples of probabilistic reparameterizations for different parameter types. Although the distribution $p(Z|\theta)$ for an ordinal random variable Z does not come from a standard probability distribution, the random variable B is Bernoulli and for any θ, the distribution $p(Z|\theta)$ is simply a Bernoulli random variable with an additive offset.

Parameter Type	Random Variable	Continuous Parameter
Binary	$Z \sim \text{Bernoulli}(\theta)$	$\theta \in [0, 1]$
Ordinal	$Z = \lfloor\theta\rfloor + B, B \sim \text{Bernoulli}(\theta - \lfloor\theta\rfloor)$	$\theta \in [0, C-1]$
Categorical	$Z \sim \text{Categorical}(\theta), \theta = (\theta^{(1)}, ..., \theta^{(C)})$	$\theta \in [0, 1]^C$

We use $\boldsymbol{z}$ to denote the vector $(z^{(1)}, ..., z^{(d_z)})$, where each element is a different (possibly vector-valued) discrete parameter). Given this reparameterization, the probabilistic objective (PO) is defined as $\mathbb{E}_{\boldsymbol{Z}\sim p(\boldsymbol{Z}|\boldsymbol{\theta})}[\alpha(\boldsymbol{x}, \boldsymbol{Z})]$. PR enables optimizing $\boldsymbol{\theta}$ and $\boldsymbol{x}$ over a continuous space to maximize the PO instead of optimizing $\boldsymbol{x}, \boldsymbol{z}$ to maximize α directly over the mixed search space $\mathcal{X} \times \mathcal{Z}$. As we will show later, maximizing the PO allows us to recover a maximizer of α over the space $\mathcal{X} \times \mathcal{Z}$. As we will show later, the maxima of the PO (when discretized via Table 1) recovers the true optimizers of the AF that would be obtained if the search over $\mathcal{X} \times \mathcal{Z}$ were to converge globally. Choosing $p(\boldsymbol{Z}|\theta)$ to be a discrete distribution over $\mathcal{Z}$ means the realizations of $\boldsymbol{Z}$ are feasible values in $\mathcal{Z}$. Hence, the AF is only evaluated for feasible discrete designs. Since $p(\boldsymbol{Z}|\boldsymbol{\theta})$ is a discrete probability distribution, we can express $\mathbb{E}_{\boldsymbol{Z}\sim p(\boldsymbol{Z}|\boldsymbol{\theta})}[\alpha(\boldsymbol{x}, \boldsymbol{Z})]$ as a linear combination where each discrete design is weighted by its probability mass:$\mathbb{E}_{\boldsymbol{Z}\sim p(\boldsymbol{Z}|\boldsymbol{\theta})}[\alpha(\boldsymbol{x}, \boldsymbol{Z})] = \sum_{\boldsymbol{z}\in\mathcal{Z}} p(\boldsymbol{z}|\boldsymbol{\theta})\alpha(\boldsymbol{x}, \boldsymbol{z})$. Example distributions for binary, ordinal, and categorical parameters are provided in Table 2. Multiple discrete parameters can be handled using an independent random variable $Z^{(i)} \sim p(Z^{(i)}|\theta^{(i)})$ for each parameter $z^{(i)}$ for $i = 1, ..., d_z$.

One important benefit of PR is that the PO is differentiable with respect to $\boldsymbol{\theta}$ (and $\boldsymbol{x}$, if the gradient of α with respect to $\boldsymbol{x}$ exists), whereas $\alpha(\boldsymbol{x}, \boldsymbol{z})$ is not differentiable with respect

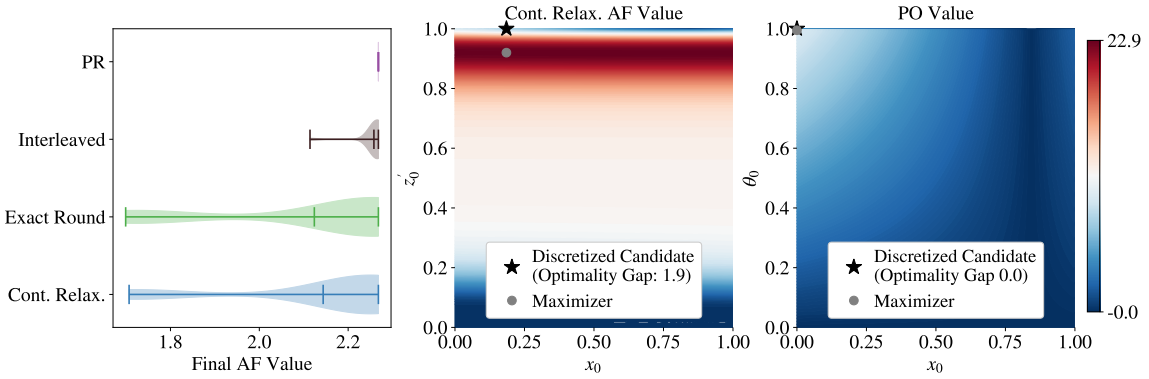

Figure 1: (**Left**) A comparison of AF optimization using different methods over a mixed search space shows that PR *outperforms alternative methods for AF optimization and has much lower variance across replications*. For each method, the best candidate across 20 restarts is selected (after discretization) and the acquisition value of the resulting feasible candidate is recorded. (**Middle/Right**) AF values with a continuous relaxation (middle) and the PO (right) for the Branin function over a mixed domain with one continuous parameter (x_0) and one binary parameter (z_0) (see Appendix E for details on Branin). (**Middle**) Under a continuous relaxation, the maximizer of the AF is an infeasible point in the domain (grey circle), which results in suboptimal AF value when rounded (black star); the optimality gap with respect to the true AF maximizer is 1.9. The maximal AF across the feasible search space is shown in white, so red regions indicated that the continuous relaxation overestimates the AF value. (**Right**) The PO is maximized at the AF unique maximizer within the valid search domain. These contours show that PR avoids the overestimation issue with a continuous relaxation.

to $\boldsymbol{z}$. See Appendix A for details. This enables optimizing the PO (line 4 of Algorithm 1) efficiently and effectively using gradient-based methods over discrete or mixed search spaces.

Theoretical Properties Here, we highlight important theoretical properties of PR. Formal statements and proofs are provided in Appendix D. Algorithm 1 outlines BO with probabilistic reparameterization. Importantly, Theorem 1 states that sampling from the distribution parameterized by a maximizer of the PO yields a maximizer of α, and therefore, Algorithm 1 enjoys the regret bounds as the BO policy that optimizes $\alpha(\cdot)$ (Corollary 3).Although the BO policy selects a maximizer of α is equivalent to the BO policy in Algorithm 1, maximizing the AF over mixed or high-dimensional discrete search spaces is challenging because commonly used gradient-based methods cannot directly be applied. The key advantage of our approach is that maximizers of the AF can be identified efficiently and effectively by optimizing the PO using gradient information instead of directly optimizing the AF. We find that optimizing PR yields significantly better results than directly optimizing α or other common relaxations as shown in Figure 1(Left), where we compare AF optimization methods on the chemical synthesis problem (Shields et al., 2021) (see Appendix E for details). The violin plots show the distribution of final AF values and the mean. "Cont. Relax." denotes optimizing a continuous relaxation of the categoricals with exact gradients. "Exact Round" refers to optimizing a continuous relaxation with

approximate gradients (via finite difference), but discretizes the relaxation before evaluating the surrogate (Garrido-Merchán and Hernández-Lobato, 2020). "Interleaved" alternates between one step of local search on the discrete parameters and one step of gradient ascent on the continuous parameters (used in CASMOPOLITAN (Wan et al., 2021)). All methods except interleaved use L-BFGS-B and the expected improvement AF (Jones et al., 1998).

Algorithm 1 BO with PR

Input: black-box objective $f : \mathcal{X} \times \boldsymbol{Z} \to \mathbb{R}$
Initialize $\mathcal{D}_0 \leftarrow \emptyset$, $\text{GP}_0 \leftarrow \text{GP}(\mathbf{0}, k)$
for $n = 1$ **to** N **do**
  $(\boldsymbol{x}_n, \boldsymbol{\theta}_n) \leftarrow \arg\max_{(\boldsymbol{x},\boldsymbol{\theta}) \in \mathcal{X} \times \Theta} \mathbb{E}_{\boldsymbol{Z} \sim p(\boldsymbol{Z}|\boldsymbol{\theta})}[\alpha(\boldsymbol{x}, \boldsymbol{Z})]$
  Sample $\boldsymbol{z}_n \sim p(\boldsymbol{Z}|\boldsymbol{\theta}_n)$
  Evaluate $f(\boldsymbol{x}_n, \boldsymbol{z}_n)$
  $\mathcal{D}_n \leftarrow \mathcal{D}_{n-1} \cup \{(\boldsymbol{x}_n, \boldsymbol{z}_n, \boldsymbol{f}(\boldsymbol{x}_n, \boldsymbol{z}_n))\}$
  Update posterior GP_n given $\mathcal{D}_n$
end for

4. Practical Computation and Optimization

Unbiased Monte Carlo Estimators As the number of discrete configurations ($|\mathcal{Z}|$) increases, the PO and its gradient may become computationally expensive to evaluate analytically because both require a summation of $|\mathcal{Z}|$ terms. Therefore, we propose to estimate the PO and its gradient using Monte Carlo (MC) sampling. We opt for using a score function gradient estimator (Kleijnen and Rubinstein, 1996) (also known as REINFORCE (Williams, 1992)) and the likelihood ratio estimator (Glynn, 1990)) because it is simple, scalable, and can be computed simply using the acquisition values $\{\alpha(\boldsymbol{x}, \tilde{\boldsymbol{z}}_i)\}_{i=1}^N$ that are already required for the MC estimator of the PO. See Appendix B for details.

Convergence Guarantee using Stochastic Gradient Ascent Since the score function gradient estimator is unbiased, we can leverage previous work on convergence in probability under stochastic gradient ascent (Robbins and Monro, 1951) to arrive at our main convergence result for acquisition optimization: namely, that optimizing the PO is guaranteed to converge in probability to a stationary point, and as the number of starting points and gradient steps increase, our approach will recover a maximizer of the (Theorem 2).

The implication is that the intended BO policy is followed and the underlying regret bounds of the AF are recovered (provided that the other conditions of the regret bound are met). Although global convergence is only guaranteed as $m \to \infty$, we observe in Figure 1(left) that PR yields stable, convergent acquisition optimization with only $m = 20$ starting points and outperforms alternative optimization approaches.

5. Summary of Empirical Results

In Appendix C, we perform an empirical analysis of PR against alternative state-of-the-art methods (see Appendix N for review of related work) on 3 synthetic and 5 real world problems. We find PR consistently delivers strong empirical performance and we demonstrate that PR is complementary to existing trust region-based approaches (Eriksson et al., 2019). Although PR is computationally intensive (due to MC integration), the computation is embarrassingly parallel and therefore exploiting GPU acceleration yields competitive wall times. See Appendix O for further discussion.

References

Maximilian Balandat, Brian Karrer, Daniel R. Jiang, Samuel Daulton, Benjamin Letham, Andrew Gordon Wilson, and Eytan Bakshy. Botorch: A framework for efficient monte-carlo bayesian optimization. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Ricardo Baptista and Matthias Poloczek. Bayesian optimization of combinatorial structures. In *Proc. of ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 471–480. PMLR, 2018.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *ArXiv preprint*, abs/1308.3432, 2013.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 2546–2554, 2011.

Felix Berkenkamp, Angela P. Schoellig, and Andreas Krause. No-regret bayesian optimization with unknown hyperparameters. *Journal of Machine Learning Research*, 20(50):1–24, 2019. URL `http://jmlr.org/papers/v20/18-213.html`.

Laurens Bliek, Arthur Guijt, Sicco Verwer, and Mathijs de Weerdt. Black-box mixed-variable optimisation using a surrogate model that satisfies integer constraints. GECCO '21, page 1851–1859. Association for Computing Machinery, 2021.

Endre Boros and Peter L. Hammer. Pseudo-boolean optimization. *Discrete Applied Mathematics*, 123(1):155–225, 2002. ISSN 0166-218X. doi: https://doi.org/10.1016/S0166-218X(01)00341-9.

Samuel Daulton, Maximilian Balandat, and Eytan Bakshy. Differentiable expected hypervolume improvement for parallel multi-objective bayesian optimization. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Samuel Daulton, David Eriksson, Maximilian Balandat, and Eytan Bakshy. Multi-objective bayesian optimization over high-dimensional search spaces. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, UAI 2022, Eindhoven, Netherlands, August 1-5, 2022*, Proceedings of Machine Learning Research. AUAI Press, 2022.

Huw ML Davies and Daniel Morton. Recent advances in c–h functionalization. *The Journal of Organic Chemistry*, 81(2):343–350, 2016.

Erik A. Daxberger, Anastasia Makarova, Matteo Turchetta, and Andreas Krause. Mixed-variable bayesian optimization. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 2633–2639. ijcai.org, 2020. doi: 10.24963/ijcai.2020/365.

Aryan Deshwal and Jana Doppa. Combining latent space and structured kernels for bayesian optimization over combinatorial spaces. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8185–8200. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper/2021/file/44e76e99b5e194377e955b13fb12f630-Paper.pdf`.

Aryan Deshwal, Syrine Belakaria, and Janardhan Rao Doppa. Bayesian optimization over hybrid spaces. In *Proc. of ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 2632–2643. PMLR, 2021a.

Aryan Deshwal, Syrine Belakaria, and Janardhan Rao Doppa. Mercer features for efficient combinatorial bayesian optimization. In *Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*, pages 7210–7218, 2021b.

Ryan M Dreifuerst, Samuel Daulton, Yuchen Qian, Paul Varkey, Maximilian Balandat, Sanjay Kasturia, Anoop Tomar, Ali Yazdan, Vish Ponnampalam, and Robert W Heath. Optimizing coverage and capacity in cellular networks using machine learning. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8138–8142. IEEE, 2021.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

Philippe Duchon, Philippe Flajolet, Guy Louchard, and Gilles Schaeffer. Boltzmann samplers for the random generation of combinatorial structures. *Comb. Probab. Comput.*, 13(4–5): 577–625, jul 2004. ISSN 0963-5483. doi: 10.1017/S0963548304006315.

M. T. M. Emmerich, K. C. Giannakoglou, and B. Naujoks. Single- and multiobjective evolutionary optimization assisted by gaussian random field metamodels. *IEEE Transactions on Evolutionary Computation*, 10(4):421–439, 2006.

David Eriksson and Matthias Poloczek. Scalable constrained Bayesian optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 730–738. PMLR, 2021.

David Eriksson, Michael Pearce, Jacob R. Gardner, Ryan Turner, and Matthias Poloczek. Scalable global optimization via local bayesian optimization. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5497–5508, 2019.

Peter I Frazier. A tutorial on bayesian optimization. *ArXiv preprint*, abs/1807.02811, 2018.

Jacob R. Gardner, Matt J. Kusner, Zhixiang Eddie Xu, Kilian Q. Weinberger, and John P. Cunningham. Bayesian optimization with inequality constraints. In *Proc. of ICML*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 937–945. JMLR.org, 2014.

Eduardo C. Garrido-Merchán and D. Hernández-Lobato. Dealing with categorical and integer-valued variables in bayesian optimization with gaussian processes. *Neurocomputing*, 380:20–35, 2020.

Peter W Glynn. Likelihood ratio gradient estimation for stochastic systems. *Communications of the ACM*, 33(10):75–84, 1990.

Nikolaus Hansen, Dimo Brockhoff, Olaf Mersmann, Tea Tusar, Dejan Tusar, Ouassim Ait ElHara, Phillipe R Sampaio, Asma Atamna, Konstantinos Varelas, Umut Batu, et al. Comparing continuous optimizers: numbbo/coco on github. 2019.

Florian Häse, Matteo Aldeghi, Riley J Hickman, Loïc M Roch, and Alán Aspuru-Guzik. Gryffin: An algorithm for bayesian optimization of categorical variables informed by expert knowledge. *Applied Physics Reviews*, 8(3):031406, 2021.

Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.

Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *Proceedings of the 5th International Conference on Learning and Intelligent Optimization*, page 507–523. Springer-Verlag, 2011. ISBN 9783642255656.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *Proc. of ICLR*. OpenReview.net, 2017.

Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13:455–492, 1998.

Kirthevasan Kandasamy, Jeff Schneider, and Barnabas Poczos. High dimensional bayesian optimisation and bandits via additive models. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 295–304, Lille, France, 07–09 Jul 2015. PMLR. URL `https://proceedings.mlr.press/v37/kandasamy15.html`.

Jack PC Kleijnen and Reuven Y Rubinstein. Optimization and sensitivity analysis of computer simulation models by the score function method. *European Journal of Operational Research*, 88(3):413–427, 1996.

Sulin Liu, Qing Feng, David Eriksson, Benjamin Letham, and Eytan Bakshy. Sparse bayesian optimization. *arXiv preprint arXiv:2203.01900*, 2022.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *Proc. of ICLR*. OpenReview.net, 2017.

Wesley J Maddox, Maximilian Balandat, Andrew G Wilson, and Eytan Bakshy. Bayesian optimization with high-dimensional outputs. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Shakir Mohamed, Mihaela Rosca, Michael Figurnov, and Andriy Mnih. Monte carlo gradient estimation in machine learning. *J. Mach. Learn. Res.*, 21(132):1–62, 2020.

Dang Nguyen, Sunil Gupta, Santu Rana, Alistair Shilton, and Svetha Venkatesh. Bayesian optimization for categorical and category-specific continuous inputs. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 5256–5263. AAAI Press, 2020.

ChangYong Oh, Jakub M. Tomczak, Efstratios Gavves, and Max Welling. Combinatorial bayesian optimization using the graph cartesian product. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 2910–2920, 2019.

Art B Owen. Quasi-monte carlo sampling. *Monte Carlo Ray Tracing: Siggraph*, 1:69–88, 2003.

Julien Pelamatti, Loïc Brevault, Mathieu Balesdent, El-Ghazali Talbi, and Yannick Guerin. Bayesian optimization of variable-size design space problems. *Optimization and Engineering*, 22(1):387–447, 2021.

Carl Edward Rasmussen. Gaussian processes in machine learning. In *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures*, 2004.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. ISSN 00034851.

Bin Xin Ru, Ahsan S. Alvi, Vu Nguyen, Michael A. Osborne, and Stephen J. Roberts. Bayesian optimisation over multiple continuous and categorical inputs. In *Proc. of ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 8276–8285. PMLR, 2020.

Soumya Ranjan Samal, Kaliprasanna Swain, Shuvabrata Bandopadhaya, Nikolay Dandanov, Vladimir Poulkov, Sidheswar Routray, and Gopinath Palai. Dynamic coverage optimization for 5g ultra-dense cellular networks based on their user densities. 2021.

Benjamin J Shields, Jason Stevens, Jun Li, Marvin Parasram, Farhan Damani, Jesus I Martinez Alvarado, Jacob M Janey, Ryan P Adams, and Abigail G Doyle. Bayesian reaction optimization as a tool for chemical synthesis. *Nature*, 590(7844):89–96, 2021.

Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias W. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proc. of ICML*, pages 1015–1022. Omnipress, 2010.

The GPyOpt authors. GPyOpt: A bayesian optimization framework in python. `http://github.com/SheffieldML/GPyOpt`, 2016.

Anh Tran, Minh Tran, and Yan Wang. Constrained mixed-integer gaussian mixture bayesian optimization and its applications in designing fractal and auxetic metamaterials. *Structural and Multidisciplinary Optimization*, 59(6):2131–2154, 2019.

Tea Tušar, Dimo Brockhoff, and Nikolaus Hansen. Mixed-integer benchmark problems for single- and bi-objective optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO '19, page 718–726. Association for Computing Machinery, 2019.

Xingchen Wan, Vu Nguyen, Huong Ha, Bin Xin Ru, Cong Lu, and Michael A. Osborne. Think global and act local: Bayesian optimisation over high-dimensional categorical and mixed search spaces. In *Proc. of ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 10663–10674. PMLR, 2021.

Xingchen Wan, Cong Lu, Jack Parker-Holder, Philip J Ball, Vu Nguyen, Binxin Ru, and Michael Osborne. Bayesian generational population-based training. In *ICLR Workshop on Agent Learning in Open-Endedness*, 2022.

Boqian Wang, Jiacheng Cai, Chuangui Liu, Jian Yang, and Xianting Ding. Harnessing a novel machine-learning-assisted evolutionary algorithm to co-optimize three characteristics of an electrospun oil sorbent. *ACS Applied Materials & Interfaces*, 12(38):42842–42849, 2020a.

Jialei Wang, Scott C Clark, Eric Liu, and Peter I Frazier. Parallel bayesian global optimization of expensive functions. *Operations Research*, 68(6):1850–1865, 2020b.

Rui Wang, Jian Xiong, Hisao Ishibuchi, Guohua Wu, and Tao Zhang. On the effect of reference point in moea/d for multi-objective optimization. *Applied Soft Computing*, 58: 25–34, 2017. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2017.04.002.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

James T. Wilson, Frank Hutter, and Marc Peter Deisenroth. Maximizing acquisition functions for bayesian optimization. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9906–9917, 2018.

Mingzhang Yin, Yuguang Yue, and Mingyuan Zhou. ARSM: augment-reinforce-swap-merge estimator for gradient backpropagation through categorical variables. In *Proc. of ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 7095–7104. PMLR, 2019.

Mingzhang Yin, Nhat Ho, Bowei Yan, Xiaoning Qian, and Mingyuan Zhou. Probabilistic Best Subset Selection by Gradient-Based Optimization. *arXiv e-prints*, 2020.

Yichi Zhang, Siyu Tao, Wei Chen, and Daniel Apley. A latent variable approach to gaussian process modeling with qualitative and quantitative factors. *Technometrics*, 62:1–19, 07 2019. doi: 10.1080/00401706.2019.1638834.

Appendix A. Analytic Gradients

Recall that the PO is given by

$$\mathbb{E}_{\boldsymbol{Z}\sim p(\boldsymbol{Z}|\boldsymbol{\theta})}[\alpha(\boldsymbol{x},\boldsymbol{Z})] = \sum_{\boldsymbol{z}\in\mathcal{Z}} p(\boldsymbol{z}|\boldsymbol{\theta})\alpha(\boldsymbol{x},\boldsymbol{z}). \tag{1}$$

The gradients of the PO with respect to $\boldsymbol{\theta}$ and $\boldsymbol{x}$ can be obtained by differentiating Equation 1:

$$\nabla_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{Z}\sim p(\boldsymbol{Z}|\boldsymbol{\theta})}[\alpha(\boldsymbol{x},\boldsymbol{Z})] = \sum_{\boldsymbol{z}\in\mathcal{Z}} \alpha(\boldsymbol{x},\boldsymbol{z})\nabla_{\boldsymbol{\theta}} p(\boldsymbol{z}|\boldsymbol{\theta}) \tag{2}$$

$$\nabla_{\boldsymbol{x}}\mathbb{E}_{\boldsymbol{Z}\sim p(\boldsymbol{Z}|\boldsymbol{\theta})}[\alpha(\boldsymbol{x},\boldsymbol{Z})] = \sum_{\boldsymbol{z}\in\mathcal{Z}} p(\boldsymbol{z}|\boldsymbol{\theta})\nabla_{\boldsymbol{x}}\alpha(\boldsymbol{x},\boldsymbol{z}) \tag{3}$$

Appendix B. Practical Monte Carlo Estimators

B.1 Unbiased estimators of the Probabilistic Reparameterization and its Gradient

As the number of discrete configurations ($|\mathcal{Z}|$) increases, the PO and its gradient may become computationally expensive to evaluate analytically because both require a summation of $|\mathcal{Z}|$ terms. Therefore, we propose to estimate the PO and its gradient using Monte Carlo (MC) sampling. The MC estimator of the PO is given by

$$\mathbb{E}_{\boldsymbol{Z}\sim p(\boldsymbol{Z}|\boldsymbol{\theta})}[\alpha(\boldsymbol{x},\boldsymbol{Z})] \approx \frac{1}{N}\sum_{i=1}^{N}\alpha(\boldsymbol{x},\tilde{\boldsymbol{z}}_i), \tag{4}$$

where $\tilde{\boldsymbol{z}}_1, ..., \tilde{\boldsymbol{z}}_N$ are samples from $p(\boldsymbol{Z}|\boldsymbol{\theta})$. This estimator is unbiased and can be computed for a large number of samples by evaluating the AF independently (or in chunks) for each input $(\boldsymbol{x}, \tilde{\boldsymbol{z}}_n)$.

MC can also be used to estimate the gradient of the PO with respect to $\boldsymbol{\theta}$. We opt for using a score function gradient estimator (Kleijnen and Rubinstein, 1996) (also known as REINFORCE (Williams, 1992)) and the likelihood ratio estimator (Glynn, 1990)) because it is simple, scalable, and can be computed simply using the acquisition values $\{\alpha(\boldsymbol{x},\tilde{\boldsymbol{z}}_i)\}_{i=1}^{N}$ that are already required for the MC estimator of the PO. Many alternative lower variance estimators (e.g. Yin et al. (2020); Yin et al. (2019)) would require many additional AF evaluations (see Mohamed et al. (2020) for a review of MC gradient estimation). The score function is the gradient of the log probability with respect to the parameters of the distribution: $\nabla_{\boldsymbol{\theta}}\log p(\boldsymbol{Z}|\boldsymbol{\theta}) = \frac{\nabla_{\boldsymbol{\theta}} p(\boldsymbol{Z}|\boldsymbol{\theta})}{p(\boldsymbol{Z}|\boldsymbol{\theta})}$. Using this score function, we can express the analytic gradient as

$$\nabla_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{Z}\sim p(\boldsymbol{Z}|\boldsymbol{\theta})}[\alpha(\boldsymbol{x},\boldsymbol{Z})] = \sum_{\boldsymbol{z}\in\mathcal{Z}} \alpha(\boldsymbol{x},\boldsymbol{z})p(\boldsymbol{z}|\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}\log p(\boldsymbol{z}|\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{Z}\sim p(\boldsymbol{Z}|\boldsymbol{\theta})}[\alpha(\boldsymbol{x},\boldsymbol{Z})\nabla_{\boldsymbol{\theta}}\log p(\boldsymbol{Z}|\boldsymbol{\theta})].$$

The unbiased MC estimator of the gradient of the PO with respect to $\boldsymbol{\theta}$ is given by

$$\nabla_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{Z}\sim p(\boldsymbol{Z}|\boldsymbol{\theta})}[\alpha(\boldsymbol{x},\boldsymbol{Z})] \approx \frac{1}{N}\sum_{i=1}^{N}\alpha(\boldsymbol{x},\tilde{\boldsymbol{z}}_i)\nabla_{\boldsymbol{\theta}}\log p(\tilde{\boldsymbol{z}}_i|\boldsymbol{\theta}). \tag{5}$$

Since the score function gradient is only defined when $p(\boldsymbol{z}|\boldsymbol{\theta}) > 0$, we reparameterize $\boldsymbol{\theta}$ to ensure $p(\boldsymbol{z}|\boldsymbol{\theta}) > 0$ for all $\boldsymbol{z}$ and $\boldsymbol{\theta}$ by using the softmax transformations provided in Table 3, which are commonly used for computational convenience and stability in probablistic reparameterization (Yin et al., 2020; Yin et al., 2019). Although θ can only be 0 or 1 when $\tau \to 0$, for any $\epsilon > 0$, there exists some τ such that the approximation error is less than ϵ. Moreover, even though $p(\boldsymbol{z}|\boldsymbol{\theta}) > 0$, when $p(\boldsymbol{z}|\boldsymbol{\theta})$ is small, a small number N of MC samples are unlikely to produce any samples where $\tilde{\boldsymbol{z}} = \boldsymbol{z}$. Instead of optimizing $\boldsymbol{\theta}$ directly, we instead optimize $\boldsymbol{\phi}$. Since the transformations $g(\cdot)$ are differentiable with respect to $\boldsymbol{\phi}$, the gradient (and MC gradient estimator) of the PO with respect to $\boldsymbol{\phi}$ are easily obtained using the gradient of the PO with respect to $\boldsymbol{\theta}$ and a simple application the chain rule (multiplying by $\nabla_{\boldsymbol{\phi}}\boldsymbol{\theta}$).

Parameter Type	Transformation ($\theta = g(\phi)$)
Binary	$\theta = \sigma((\phi - \frac{1}{2})/\tau)$
Ordinal	$\theta = \lfloor\phi\rfloor + \sigma((\phi - \lfloor\phi\rfloor - \frac{1}{2})/\tau)$
Categorical	$\theta^{(c)} = \text{softmax}((\boldsymbol{\phi} - 0.5)/\tau)^{(c)}$

Table 3: Transformations where $\tau \in \mathbb{R}_+$ and $\phi, \theta \in \Theta$.

B.2 Deterministic Optimization via Sample Average Approximation

Although multi-start stochastic ascent is provably convergent (Theorem 2), we opt instead to use lower variance deterministic MC estimators using common random numbers. In order maintain valid samples $\tilde{\boldsymbol{z}}_i \sim p(\boldsymbol{Z}|\boldsymbol{\theta})$ for $i = 1, ..., N$ as $\boldsymbol{\theta}$ changes over the course of the acquisition optimization, we reparameterize $\boldsymbol{Z}$ as a deterministic function $h(\cdot, \cdot)$ that operates component-wise on $\boldsymbol{\theta}$ and the random variable $\boldsymbol{U} = (u^{(1)}, ..., u^{(d_z)}), u^{(i)} \sim \text{Uniform}(0, 1)$: $\boldsymbol{Z} = h(\boldsymbol{\theta}, \boldsymbol{U})$. Descriptions of $h(\cdot, \cdot)$ for provided for each parameter type in Appendix E.3. Using a fixed a set of base samples $\{\tilde{\boldsymbol{u}}_i\}_{i=1}^N$, the sample average approximation estimators are obtained by substituting $\boldsymbol{z}_i = h(\boldsymbol{\theta}, \tilde{\boldsymbol{u}}_i)$ into (4) and (5). Although biased, sample average approximation estimators are deterministic, enabling the use of second-order numerical methods for efficient acquisition optimization (Balandat et al., 2020). We reduce the variance further by leveraging quasi-MC sampling (Owen, 2003) instead of i.i.d. sampling.

Appendix C. Experiments

In this section, we provide an empirical evaluation of PR using SAA estimators on a suite of synthetic problems and real world applications. We use $N = 1,024$ MC samples in our experiments and demonstrate that PR is robust with respect to the number of MC samples (and compare against analytic PR, where computationally feasible) in Appendix H. We compare PR against two alternative acquisition optimization strategies: using a continuous relaxation (Cont. Relax.) and using exact discretization and approximate gradients (Exact Round) (Garrido-Merchán and Hernández-Lobato, 2020). In addition, we compare against two state-of-the-art methods for discrete/mixed BO: a modified version of Casmopolitan (Wan et al., 2021) that additionally supports ordinal variables introduced in Wan et al. (2022) and HyBO (Deshwal et al., 2021a), both of which are shown to outperform the other related works discussed in Section N. In addition, we showcase how PR is complementary to existing methods such as trust region methods (Eriksson et al., 2019). We demonstrate this by using PR with a trust region for the continuous and discrete ordinal parameters and optimize PR within this trust region. In Appendix I, we provide comparison of TR

methods with alterative optimizers and find that PR is the best optimizer for TR on 6 of the 7 benchmark problems. See Appendix E for additional discussion of PR + TR. For PR, Exact Round, and PR + TR we use the sum of a product kernel and a sum kernel of a categorical kernel (Ru et al., 2020) for the categorical parameters and Matérn-5/2 kernel for all other parameters.[3]

Cont. Relax., Exact Round, PR, and PR + TR use expected improvement (Jones et al., 1998; Gardner et al., 2014) for single objective (constrained problems) and expected hypervolume improvement (Emmerich et al., 2006) for the multi-objective oil sorbent problem (where exact gradients with respect to continuous parameters are computed using auto-differentiation (Daulton et al., 2020)). We report the mean for each method ± 2 standard errors across 20 replications. Performance is evaluated in terms of regret (feasible regret for constrained problems and hypervolume regret for multi-objective problems). Casmopolitan and HyBO are not run on Welded Beam and Oil Sorbent as they do not support constrained and multi-objective optimization. We also leave the multi-objective extension of PR+TR to future work because it would add additional complexity (Daulton et al., 2022). For HyBO, we only run 60 BO iterations on SVM due to the large wall time (see Figure 4) and only report partial results on Cellular Network due to a singular covariance matrix error. See Appendix E for details on the experiment setup, regret metrics, benchmark problems, and methodological details. We leverage existing open source implementations of Casmopolitan and HyBO (see Appendix E for links). Our implementations of all of other methods are available at `github.com/facebookresearch/bo_pr`.

C.1 Synthetic Problems

We evaluate all methods on 3 synthetic problems: **Ackley** is a 13-dimensional function with 10 binary and 3 continuous parameters (a modified version of the problem in Bliek et al. (2021)). **Mixed Int F1** is a 16-dimensional variant of the F1 function from Tušar et al. (2019) with 2 binary, 6 discrete ordinal parameters, and 8 continuous parameters. The discrete ordinal parameters have following cardinalities: 2 parameters with 3 values, 2 with 5 values, and 2 with 7 values. **Rosenbrock** is a 10-dimensional Rosenbrock function with 6 discrete ordinal parameters with 4 values each and 4 continuous parameters.

C.2 Real World Problems

We consider 5 real world applications including a problem with 5 black-box outcome constraints and a 3-objective problem (see Appendix F for details on constrained and multi-objective BO).

Welded Beam Optimizing the design of a welded steel beam is a classical engineering optimization. In this problem, the goal is to minimize manufacturing cost subject to 5 black-box constraints on structural properties of the beam (including shear stress, bending stress, and buckling load) my tuning 6 parameters: the welding configuration (binary), the metal material type (categorical with 4 options), and 4 ordinal parameters controlling the dimensions of the beam (Tran et al., 2019).

SVM Feature Selection This problem involves jointly performing feature selection and hyperparameter optimization for a Support Vector Machine (SVM) trained on the CTSlice UCI data set (Dua and Graff, 2017; Liu et al., 2022). This design space for this

[3] Cont. Relax. is incompatible with a categorical kernel so we use Matérn-5/2 with one-hot encoding.

problem involves 50 binary parameters controlling whether a particular feature is included or not, and 3 continuous hyperparameters of the SVM.

Cellular Network Optimization In this 30-dimensional problem, the goal is to tune the tilt (ordinal with 6 values) and transmission power (continuous) for a set of 15 antennas (Samal et al., 2021) to maximize a coverage quality metric that is a function of signal power and interference (Maddox et al., 2021) over a geographic region of interest. We use the simulator from Dreifuerst et al. (2021).

Direct Arylation Chemical Synthesis Palladium-catalysed direct arylation has generating significant interest in pharamceutical development (Davies and Morton, 2016). In this problem, the goal is maximize yield for a direct arylation chemical reaction by tuning the 3 categorical parameters corresponding to the choice of solvent, base, and ligand, as well 2 continuous parameters controlling the temperature and concentration. We fit a surrogate model to the direct arylation dataset from Shields et al. (2021) in order to facilitate continuous optimization of temperature and concentration.

Electrospun Oil Sorbent Marine oil spills can cause ecological catastrophe. One avenue for mitigating environmental harm is to design and deploy absorbent materials to capture the spilled oil. In this problem, we tune 5 ordinal parameters (3 parameters with 5 values and 2 with 4 values) and 2 continuous parameters controlling the composition and manufacturing conditions for an electrospun oil sorbent material to maximize 3 competing objectives: the oil absorbing capacity, the mechanical strength, and the water contact angle (Wang et al., 2020a).

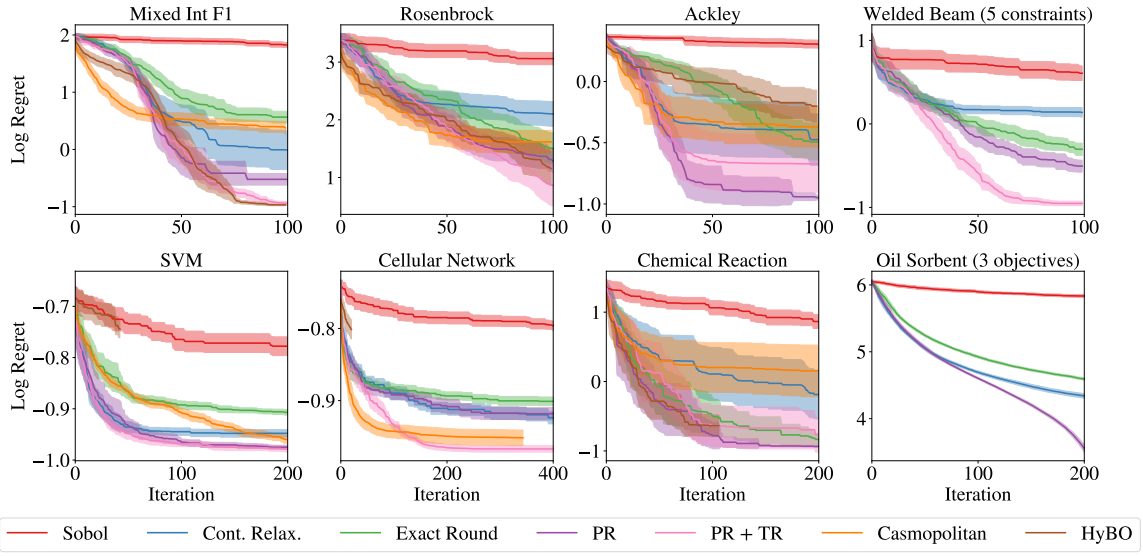

Figure 2: PR (or PR + TR) consistently outperforms alternatives with respect to log regret.

Appendix D. Theoretical Results and Proofs

D.1 Results

Let $\mathcal{P}_{\mathcal{Z}}^{(i)} := \mathcal{P}(\mathcal{Z}^{(i)})$ denote the set of probability measures on $\mathcal{Z}^{(i)}$ for each $i = 1, ..., d_z$, and let $\mathcal{P}_{\mathcal{Z}} := \prod_{i=1}^{d_z} \mathcal{P}_{\mathcal{Z}}^{(i)}$. For any $\alpha : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$, define $\tilde{\alpha} : \mathcal{X} \times \mathcal{P} \to \mathbb{R}$ as

$$\tilde{\alpha}(\boldsymbol{x}, p) = \int_{\mathcal{Z}} \alpha(\boldsymbol{x}, \boldsymbol{z}) dp(\boldsymbol{z}) = \sum_{\boldsymbol{z} \in \mathcal{Z}} \alpha(\boldsymbol{x}, \boldsymbol{z}) p(\{\boldsymbol{z}\}). \tag{6}$$

Let Θ be a compact metric space, and consider the set of functionals $\Phi = \{\varphi \; s.t. \; \varphi : \Theta \to \mathcal{P}_{\mathcal{Z}}\}$. Let

$$\hat{\alpha}(\boldsymbol{x}, \boldsymbol{\theta}) := \tilde{\alpha}(\boldsymbol{x}, \boldsymbol{\varphi}(\boldsymbol{\theta})) = \int_{\mathcal{Z}} \alpha(\boldsymbol{x}, \boldsymbol{z}) dp_{\varphi(\boldsymbol{\theta})}(\boldsymbol{z}) = \sum_{\boldsymbol{z} \in \mathcal{Z}} \alpha(\boldsymbol{x}, \boldsymbol{z}) p_{\varphi(\boldsymbol{\theta})}(\{\boldsymbol{z}\}) \tag{7}$$

Since $\mathcal{Z}$ is finite, each element of $\varphi \in \Phi$ can be expressed as a mapping from Θ to $\mathbb{R}^{|\mathcal{Z}|}$. Namely, each $\varphi(\boldsymbol{\theta})$ corresponds to a vector with $|\mathcal{Z}|$ elements containing the probability mass for each element of $\mathcal{Z}$ under $p_{\varphi}(\boldsymbol{\theta})$. Thus $(\mathcal{P}_{\mathcal{Z}}, \|\cdot\|)$ is a metric space under any norm $\|\cdot\|$ on $\mathbb{R}^{|\mathcal{Z}|}$. Let $\alpha^* := \max_{(\boldsymbol{x}, \boldsymbol{z}) \in (\mathcal{X} \times \mathcal{Z})} \alpha(\boldsymbol{x}, \boldsymbol{z})$ and let $\mathcal{H}^* := \arg\max_{(\boldsymbol{x}, \boldsymbol{z}) \in (\mathcal{X} \times \mathcal{Z})} \alpha(\boldsymbol{x}, \boldsymbol{z})$ denote the set of maximizers of α.

Lemma 1 *Suppose α is continuous in $\boldsymbol{x}$ for every $\boldsymbol{z} \in \mathcal{Z}$ and that $\varphi : \Theta \mapsto (\mathcal{P}_{\mathcal{Z}}, \|\cdot\|)$ is continuous with $\varphi(\Theta) = \mathcal{P}_{\mathcal{Z}}$. Let $\mathcal{J}^* := \arg\max_{(\boldsymbol{x}, \boldsymbol{\theta}) \in \mathcal{X} \times \Theta} \hat{\alpha}(\boldsymbol{x}, \boldsymbol{\theta})$. Then for any $(\boldsymbol{x}^*, \boldsymbol{\theta}^*) \in \mathcal{J}^*$, it holds that $(\boldsymbol{x}^*, \boldsymbol{z}) \in \mathcal{H}^*$ for all $z \in \operatorname{supp} p_{\varphi(\boldsymbol{\theta}^*)}$.*

Proof First, note that $\hat{\alpha} : \mathcal{X} \times \Theta \to \mathbb{R}$ is continuous (using that φ is continuous and α is bounded). Since both $\mathcal{X}$ and Θ are compact $\hat{\alpha}$ attains its maximum, i.e., $\mathcal{J}^*$ exists. Let $(\boldsymbol{x}^*, \boldsymbol{\theta}^*) \in \mathcal{J}^*$. Clearly, there exists $\boldsymbol{z}^* \in \arg\max_{\boldsymbol{z} \in \mathcal{Z}} \alpha(\boldsymbol{x}^*, \boldsymbol{z})$ such that $\alpha(\boldsymbol{x}^*, \boldsymbol{z}^*) = \alpha^*$. Suppose there exists $\boldsymbol{z}' \in \operatorname{supp} p_{\varphi(\boldsymbol{\theta}^*)}$ such that $(\boldsymbol{x}^*, \boldsymbol{z}') \notin \mathcal{H}^*$. Then $\alpha(\boldsymbol{x}^*, \boldsymbol{z}') < \alpha^*$ and, since $\mathcal{Z}$ is finite, $p_{\varphi(\boldsymbol{\theta}^*)}(\{\boldsymbol{z}'\}) > 0$. Consider the probability measure p' given by

$$p'(\{\boldsymbol{z}\}) = \begin{cases} 0 & \text{if } \boldsymbol{z} = \boldsymbol{z}' \\ p_{\varphi(\boldsymbol{\theta}^*)}(\{\boldsymbol{z}^*\}) + p_{\varphi(\boldsymbol{\theta}^*)}(\{\boldsymbol{z}'\}) & \text{if } \boldsymbol{z} = \boldsymbol{z}^* \\ p_{\varphi(\boldsymbol{\theta}^*)}(\{\boldsymbol{z}\}) & \text{otherwise} \end{cases}$$

Then

$$\begin{aligned} \tilde{\alpha}(\boldsymbol{x}^*, p') - \hat{\alpha}(\boldsymbol{x}^*, \boldsymbol{\theta}^*) &= \sum_{\boldsymbol{z} \in \mathcal{Z}} \alpha(\boldsymbol{x}^*, z) p'(\{\boldsymbol{z}\}) - \hat{\alpha}(\boldsymbol{x}^*, \boldsymbol{\theta}^*) \\ &= \sum_{\boldsymbol{z} \in \mathcal{Z}} \alpha(\boldsymbol{x}^*, \boldsymbol{z}) p_{\varphi(\boldsymbol{\theta}^*)}(\{\boldsymbol{z}\}) + p_{\varphi(\boldsymbol{\theta}^*)}(\{\boldsymbol{z}'\})(\alpha(\boldsymbol{x}^*, \boldsymbol{z}^*) - \alpha(\boldsymbol{x}^*, \boldsymbol{z}')) \\ &\qquad - \hat{\alpha}(\boldsymbol{x}^*, \boldsymbol{\theta}^*) \\ &= p_{\varphi(\boldsymbol{\theta}^*)}(\{\boldsymbol{z}'\})(\alpha(\boldsymbol{x}^*, \boldsymbol{z}^*) - \alpha(\boldsymbol{x}^*, \boldsymbol{z}')) \\ &> 0 \end{aligned}$$

Now $p' \in \mathcal{P}_{\mathcal{Z}}$, and so $p' = \varphi(\boldsymbol{\theta}')$ for some $\boldsymbol{\theta}' \in \Theta$. But then $\hat{\alpha}(\boldsymbol{x}^*, \boldsymbol{\theta}') > \hat{\alpha}(\boldsymbol{x}^*, \boldsymbol{\theta}^*)$. This is a contradiction. ■

Corollary 1 *Suppose the optimizer of g is unique, i.e., that $\mathcal{H}^* = \{(\boldsymbol{x}^*, \boldsymbol{z}^*)\}$ is a singleton. Then the optimizer of $\hat{\alpha}$ is also unique and $\mathcal{J}^* = \{(\boldsymbol{x}^*, \boldsymbol{\theta}^*)\}$, with $p_{\varphi(\boldsymbol{\theta}^*)}(\{\boldsymbol{z}^*\}) = 1$.*

Corollary 2 *Consider the following mappings:*

- ***Binary:*** $\varphi : [0,1] \to \mathcal{P}_{\{0,1\}}$ *with* $p_{\varphi(\theta)}(\{1\}) = \theta$ *and* $p_{\varphi(\theta)}(\{0\}) = 1 - \theta$.
- ***Ordinal:*** $\varphi : [0, C-1] \to \mathcal{P}_{\{0,1,\ldots,C\}}$ *with* $p_{\varphi(\theta)}(\{i\}) = (1 - |i - \theta|)\,\mathbf{1}\{|i - \theta| \leq 1\}$ *for* $i = 1, \ldots, C$.
- ***Categorical:*** $\varphi : [0,1]^C \to \mathcal{P}_{\{0,1,\ldots,C\}}$ *with* $p_{\varphi(\theta)}(\{i\}) = \frac{\theta_i}{\sum_{i=1}^{C} \theta_i}$.

These mappings satisfy the conditions for Lemma 1. In the setting with multiple discrete parameters where the above mappings are applied in component-wise fashion for each discrete parameter, the component-wise mappings also satisfy the conditions for Lemma 1.

Clearly, the mappings given in Corollary 2 are continuous functions of θ. In the setting with multiple discrete parameters , the component-wise function is also continuous with respect to the distribution parameters for each discrete parameter. Hence, the mappings satisfy the conditions for Lemma 1.

Lemma 2 *If* $(\boldsymbol{x}^*, \boldsymbol{z}^*) \in \mathcal{H}^* = \arg\max_{(\boldsymbol{x},\boldsymbol{z}) \in \mathcal{X} \times \mathcal{Z}} \alpha(\boldsymbol{x}, \boldsymbol{z})$, *then*

$$\alpha(\boldsymbol{x}^*, \boldsymbol{z}^*) = \max_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{Z} \sim p(\boldsymbol{Z}|\boldsymbol{\theta})}[\alpha(\boldsymbol{x}^*, \boldsymbol{Z})].$$

Proof For any $\boldsymbol{z}^*$, let $\boldsymbol{\theta}^*$ be the parameters such that $p(\boldsymbol{z}^*|\boldsymbol{\theta}^*) = 1$ (i.e. a point mass on $\boldsymbol{z}^*$). From Equation (1),

$$\mathbb{E}_{\boldsymbol{Z} \sim p(\boldsymbol{Z}|\boldsymbol{\theta}^*)}[\alpha(\boldsymbol{x}^*, \boldsymbol{Z})] = \sum_{\boldsymbol{z} \in \mathcal{Z}} \alpha(\boldsymbol{x}^*, \boldsymbol{z}) p(\boldsymbol{z}|\boldsymbol{\theta}^*) = \alpha(\boldsymbol{x}^*, \boldsymbol{z}^*).$$

Claim: $\mathbb{E}_{\boldsymbol{Z} \sim p(\boldsymbol{Z}|\boldsymbol{\theta}^*)}[\alpha(\boldsymbol{x}^*, \boldsymbol{Z})] = \max_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{Z} \sim p(\boldsymbol{Z}|\boldsymbol{\theta})}[\alpha(\boldsymbol{x}^*, \boldsymbol{Z})]$.

Suppose there exists $\boldsymbol{\theta}'$ such that $\mathbb{E}_{\boldsymbol{Z} \sim p(\boldsymbol{Z}|\boldsymbol{\theta}')}[\alpha(\boldsymbol{x}^*, \boldsymbol{Z})] > \mathbb{E}_{\boldsymbol{Z} \sim p(\boldsymbol{Z}|\boldsymbol{\theta}^*)}[\alpha(\boldsymbol{x}^*, \boldsymbol{Z})]$. Since $(\boldsymbol{x}^*, \boldsymbol{z}^*) \in \mathcal{H}^*$, $\alpha(\boldsymbol{x}^*, \boldsymbol{z}^*) = \max_{(\boldsymbol{x},\boldsymbol{z}) \in \mathcal{X} \times \mathcal{Z}} \alpha(\boldsymbol{x}, \boldsymbol{z})$. Hence, there is no convex combination of values of α that is greater than $\alpha(\boldsymbol{x}^*, \boldsymbol{z}^*)$. This is a contradiction. ■

Theorem 1 (Consistent Maximizers) *Suppose that α is continuous in $\boldsymbol{x}$ for every $\boldsymbol{z} \in \mathcal{Z}$. Let $\mathcal{H}^*$ be the maximizers of $\alpha(\boldsymbol{x}, \boldsymbol{z})$: $\mathcal{H}^* = \{(\boldsymbol{x}, \boldsymbol{z}) \in \arg\max_{(\boldsymbol{x},\boldsymbol{z}) \in \mathcal{X} \times \mathcal{Z}} \alpha(\boldsymbol{x}, \boldsymbol{z})\}$. Let $\mathcal{J}^* \subseteq \mathcal{X} \times \Theta$ be defined as: $\mathcal{J}^* = \{(\boldsymbol{x}, \boldsymbol{\theta}) \in \arg\max_{(\boldsymbol{x},\boldsymbol{\theta}) \in \mathcal{X} \times \Theta} \mathbb{E}_{\boldsymbol{Z} \sim p(\boldsymbol{Z}|\boldsymbol{\theta})}[\alpha(\boldsymbol{x}, \boldsymbol{Z})]\}$, where Θ is the domain of $\boldsymbol{\theta}$. Let $\hat{\mathcal{H}}^* \subseteq \mathcal{X} \times \mathcal{Z}$ be defined as: $\hat{\mathcal{H}}^* = \{(\boldsymbol{x}, \tilde{\boldsymbol{z}}) : (\boldsymbol{x}, \boldsymbol{\theta}) \in \mathcal{J}^*, \tilde{\boldsymbol{z}} \sim p(\boldsymbol{Z}|\boldsymbol{\theta})\}$. Then, $\hat{\mathcal{H}}^* = \mathcal{H}^*$.*

Proof From Lemma 1, we have that for any $(\boldsymbol{x}^*, \boldsymbol{\theta}^*) \in \mathcal{J}^*$, it holds that $(\boldsymbol{x}^*, \boldsymbol{z}) \in \mathcal{H}^*$ for all $z \in \operatorname{supp} p_{\varphi(\boldsymbol{\theta}^*)}$. Hence, $\hat{\mathcal{H}}^* \subseteq \mathcal{H}^*$.

Now, let $(\boldsymbol{x}^*, \boldsymbol{z}^*) \in \mathcal{H}^*$. Let $\boldsymbol{\theta}^* \in \Theta$ such that $p(\boldsymbol{z}^*|\boldsymbol{\theta}^*) = 1$. From the proof of Lemma 2, we have that $\mathbb{E}_{\boldsymbol{Z} \sim p(\boldsymbol{Z}|\boldsymbol{\theta}^*)}[\alpha(\boldsymbol{x}^*, \boldsymbol{Z})] = \alpha(\boldsymbol{x}^*, \boldsymbol{z}^*)$. As in the proof of Lemma 2, there

is no convex combination of values of α greater than $\alpha(\boldsymbol{x}^*, \boldsymbol{z}^*)$. So $\mathbb{E}_{\boldsymbol{Z}\sim p(\boldsymbol{Z}|\boldsymbol{\theta}^*)}[\alpha(\boldsymbol{x}^*, \boldsymbol{Z})] = \max_{(\boldsymbol{x},\boldsymbol{\theta})\in\mathcal{X}\times\Theta} \mathbb{E}_{\boldsymbol{Z}\sim p(\boldsymbol{Z}|\boldsymbol{\theta})}[\alpha(\boldsymbol{x}, \boldsymbol{Z})]$, and therefore, $\boldsymbol{x}^*, \boldsymbol{\theta}^* \in \mathcal{J}^*$. Hence $(\boldsymbol{x}^*, \boldsymbol{z}^*) \in \hat{\mathcal{H}}^*$. So $\mathcal{H}^* \subseteq \hat{\mathcal{H}}^*$, and hence, $\hat{\mathcal{H}}^* = \mathcal{H}^*$. ■

Lemma 3 *Suppose that $\alpha : (\boldsymbol{x}, \boldsymbol{z}) \mapsto \mathbb{R}$ is differentiable with respect to $\boldsymbol{x}$ for all $\boldsymbol{z} \in \mathcal{Z}$, and that the mapping $\varphi : \boldsymbol{\theta} \mapsto \mathcal{P}_{\mathcal{Z}}$ is such that $p_{\varphi(\boldsymbol{\theta})}(\{\boldsymbol{z}\})$ is differentiable with respect to $\boldsymbol{\theta}$ for all $z \in \mathcal{Z}$. Then the probabilistic objective $\mathbb{E}_{\boldsymbol{Z}\sim p(\boldsymbol{Z}|\boldsymbol{\theta})}[\alpha(\boldsymbol{x}, \boldsymbol{Z})]$ is differentiable with respect to $(\boldsymbol{x}, \boldsymbol{\theta})$.*

Proof For any $\boldsymbol{z} \in \mathcal{Z}$, the function $p(\boldsymbol{z}, \boldsymbol{\theta})\alpha(\boldsymbol{x}, \boldsymbol{z}) = p_{\varphi(\boldsymbol{\theta})}(\{\boldsymbol{z}\})\alpha(\boldsymbol{x}, \boldsymbol{z})$ is the product of two differentiable functions, hence differentiable. Therefore the probabilistic objective is a (finite) linear combination of differentiable functions, hence differentiable. ■

Theorem 2 (Convergence Guarantee) *Let $\alpha : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ be differentiable in $\boldsymbol{x}$ for every $\boldsymbol{z} \in \mathcal{Z}$. Let $(\hat{\boldsymbol{x}}_{t,m}, \hat{\boldsymbol{\theta}}_{t,m})$ be the best solution after running stochastic gradient ascent for t time steps on the probabilistic objective $\mathbb{E}_{\boldsymbol{Z}\sim p(\boldsymbol{Z}|\boldsymbol{\theta})}[\alpha(\boldsymbol{x}, \boldsymbol{Z})]$ from m starting points with its unbiased MC estimators proposed above. Let $\{a_t\}_{t=1}^{\infty}$ be a sequence of positive step sizes such that $0 < \sum_t^{\infty} a_t^2 = A < \infty$, where a_t is the step size used in stochastic gradient ascent at time step t. Let $\hat{\boldsymbol{z}}_{t,m} \sim p(\boldsymbol{Z}|\hat{\boldsymbol{\theta}}_{t,m})$. Then as $t \to \infty$ and $m \to \infty$, $(\hat{\boldsymbol{x}}_{t,m}, \hat{\boldsymbol{z}}_{t,m}) \to (\boldsymbol{x}^*, \boldsymbol{z}^*) \in \arg\max_{(\boldsymbol{x},\boldsymbol{z})\in\mathcal{X}\times\mathcal{Z}} \alpha(\boldsymbol{x}, \boldsymbol{z})$ in probability.*

Proof The binary and categorical mappings in Corollary 2 are differentiable in θ (the ordinal mapping is differentiable almost everywhere[4]). Since the acquisition function $\alpha : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ is differentiable in $\boldsymbol{x}$ for every $\boldsymbol{z} \in \mathcal{Z}$, this means that the PO is differentiable. Using the prescribed sequence of step sizes, optimizing the PO using stochastic gradient ascent will converge almost surely to a local maximum after a sufficient number of steps (Robbins and Monro, 1951). As we increase the number of randomly distributed starting points, the probability of not finding the global maximum of the PO will converge to zero (Wang et al., 2020b). From Theorem 1, the PO and the AF have the same set of maximizers. Hence, convergence in probability to a global maximizer of the PO means convergence in probability to a global maximizer of the AF. ■

Corollary 3 (Regret Bounds) *Let $\alpha(\boldsymbol{x}, \boldsymbol{z})$ be an acquisition function with bounded regret over a search space $\mathcal{X} \times \mathcal{Z}$. If the conditions for the regret bounds of $\alpha(\boldsymbol{x}, \boldsymbol{z})$ are satisfied, then Algorithm 1 with α enjoys the same regret bound.*

[4]Technically, the arguments presented here do not prove convergence under the ordinal mapping, but we have found this to work well and reliably in practice. Alternatively, ordinal parameters could also just be treated as categorical ones in which case the convergence results hold. In practice, however, this introduces additional optimization variables that make the problem unnecessarily hard by removing the ordered structure from the problem.

Appendix E. Experiment Details

For each BO optimization replicate, we use $N_{\text{init}} = \min(20, 2 * d_{\text{eff}})$ points from a scrambled sobol sequence, where d_{eff} is the "effective dimensionality" after one-hot encoding categorical parameters. Unless otherwise noted, all experiments use 20 replications and confidence intervals represent 2 standard errors of the mean. The same initial points are used for all methods for that replicate and different initial points are used for each replicate. For each method we report the $\log_{10}$ regret. Since the optimal value is unknown for many problems, we set the optimal value to be $\max(1.001 \cdot f^*, f^* + 0.1)$ where f^* is the best observed value across all methods and all replications. For constrained optimization f^* is the best feasible observed value and for multi-objective optimization f^* is the maximum hypervolume across all methods and replications. In total, the experiments in the main text (excluding HyBO and Casmopolitan) ran for an equivalent of 2,009.82 hours on a single Tesla V100-SXM2-16GB GPU. The baseline experiments (HyBO and Casmopolitan) ran for an equivalent of 745.10 hours on a single Intel Xeon Gold 6252N CPU.

E.1 Additional Problem Details

In this section, we describe the details of each synthetic problem considered in the experiments (the details of the remaining real-world problems are already described in Section C.2).

Ackley. We use an adapted version of the 13-dimensional Ackley function modified from Bliek et al. (2021). The function is given by:

$$f(\mathbf{x}) = -a \exp\Big(-b\sqrt{\frac{1}{d}\sum_{i=1}^{d} x_i^2)}\Big) - \exp\Big(\frac{1}{d}\sum_{i=1}^{d}\cos(cx_i)\Big) + a + \exp(1), \tag{8}$$

where in this case $a = 20, n = 0.2, c = 2\pi$ and $d = 13$ and $\mathbf{x} \in [-1, 1]^{13}$. We discretize the first 10 dimensions to be binary with the choice $\{-1, 1\}$, and the final 3 dimensions are unmodified with the original bounds.

Mixed Int F1. Mixed Int F1 is a partially discretized version of the 16-dimensional Sphere optimization problem (Hansen et al., 2019), given by:

$$f(\mathbf{x}) = \sum_{i=1}^{d}(x_i - x_{\text{opt},i})^2 + f_{\text{opt}}, \tag{9}$$

where f_{opt} is sampled from a Cauchy distribution with median $= 0$ and roughly 50% of the values between -100 and 100. The sampled f_{opt} is then clamped to be between $[-1000, 1000]$ and rounded to the nearest integer. $\mathbf{x}_{\text{opt}}$ is sampled uniformly in $[-4, 4]^d$, and in this case $d = 16$. We discretize the first 8 dimensions as follows: the first 2 dimensions are binary with 2 choices $\{-5, 5\}$; the next 2 dimensions are ordinal with 3 choices $\{-5, 0, 5\}$; the next 2 dimensions are ordinal with 5 choices $\{-5, -2.5, 0, 2.5, 5\}$; the final 2 dimensions are ordinal with 7 choices $\{-5, -\frac{10}{3}, -\frac{5}{3}, 0, \frac{5}{3}, \frac{10}{3}, 5\}$. The remaining 8 dimensions are continuous with bounds $[-5, 5]^8$.

Rosenbrock. We use an adapted version of the Rosenbrock function, given by:

$$f(\mathbf{x}) = \Big(\sum_{i=1}^{d-1} \big(100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2\big) \Big), \tag{10}$$

where in this case $d = 10$. The first 6 dimensions are discretized to be ordinal variables, with 4 possible values each $x_i \in \{-5, 0, 5, 10\} \forall i \in [1, 6]$. The final 4 dimensions are continuous with bounds $[-5, 10]^4$.

Chemical Reaction (Direct Arylation Chemical Synthesis). For this problem, we fit a GP surrogate (with the same kernel used by the BO methods) to the dataset from Shields et al. (2021) (available at `https://github.com/b-shields/edbo/tree/master/experiments/data/direct_arylation` under the MIT license) in order to facilitate continuous optimization of temperature and concentration. The surrogate is included with our source code.

Oil Sorbent. We set the reference point for this problem to be $[-125.3865, -57.8292, 43.2665]$, which we choose using a commonly used heuristic to scale the nadir point (component-wise worst objective values across the Pareto frontier) (Wang et al., 2017).

E.2 Method details

PR, Cont. Relax., Exact Round, PR + TR, and Exact Round + STE. We implemented all of these methods using BoTorch (Balandat et al., 2020), which is available under the MIT license at `https://github.com/pytorch/botorch`. All AFs are optimized via L-BFGS-B with 20 random restarts, run for a maximum of 200 iterations. We follow the default initialization heuristic in BoTorch (Balandat et al., 2020), which initializes the optimizer by evaluating the acquisition function at a large number of starting points (here, 1024, chosen from a scrambled Sobol sequence), and selecting (20) points using Botlzmann sampling (Duchon et al., 2004) of the 1024 initial points, according to their acquisition function utilities.

Combining PR with trust regions: When combining PR with the trust regions used in TuRBO we only use a trust region over the continuous parameters and discrete ordinals with at least 3 values. While methods like Casmopolitan uses a Hamming distance for the trust regions over the categorical parameters, we choose not to do so as there is no natural way of efficiently optimizing PR using gradient-based methods. Finally, we do not use a trust region over the Boolean parameters as the trust region will quickly shrink to only include one possible value. We use the same hyperparameters as TuRBO (Eriksson et al., 2019) for unconstrained problems and SCBO (Eriksson and Poloczek, 2021) in the presence of outcome constraints, including default trust region update settings.

Casmopolitan: We use the official implementation of Casmopolitan—which is available at `https://github.com/xingchenwan/Casmopolitan` under the MIT licence—but modify it where appropriate to additionally handle the ordinal variables. Specifically, the ordinal variables are treated as continuous when computing the kernel. However, during interleaved search, ordinal variables are searched via local search similar to the categorical variables. We use a set of Casmopolitan hyperparameters (i.e. success/failure sensitivity, initial trust region sizes and expansion factor) recommended by the authors.

HyBO: We use the official implementation of HyBO at `https://github.com/aryandeshwal/HyBO`, which is licensed by the University of Amsterdam. We use the default hyperparameters recommended by the authors in all the experiments, and we use the full HyBO method with marginalization treatment of the hyperparameters as it has been shown to perform stronger empirically (Deshwal and Doppa, 2021).

E.3 Gaussian process regression

When there are no categorical variables we use k_{ordinal} which is a product of an isotropic Matern-5/2 kernel for the binary parameters and a Matern-5/2 kernel with ARD for the remaining ordinal parameters. In the presence of categorical parameters, this kernel is combined with a categorical kernel (Ru et al., 2020) k_{cat} as $k_{\text{cat}} \times k_{\text{ordinal}} + k_{\text{cat}} + k_{\text{ordinal}}$. We use a constant mean function. The GP hyperparameters are fitted using L-BFGS-B by optimizing the log-marginal likelihood. The ranges for the ordinal parameters are rescaled to $[0, 1]$ and the outcomes are standardized before fitting the GP.

Table 4: Discrete random variables and their reparameterizations in terms of a Uniform random variable $U \sim \text{Uniform}(0, 1)$ and θ via a deterministic function $h(\cdot, \cdot)$.

Type	Random Variable	Reparameterization ($Z = h(\theta, U)$)
Binary	$Z \sim \text{Bernoulli}(\theta)$	$h(\theta, U) = \mathbb{1}(U < \theta)$
Ordinal	$Z = \lfloor\theta\rfloor + B$, $B \sim \text{Bernoulli}(\theta - \lfloor\theta\rfloor)$	$h(\theta, U) = \lfloor\theta\rfloor + \mathbb{1}(U < \theta - \lfloor\theta\rfloor)$
Categorical	$Z \sim \text{Categorical}(\boldsymbol{\theta})$	$h(\theta, U) = \min(\arg\max_{i=0}^{C-1} \mathbb{1}(U < \sum_c^i \theta^{(c)}))$

Appendix F. Constrained and Multi-Objective Bayesian Optimization

In many practical problems, the black-box objective must be maximized subject to $V > 0$ black-box outcome constraints $f_c^{(v)}(\boldsymbol{x}, \boldsymbol{z}) \geq 0$ for $v = 1, ..., V$. See Gardner et al. (2014) for a more in depth review of black-box optimization with black-box constraints and BO techniques for this class of problems.

In the multi-objective setting, the goal is to maximize (without loss of generality) a set of M objectives $f^{(1)}, ..., f^{(M)}$. Typically there is no single best solution, and hence the goal is to learn the Pareto frontier (i.e. the set of optimal trade-offs between objectives). In the multi-objective setting, the hypervolume indicator is a common metric for evaluating the quality of a Pareto frontier. See (Emmerich et al., 2006) for a review of multi-objective optimization.

Appendix G. Wall times

In this section, we reported the AF optimization wall times for the experiments in the main text.

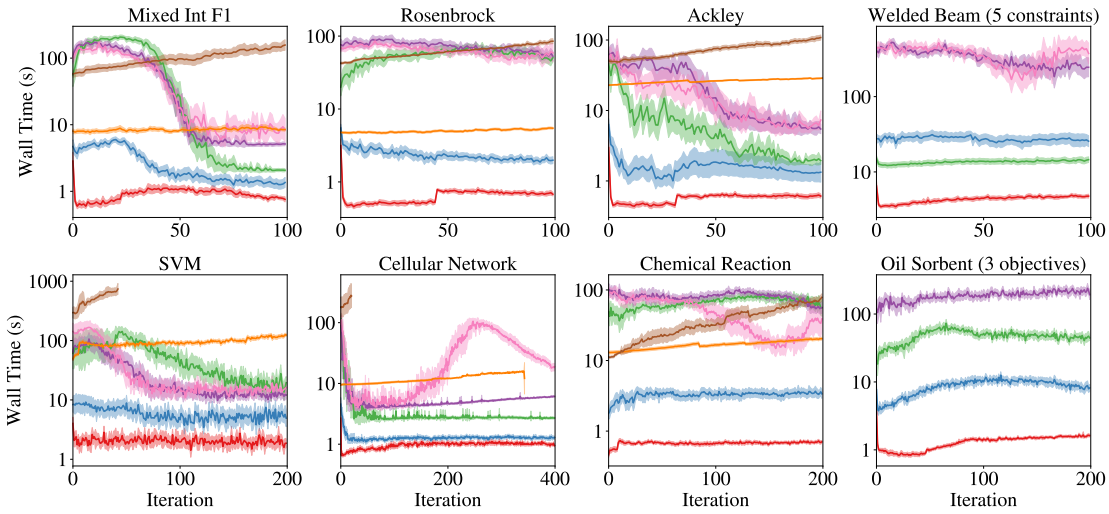

Figure 3: Wall time for candidate generation at each BO iteration in seconds. Cont. Relax., Exact Round, PR, and PR + TR are run on a single Tesla V100-SXM2-16GB GPU and other methods are run on an Intel Xeon Gold 6252N CPU.

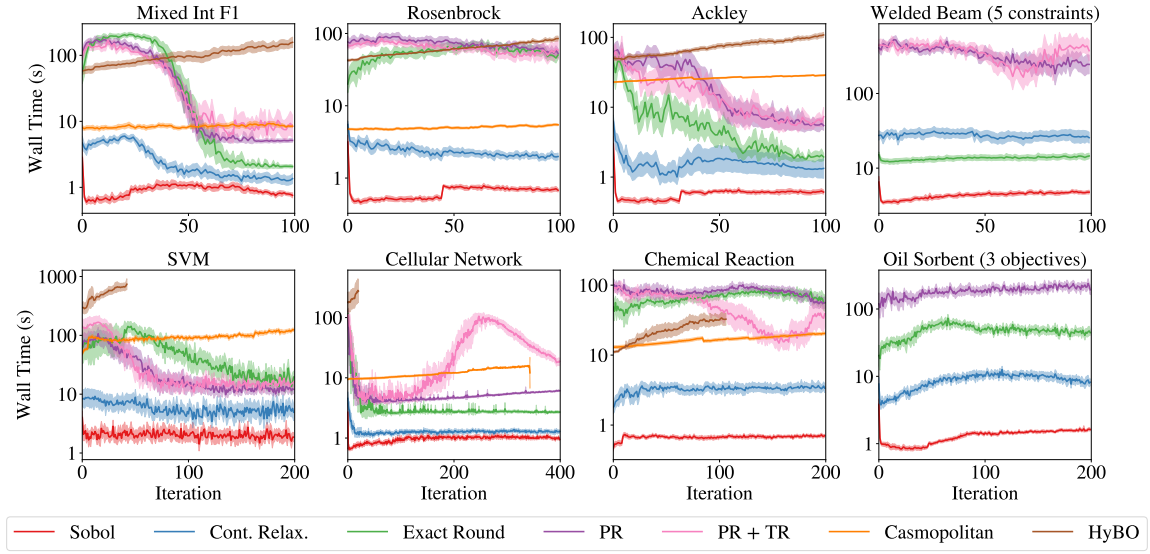

Figure 4: Wall time for candidate generation at each BO iteration in seconds. Cont. Relax., Exact Round, PR, and PR + TR are run on a single Tesla V100-SXM2-16GB GPU and other methods are run on an Intel Xeon Gold 6252N CPU.

Appendix H. Probabilistic Reparameterization Sensitivity Analyses

H.1 Monte Carlo samples

The main text considers 1024 MC for PR. We consider 128, 256, and 512 samples, in addition to the default of 1024. For problems with discrete spaces that are enumerable, we

also consider analytic PR. We do not find statistically significant differences between the final regret of any of these configurations (Figure 5). Runtime is linear with respect to MC samples, and so substantial compute savings are possible when fewer MC samples are used (Figure 6).

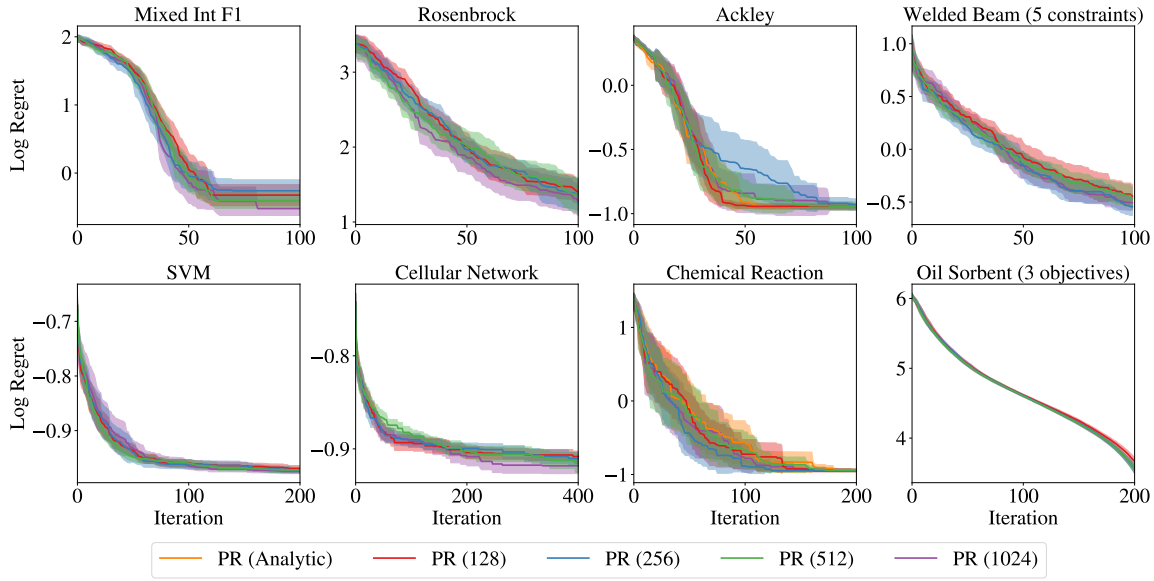

Figure 5: A sensitivity analysis of the optimization performance of PR with respect to the number of MC samples. We find that PR is robust to the number of MC samples, and the MC PR matches the performance of the analytic PR.

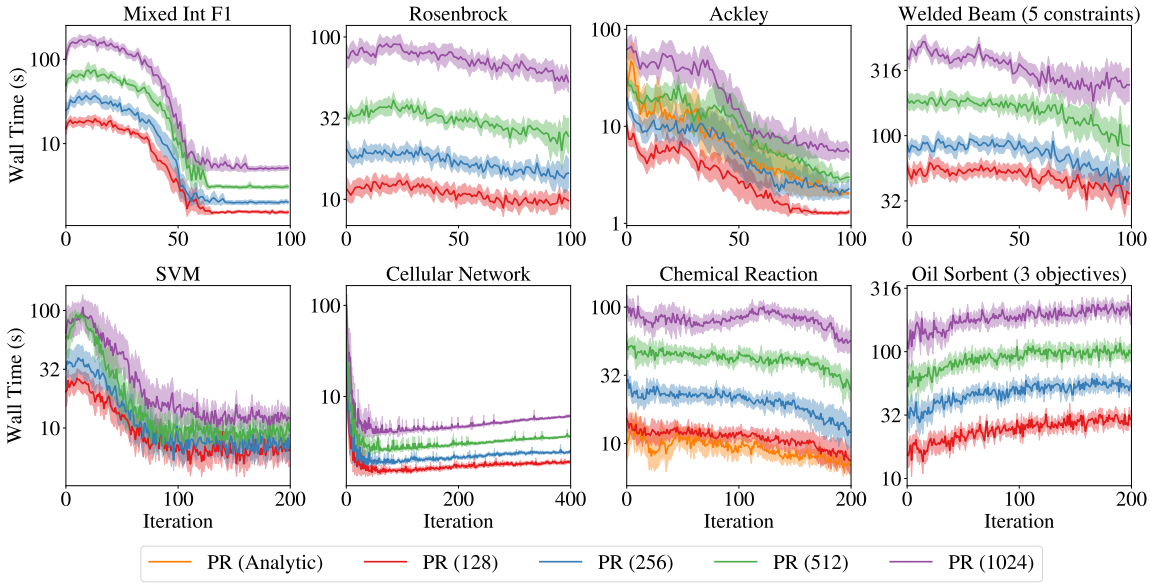

Figure 6: A sensitivity analysis of the wall time of PR with respect to the number of MC samples. We observe that wall time scales linearly with the number of MC samples, which is expected since compute PR in $\frac{N}{32}$ chunks to avoid overflowing GPU memory.

H.2 Effect of τ in Transformation

Throughout the main text, we use $\tau = 0.1$, which we selected based on the observation that it provides a reasonable balance between retaining non-zero gradients of $g(\phi)$ with respect to ϕ and allowing θ to become close to 0 or 1. This observation is obvious from Figure 7.

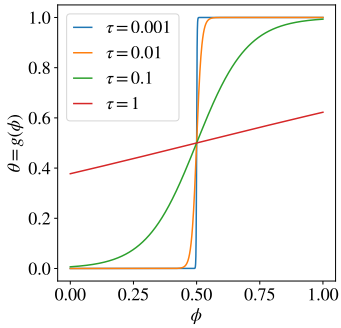

Figure 7: A comparison of the reparameterization of θ under various choices of τ. We observe that $\tau = 0.1$ provides a reasonable balance between retaining non-zero gradients of $g(\phi)$ with respect to ϕ and allowing θ to become close to 0 or 1.

We further examine the effects of this hyperparameter by testing values of τ across four orders of magnitude from 10^0 to 10^{-3} in the following sensitivity analysis. We find that our default choice of 0.1 tends to perform well across all benchmark problems considered in this work.

As $\tau \to 0$, the θ can take more extreme values, but the gradient of the transformation with respect to ϕ also moves closer to zero. For larger values of τ, the gradient of the transformation with respect to ϕ is larger, but θ has a more limited domain with less extreme values. We find that $\tau = 0.1$ is a robust setting across all experiments.

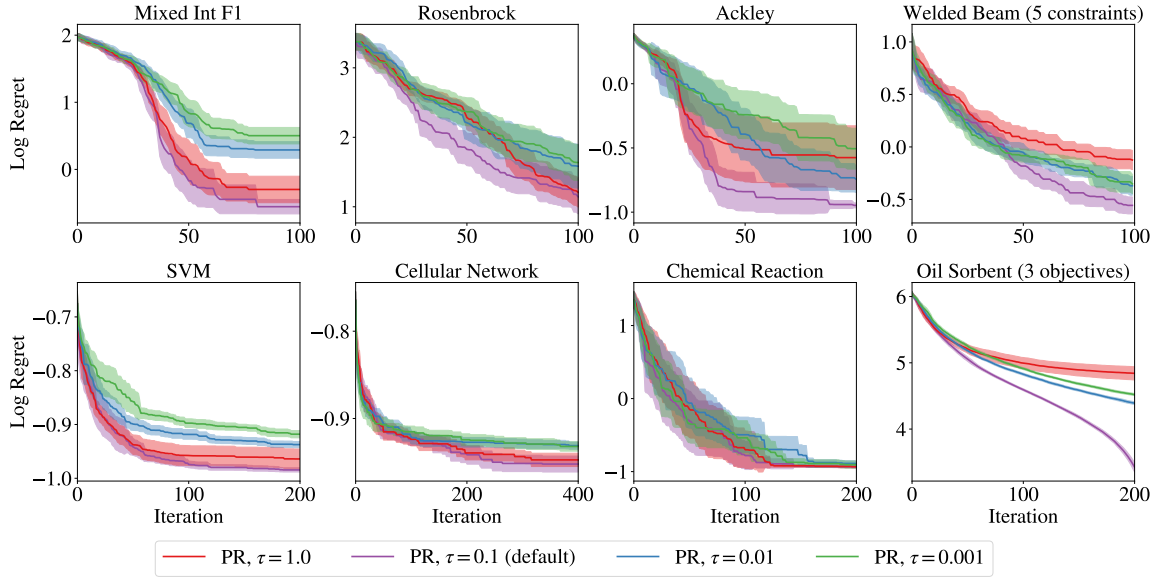

Figure 8: A sensitivity analysis of the log regret of PR with respect to the τ. We find $\tau = 0.1$ to be quite robust across experiments.

Appendix I. Alternative methods

I.1 Straight-through gradient estimators

An alternative approach to using approximating the gradients under exact rounding using finite differences is to approximate the gradients using straight-through gradient estimation (STE) (Bengio et al., 2013). The idea of STE is to approximate the gradient of a function with the identity function. In our setting, the gradient of the discretization function with respect to its input is estimated using an identity function. Using this estimator enables gradient-based AF optimization, even though the true gradient of the discretization function is zero everywhere that it is defined. Although STEs have been shown to work well empirically, these estimators are not well-grounded theoretically. Their robustness and potential pitfalls in the context of AF optimization have not been well studied. Below, we evaluate the aforementioned Exact Round + STE approach and show that it offers competitive optimization performance (Figure 9) with fast wall times (Figure 10), but does not quite match the optimization performance of PR on several benchmark problems.

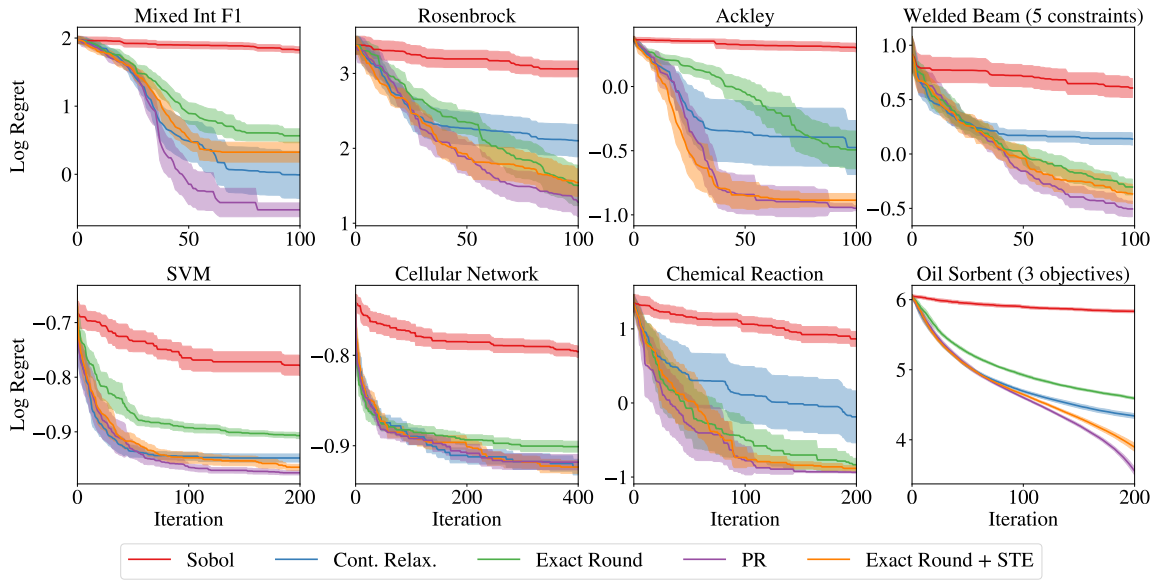

Figure 9: A comparison of exact rounding with straight-through gradient estimators versus other acquisition optimization strategies. Log regret on each problem. We report log hypervolume regret for Oil Sorbent and report the log regret of the best feasible objective for Welded beam.

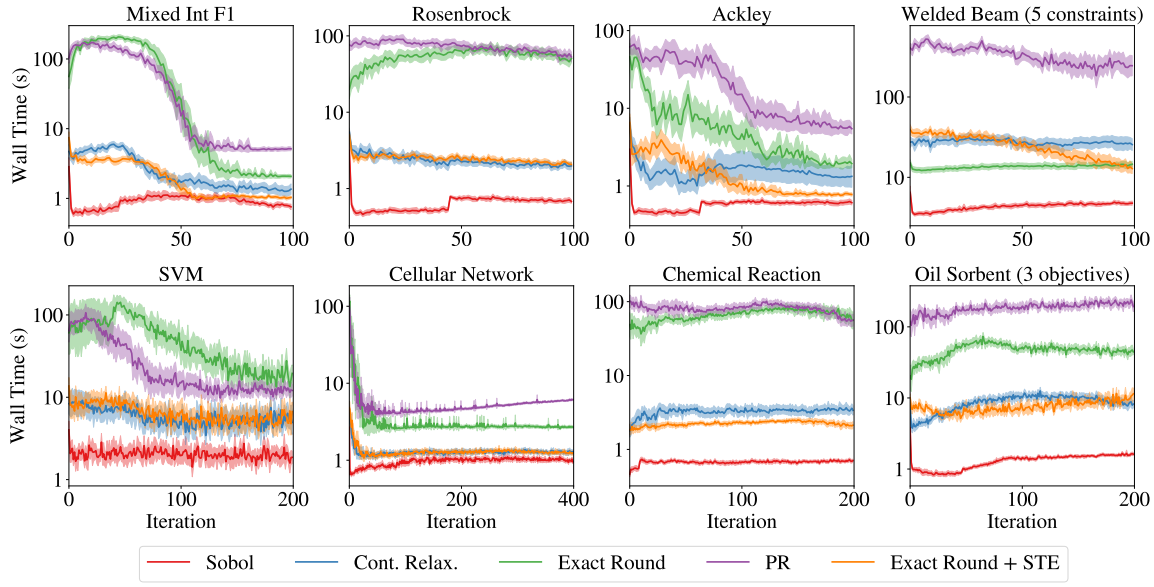

Figure 10: A comparison of wall times of exact rounding with straight-through gradient estimators versus other acquisition optimization strategies.

I.2 TR methods with alternative optimizers

In this section, we consider alternative methods to PR for optimizing AFs using within trust regions. The results in Figure 11 show that PR is a consistent best optimizer using TRs, but that STEs work quite well with TRs in many scenarios.

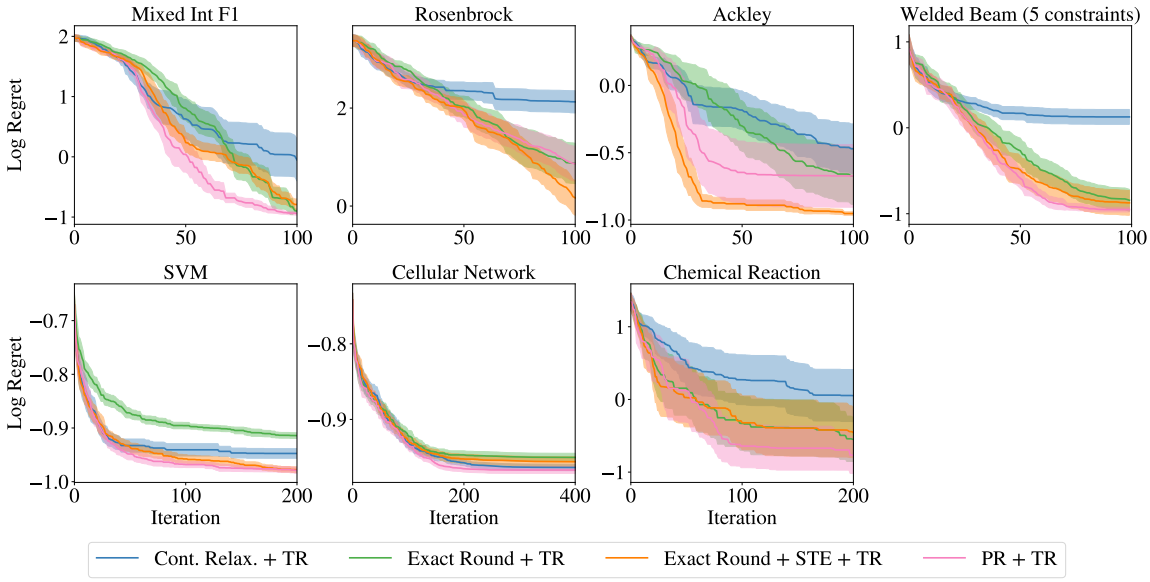

Figure 11: A comparison of TR methdods with different acquisition optimization strategies. Log regret on each problem. We report log hypervolume regret for Oil Sorbent and report the log regret of the best feasible objective for Welded beam.

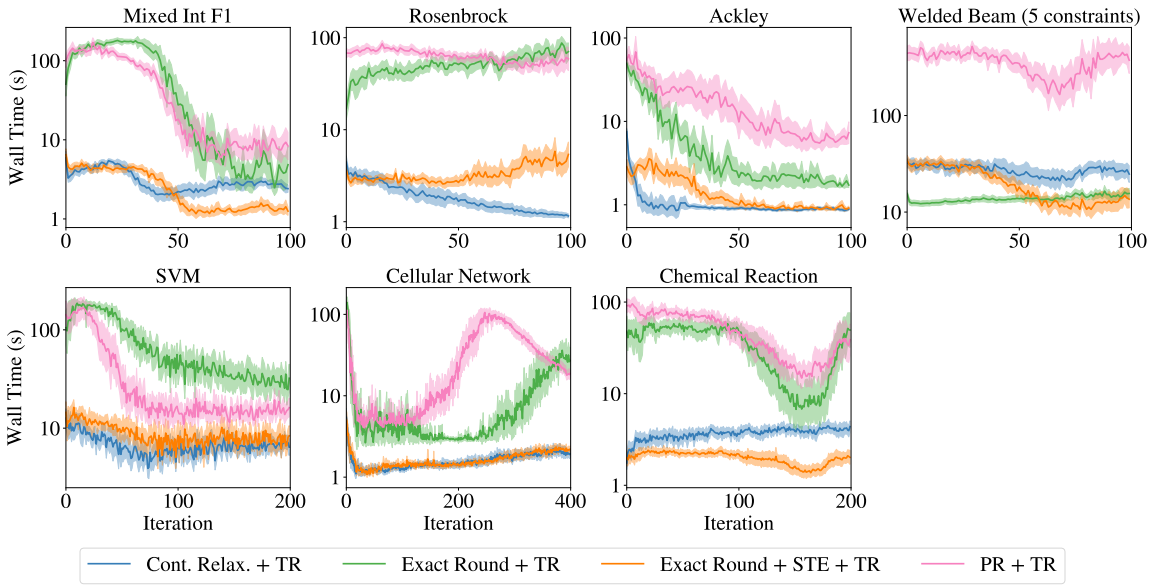

Figure 12: A comparison of wall times of TR methdods with different acquisition optimization strategies.

Appendix J. Alternative categorical kernels

In this section, we demonstrate that PR can be used with arbitrary kernels over the categorical parameters including those that require discrete inputs (which Cont. Relax. is incompatible with). Specifically for the categorical parameters, we compare using a Matérn-5/2 Kernel with one-hot encoded categoricals, a Categorical kernel (default), and a latent embedding kernel (Zhang et al., 2019). For the latent embedding kernel, we follow Pelamatti et al. (2021) and use a 1-d latent embedding for categorical parameters where the cardinality is less than or equal to 3 and a 2-d embedding for categorical parameters where the cardinality is greater than 3. For each latent embedding, we use an isotropic Matérn-5/2 kernel and use product kernel across the kernels for the categorical, binary, ordinal, and continuous parameters. The results presented in Figure 14 show that latent embedding and categorical kernels have much lower variance across replications on the chemical reaction problem.

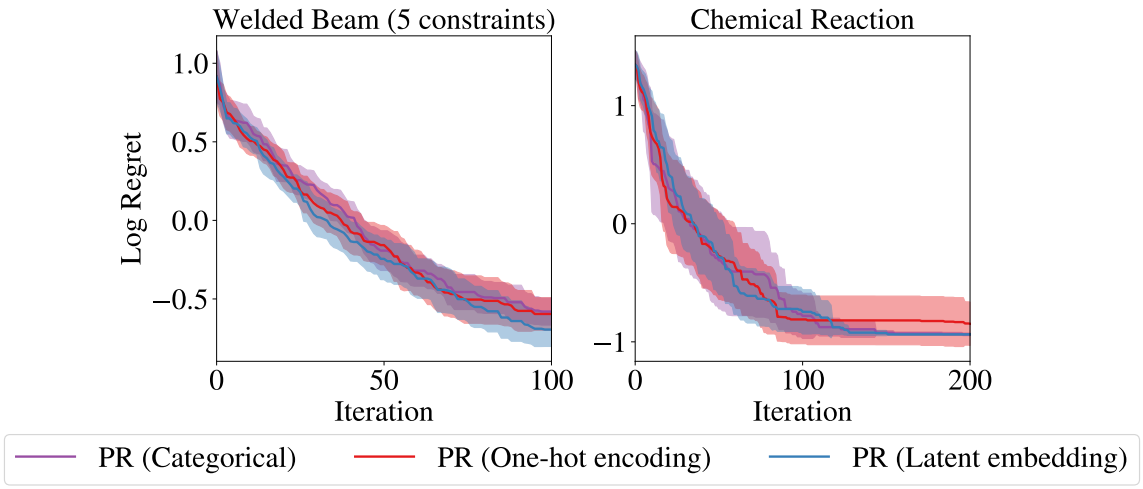

Figure 13: A comparison of different kernels over categorical parameters. Left: Welded beam has one categorical parameter, metal type (4 levels). Right: Chemical reaction has three categorical parameters, solvent, base, and ligand (with 4, 12, and 4 levels, respectively).

Appendix K. Alternative Acquisition Functions

In this section, we compare PR with expected improvement (EI) against PR with upper confidence bound (UCB). For UCB, we set the hyperparameter β in each iteration using the method in Kandasamy et al. (2015). Although UCB comes enjoys bounded regret (Srinivas et al., 2010), we find empirically that EI works better on most problems.

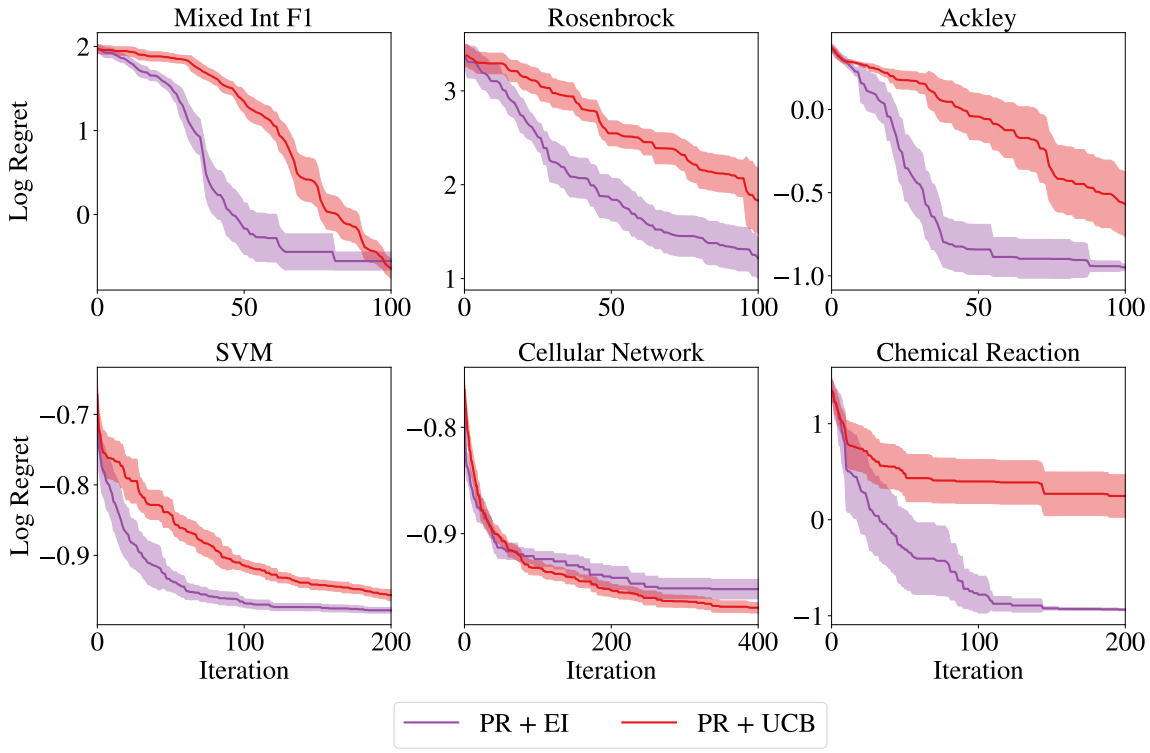

Figure 14: A comparison of different kernels over categorical parameters.

Appendix L. Additional Results on Optimizing Acquisition Functions

In this section, we provide additional results on various approaches for optimizing acquisition functions using the same evaluation procedure as in the main text. We use 50 replications.

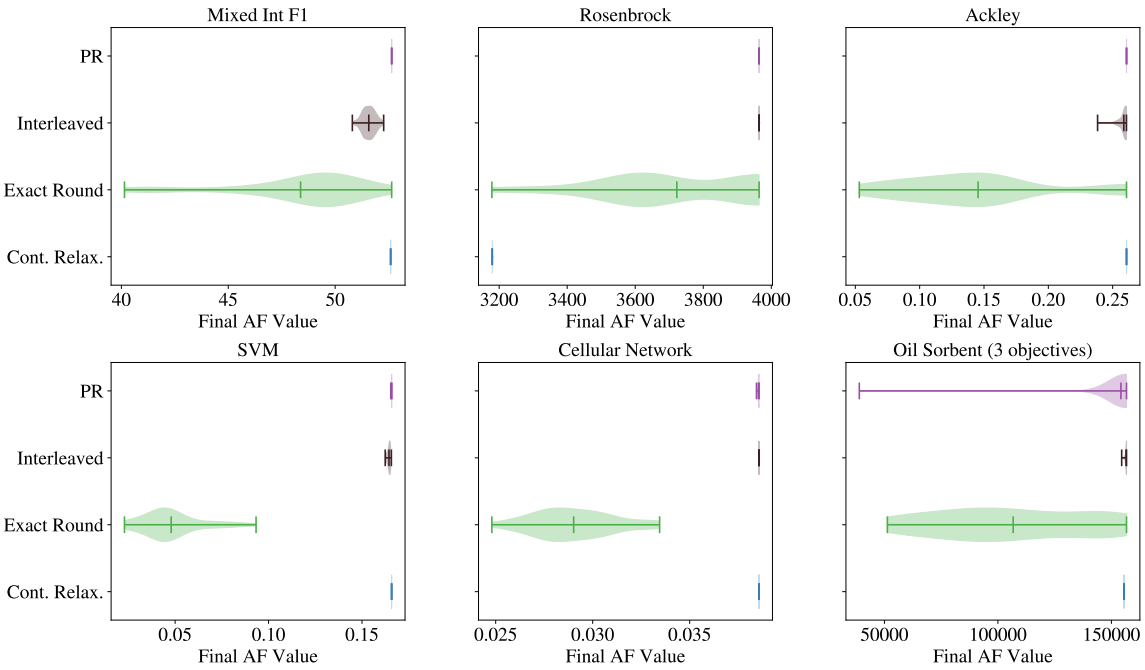

Figure 15: A comparison of methods for optimizing acquisition functions.

Appendix M. Stochastic vs Deterministic Optimization

In this section, we compare optimizing PR with stochastic and deterministic optimization methods. For stochastic methods, we use stochastic gradient ascent (SGA) with various initial learning rates between 10^{-1} and 10^{-3}. The learning rate is multiplied by 0.1 every 30 steps following the approach in Huang et al. (2017). For SGA, the MC estimators of PR and its gradient resample the base samples at each evaluation. For deterministic optimization, base samples are kept fixed. SGA is run for 200 iterations and L-BFGS-B is run for a maximum of 200 iterations. In Figure 16, we observe that SGA works well on some problems with some learning rates, but SGA is sensitive to the choice of learning rate. Furthermore, we observe that the best learning rate varies from problem to problem. Deterministic optimization consistently performs better than stochastic optimization and has lower variance across replications.

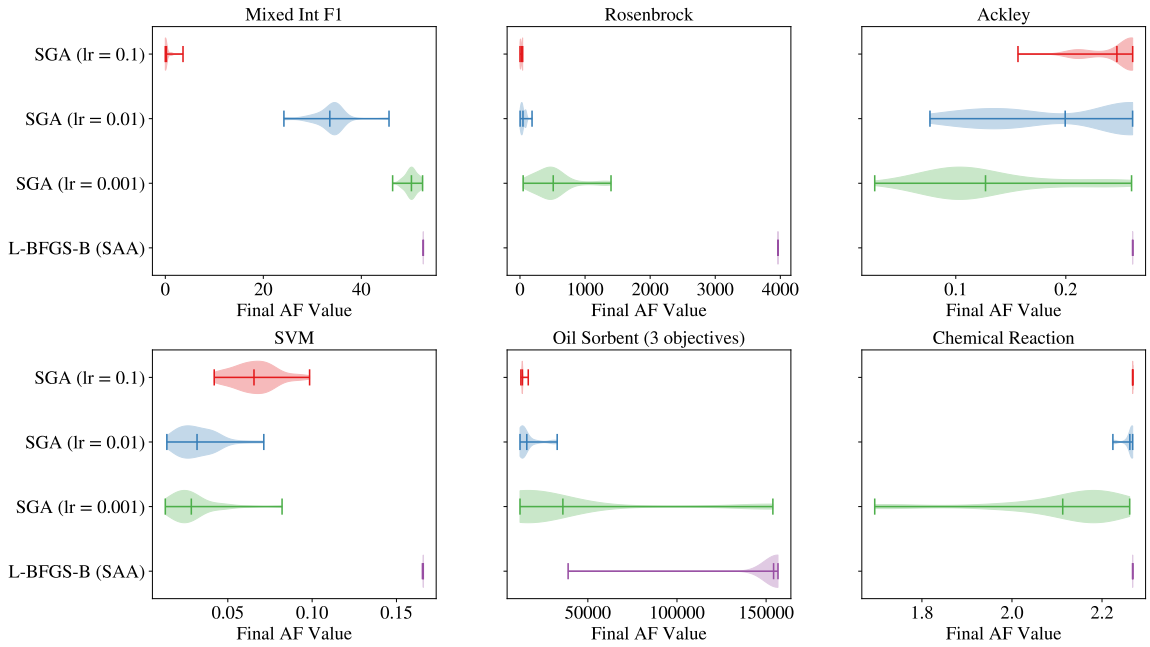

Figure 16: A comparison of PR using stochastic and deterministic optimization methods. The initial learning rate for stochastic gradient ascent is given in parentheses.

Appendix N. Related Work

Many methods for BO over discrete and mixed search spaces have been proposed. Previous work has largely focused on improving BO by either (i) improving the surrogate models or (ii) improving techniques for optimizing AFs.

Improving models: Historically, methods leveraging tree-based surrogate models, e.g., SMAC (Hutter et al., 2011) and TPE (Bergstra et al., 2011), have been popular for optimizing discrete or mixed search spaces. BOCS encodes categorical parameters as binary variables and uses Bayesian linear regression with pairwise interactions (Baptista and Poloczek, 2018). COMBO uses a diffusion kernel on the graph built from a graph Cartesian product of discrete parameters (Oh et al., 2019). MERCBO similarly exploits the combinatorial graph, but with Mercer features and Thompson sampling (Deshwal et al., 2021b). HYBO extends the diffusion kernels to mixed continuous-discrete spaces (Deshwal et al., 2021a). However, these methods scale poorly with respect to the number of data points and parameters. Moreover, of the methods listed above, only HYBO supports continuous parameters without discretization. HYBO enjoys a universal approximation property, but relies on summing over all possible orders of interactions between base kernels for each parameter which results in exponential complexity with respect to the number of parameters and limits its applicability to low-dimensional problems. GRYFFIN (Häse et al., 2021) uses kernel density estimation, but is limited to categorical search spaces. MIVABO uses a linear combination of basis functions (e.g. pseudo Boolean features (Boros and Hammer, 2002) for discrete parameters) with interaction terms(Daxberger et al., 2020). MVRSM (Bliek

et al., 2021) uses ReLU-based surrogates for computational efficiency, but is limited by the expressiveness of these models.

Optimizing acquisition functions: As discussed previously, Garrido-Merchán and Hernández-Lobato (2020) propose using continuous relaxation and discretize the inputs before evaluating the AF. However, the resulting AF after discretization is piece-wise-flat along slices of the discrete parameters and is difficult to optimize. CoCaBO (Ru et al., 2020) sample discrete parameters using a multi-armed bandit and optimize the continuous parameters conditioned on the sampled discrete parameters. However, CoCaBO's performance degrades as number of discrete configurations increases. Casmopolitan (Wan et al., 2021) uses local trust regions combined with an interleaved AF optimization strategy that alternates between local search steps on the discrete parameters and gradient ascent for the continuous parameters. Furthermore, both CoCaBO and Casmopolitan do not inherently exploit ordinal structure. Moreover, the computational issues of many existing approaches make it difficult to apply them to multi-objective and constrained optimization.

Probabilistic reparameterization: PR has been considered for optimizing discrete parameters in other domains such as reinforcement learning (Williams, 1992) and sparse regression (Yin et al., 2020). However, PR has not been leveraged for BO (although a different reparameterization is commonly used in MC AFs (Wilson et al., 2018)).

Alternative methods for propagating gradients: Alternative methods for propagating gradients through discrete structures have been considered in the deep learning community (among others). One approach is to use approximate discrete Concrete distributions (Maddison et al., 2017; Jang et al., 2017), which admit sample-path gradients. However, samples from Concrete distributions are not discrete and significant approximation error can result in pathologies similar to evaluating the AF using continuous relaxation. Moreover, approximately discrete samples prohibit using surrogate models that require discrete inputs, e.g., GPs with Hamming distance kernels (Ru et al., 2020). Another approach for gradient propagation in the deep learning community is to use straight-through gradient estimators (STE) (Bengio et al., 2013), where the gradient of the discretization function with respect to its input is estimated using an identity function. This approach works well empirically, these estimators are not well-grounded theoretically. Nevertheless, we discuss and evaluate using STE for AF optimization in Appendix I.

Appendix O. Discussion

The performance and regret properties of BO depend critically on adequately maximizing the AF. For problems with discrete features, exhaustively trying all possible combinations of discrete values quickly becomes infeasible as the number of combinations grows. Alternatives such as trying a subset of the possible combinations or resorting to continuous relaxations often leads to a failure to effectively optimizing the AF which may result in sub-optimal BO performance. As an alternative, we propose using PR which allows us to better optimize the AF and we showed that PR shows strong performance on a large number of real-world problems. Our approach is complementary to other approaches and can easily be combined with other approaches such as trust region methods and specialized kernels for discrete parameters. One limitation of PR is that it requires introduces computationally-demanding

integration. However, given that the computation in PR is embarrassingly parallel, it motivates for future research on optimizing AFs on distributed hardware.