Think Global and Act Local: Bayesian Optimisation over High-dimensional Categorical and Mixed Search Spaces

Xingchen Wan[1], Vu Nguyen[2], Huong Ha[3], Binxin Ru[1], Cong Lu[1], Michael A. Osborne[1]

[1]Machine Learning Research Group, University of Oxford, Oxford, UK
[2]Amazon, Adelaide, Australia
[3]RMIT University, Melbourne, Australia

ICML International Conference On Machine Learning

UNIVERSITY OF OXFORD amazon RMIT UNIVERSITY

Bayesian optimisation (BO)

- Bayesian Optimisation (BO) has been a successful black-box optimiser
  - Expressive surrogate models and sample efficiency
  - Useful for applications where evaluations are difficult, e.g., AutoML

Bayesian optimisation (BO)

- Bayesian Optimisation (BO) has been a successful black-box optimiser
  - Expressive surrogate models and sample efficiency
  - Useful for applications where evaluations are difficult, e.g., AutoML

- However, BO in **high dimensions** and heterogenous search spaces (i.e., some variables are not continuous) is still challenging
  - Search space grows exponentially with dimension, making GP surrogate difficult to cover
  - Categorical variables do not have natural ordering; one-hot transform makes problems even higher-dimensional
  - Some examples: combinatorial optimisation problems such as maximum satisafiability, Neural network tuning with both continuous (e.g., learning rate) and categorical (e.g., choice of optimiser) hyperparameters

Contributions

- **Our method: CASMOPOLITAN**
  - Use *local trust regions* and *tailored kernels* to effectively handle high dimensions and categorical/mixed search spaces.
  - Derive guarantee under some assumptions that the method converges.
  - Empirically show that our method achieves better performance, better sample efficiency or both.
  - Code implementation is open-sourced.

CASMOPOLITAN in Categorical Space

- Tailored kernel: GP with Exponentiated overlap kernel.

$$k_h(\mathbf{h}, \mathbf{h}') = \exp\left(\frac{1}{d_h}\sum_{i=1}^{d_h} \ell_i \delta(h_i, h'_i)\right),$$

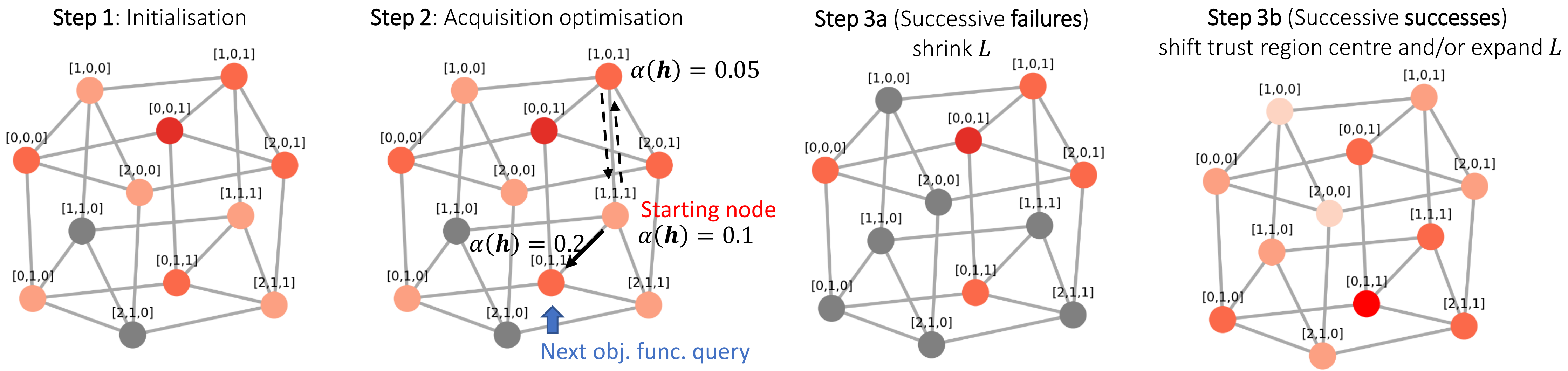
CASMOPOLITAN in Categorical Space

- Tailored kernel: GP with Exponentiated overlap kernel.

$$k_h(\mathbf{h}, \mathbf{h}') = \exp\left(\frac{1}{d_h}\sum_{i=1}^{d_h} \ell_i \delta(h_i, h'_i)\right),$$

- Use of **Trust Regions** defined in terms of Hamming distance from the best location seen so far:

$$\mathrm{TR}_h(\mathbf{h}^*)_{L^h} = \left\{\mathbf{h} \mid \sum_{i=1}^{d_h} \delta(h_i, h^*_i) \leq L^h\right\}.$$



CASMOPOLITAN in Categorical Space

- Tailored kernel: GP with Exponentiated overlap kernel.

$$k_h(\mathbf{h}, \mathbf{h}') = \exp\left(\frac{1}{d_h}\sum_{i=1}^{d_h} \ell_i \delta(h_i, h_i')\right),$$

- Use of **Trust Regions** defined in terms of Hamming distance from the best location seen so far:

$$\mathrm{TR}_h(\mathbf{h}^*)_{L^h} = \left\{\mathbf{h} \mid \sum_{i=1}^{d_h} \delta(h_i, h_i^*) \leq L^h\right\}.$$

- Trust regions are restarted using the UCB principle to ensure guarantee in terms of trust region restarts.

CASMOPOLITAN in Mixed Space

- Tailored kernel: GP now handles both categorical variables $\boldsymbol{h}$ continuous variables x.

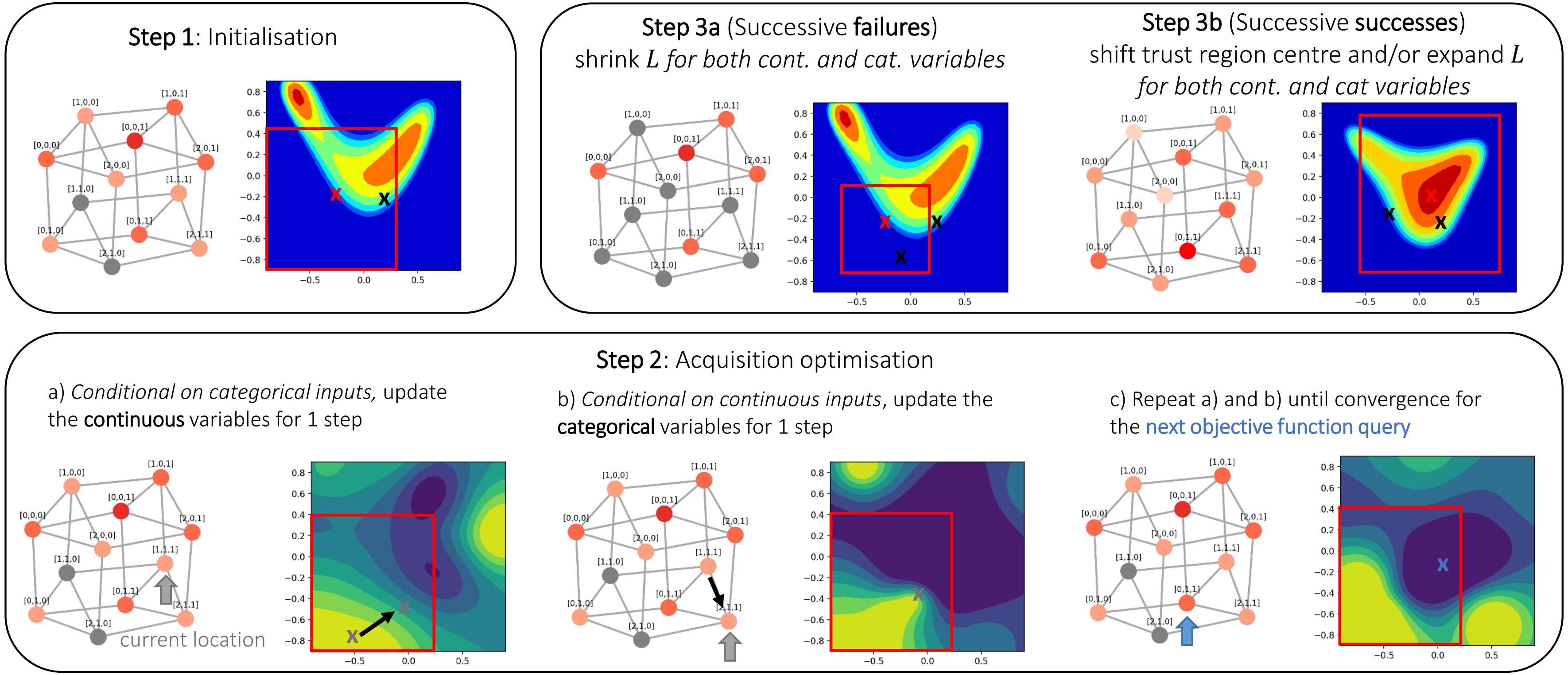
$$k(\mathbf{z},\mathbf{z}') = \lambda\Big(k_x(\mathbf{x},\mathbf{x}')k_h(\mathbf{h},\mathbf{h}')\Big) + (1-\lambda)\Big(k_h(\mathbf{h},\mathbf{h}') + k_x(\mathbf{x},\mathbf{x}')\Big),$$

CASMOPOLITAN in Mixed Space

- Tailored kernel: GP now handles both categorical variables $\boldsymbol{h}$ continuous variables x.

$$k(\mathbf{z}, \mathbf{z}') = \lambda\Big(k_x(\mathbf{x}, \mathbf{x}')k_h(\mathbf{h}, \mathbf{h}')\Big) + (1-\lambda)\Big(k_h(\mathbf{h}, \mathbf{h}') + k_x(\mathbf{x}, \mathbf{x}')\Big),$$

- Use of **separate trust regions** for both categorical and continuous variables



CASMOPOLITAN in Mixed Space

- Tailored kernel: GP now handles both categorical variables $\boldsymbol{h}$ continuous variables x.

$$k(\mathbf{z}, \mathbf{z}') = \lambda\Big(k_x(\mathbf{x}, \mathbf{x}')k_h(\mathbf{h}, \mathbf{h}')\Big) + (1-\lambda)\Big(k_h(\mathbf{h}, \mathbf{h}') + k_x(\mathbf{x}, \mathbf{x}')\Big),$$

- Use of **separate trust regions** for both categorical and continuous variables

- May be extended to handle other discrete inputs such as integer-valued and ordinal variables.

Proof sketch of guarantee derivation

- We first prove that both kernels have bounded maximum information gain.

- We make assumptions on the properties of the trust region.
  - Objective function f is bounded
  - Given a small enough region, the surrogate model may approximate f

- Using the UCB restarting of the trust region, we may derive sub-linear regret in terms of **number of restarts.**

Experiments

- Categorical problems
  - Contamination control (Hu et al, 2010): 25 binary variables and >3.35×10^7 configurations
  - Pest control (Oh et al, 2019): 25 variables, 5 choices each and > 2.98×10^{17} configurations
  - Weighted maximum satisfiability: 60 binary variables and > 1.15×10^{18} configurations

- Mixed problems
  - Func2C (2 categorical and 2 continuous) and Func3C (3 categorical and 3 continuous) (Ru et al, 2020a)
  - Hyperparameter tuning of the XGBoost model
  - 53-dimensional Ackley with 50 binary and 3 continuous dimensions
  - 200-dimensional Rosenbrock with 100 binary and 100 continuous dimensions
  - Black-box adversarial attack on CIFAR-10 on a sparse attack setup:
    - Need to choose the pixel location (categorical dimension) and the amount of noise to inject (continuous dimension)
    - 43 categorical dimension with 15 choices each, and 43 continuous dimensions.

Results

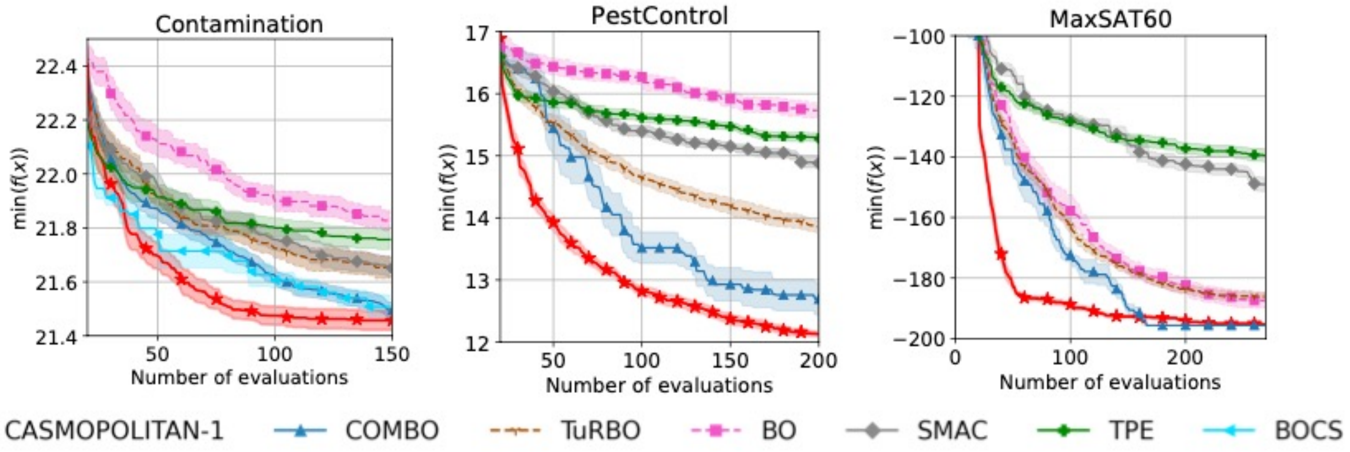


Figure 3. Results on various categorical optimisation problems. Lines and shaded area denote mean ± 1 standard error.

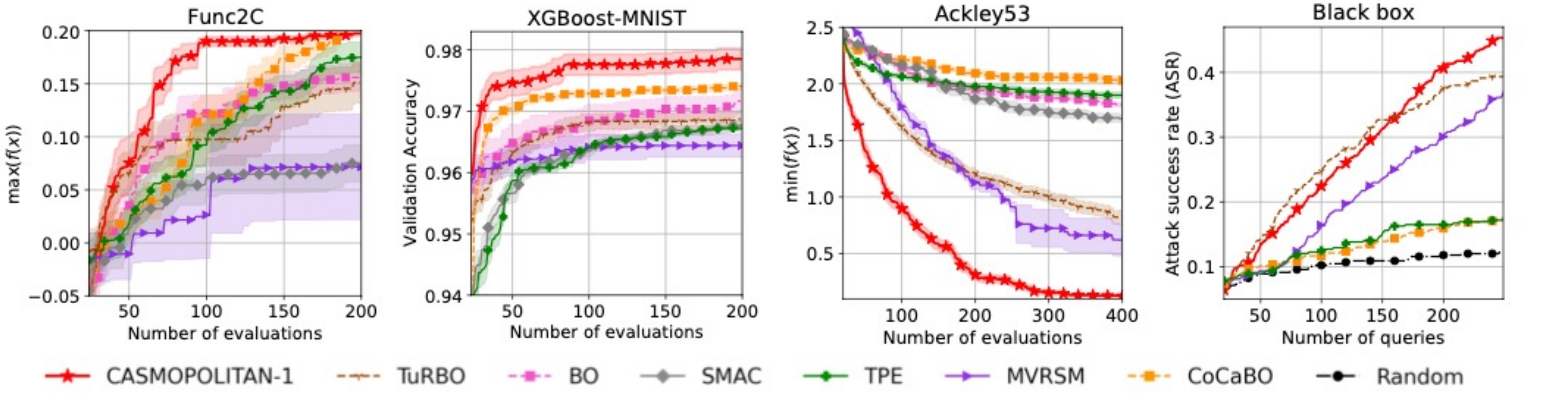


Figure 4. Results on various mixed optimisation problems. Lines and shaded area denote mean ± 1 standard error (except for Black-box where we show the ASR against number of queries). Additional experiment results in App. B.

Summary

- **CASMOPOLITAN**
  - Effective BO method applicable for high-dimensional problems that are categorical or mixed in nature
  - Use a combination of trust-region-based local optimization and tailored kernel to adapt to the setup
  - Features theoretical guarantee and state-of-the-art empirical performance

- **Future Directions**
  - Other types of structured search space: e.g., graphs, trees, conditional search spaces
  - Improvements on theories, e.g., simplifying assumptions in the certain combinatorial problems leading to better bounds

- **Paper Link**: https://arxiv.org/pdf/2102.07188.pdf
- **Code**: https://github.com/xingchenwan/Casmopolitan.
- **Email**: xwan@robots.ox.ac.uk