c-TPE: Generalizing Tree-structured Parzen Estimator with Inequality Constraints for Continuous and Categorical Hyperparameter Optimization

Shuhei Watanabe, Frank Hutter
{ watanabs | fh }@cs.uni-freiburg.de
Department of Computer Science, University of Freiburg, Germany

AutoML.org

UNI FREIBURG

Summary

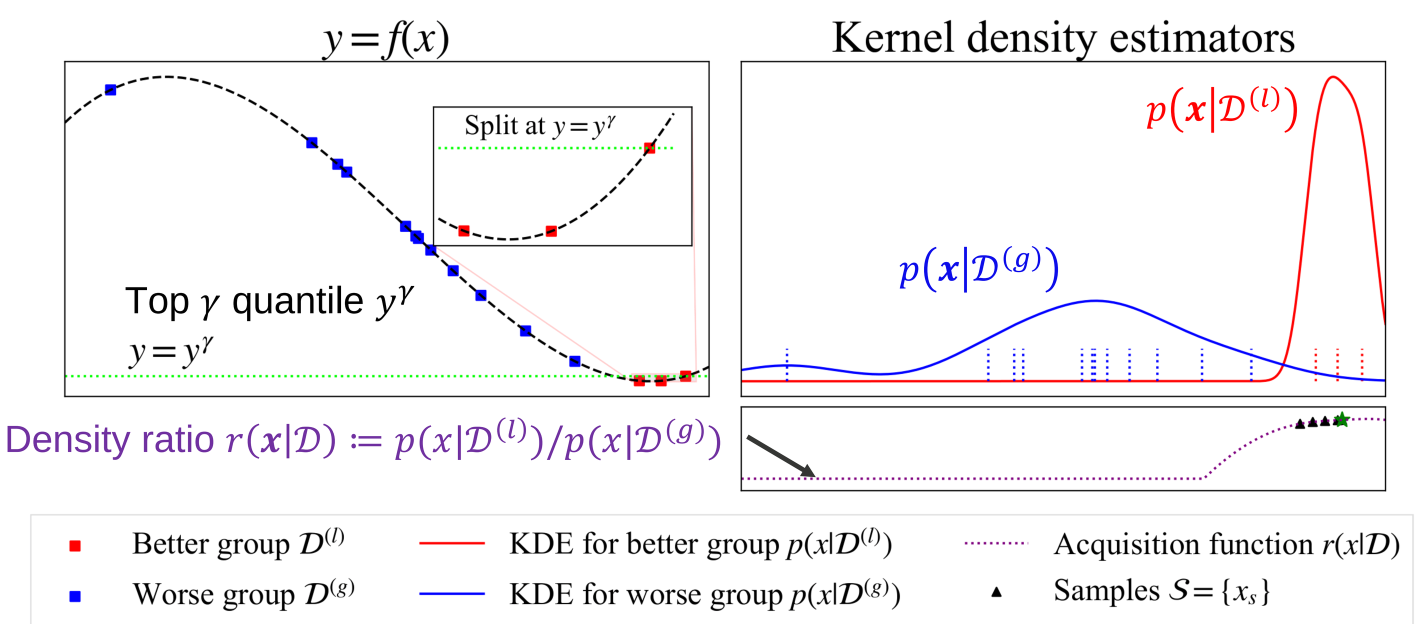
- Propose an extension of **tree-structured Parzen estimator** (**TPE**), which uses the density ratio of good and bad groups, to inequality constrained optimizations
- Integrate the **acquisition function** (**AF**) of constrained **Bayesian optimization** (**BO**) by Gardner et al.
- Modify the AF and the split of good and bad groups to enhance the performance
  1. Use relative density ratios instead of density ratio
  2. Take a certain number of feasible solutions instead of just taking top solutions
- Demonstrate that our method exhibits:
  1. much better performance than a naïve extension,
  2. the best average rank among various methods.

Tree-structured Parzen estimator (TPE)

- Assume we **minimize** $y = f(\boldsymbol{x})$ and have a set of observations $\mathcal{D} \coloneqq \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$
- Define a **lower group** $\mathcal{D}^{(l)}$ as **top-γ quantile** and a **greater group** $\mathcal{D}^{(g)}$ as **the rest**
- Build kernel density estimators (KDEs) using $\mathcal{D}^{(l)}$ and $\mathcal{D}^{(g)}$ ($N^{(l)} \coloneqq |\mathcal{D}^{(l)}|, N^{(g)} \coloneqq |\mathcal{D}^{(g)}|$):

$$p(\boldsymbol{x}|\mathcal{D}^{(l)}) = \frac{1}{N^{(l)}} \sum_{\boldsymbol{x}_n \in \mathcal{D}^{(l)}} k(\boldsymbol{x}, \boldsymbol{x}_n), p(\boldsymbol{x}|\mathcal{D}^{(g)}) = \frac{1}{N^{(g)}} \sum_{\boldsymbol{x}_n \in \mathcal{D}^{(g)}} k(\boldsymbol{x}, \boldsymbol{x}_n)$$

- At each iteration, pick the configuration with the best **density ratio** $p(\boldsymbol{x}|\mathcal{D}^{(l)})/p(\boldsymbol{x}|\mathcal{D}^{(g)})$

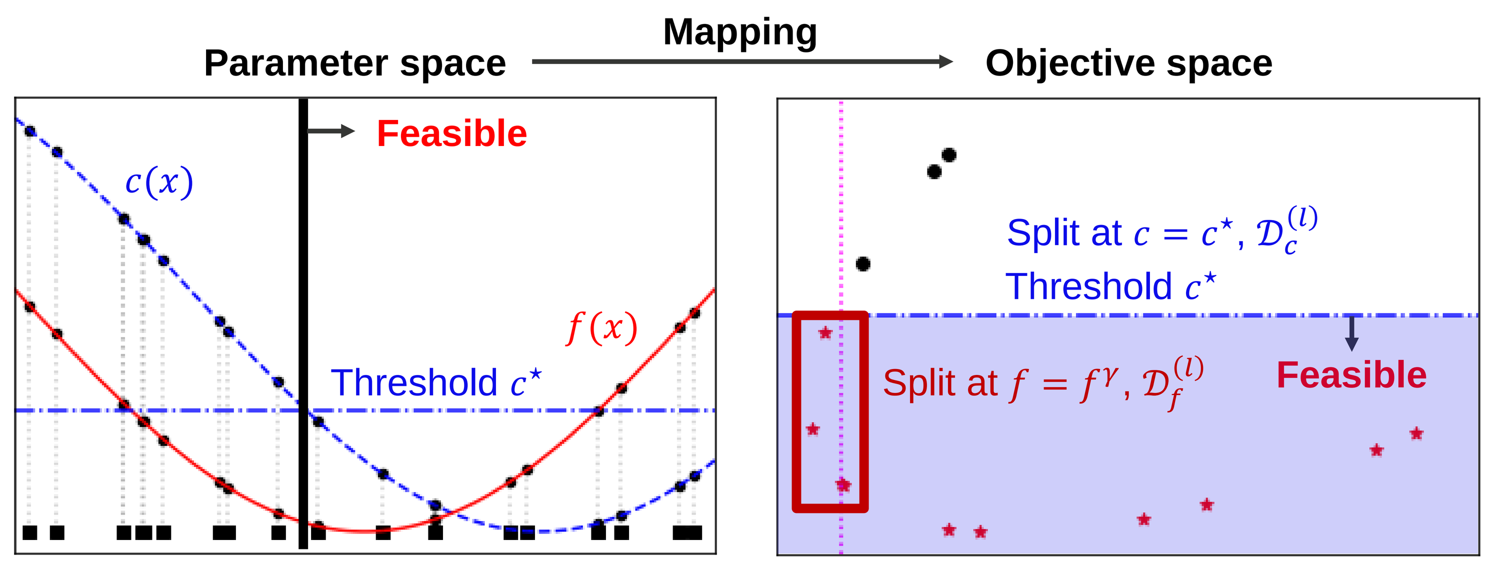


Naïve constrained TPE (Naïve c-TPE)

- The AF of TPE (density ratio) is known as **expected improvement** (**EI**), but the AF is, in fact, **probability of improvement** (**PI**) at the same time (proof in the paper)
- Constrained BO by Gardner et al. computes the AF via the product of the AFs for the objective f and constraints c_i (for $i = 1, \dots, C$) (expected constraint improvement (**ECI**))
- Hence, just taking the product of density ratios would be the naïve version:

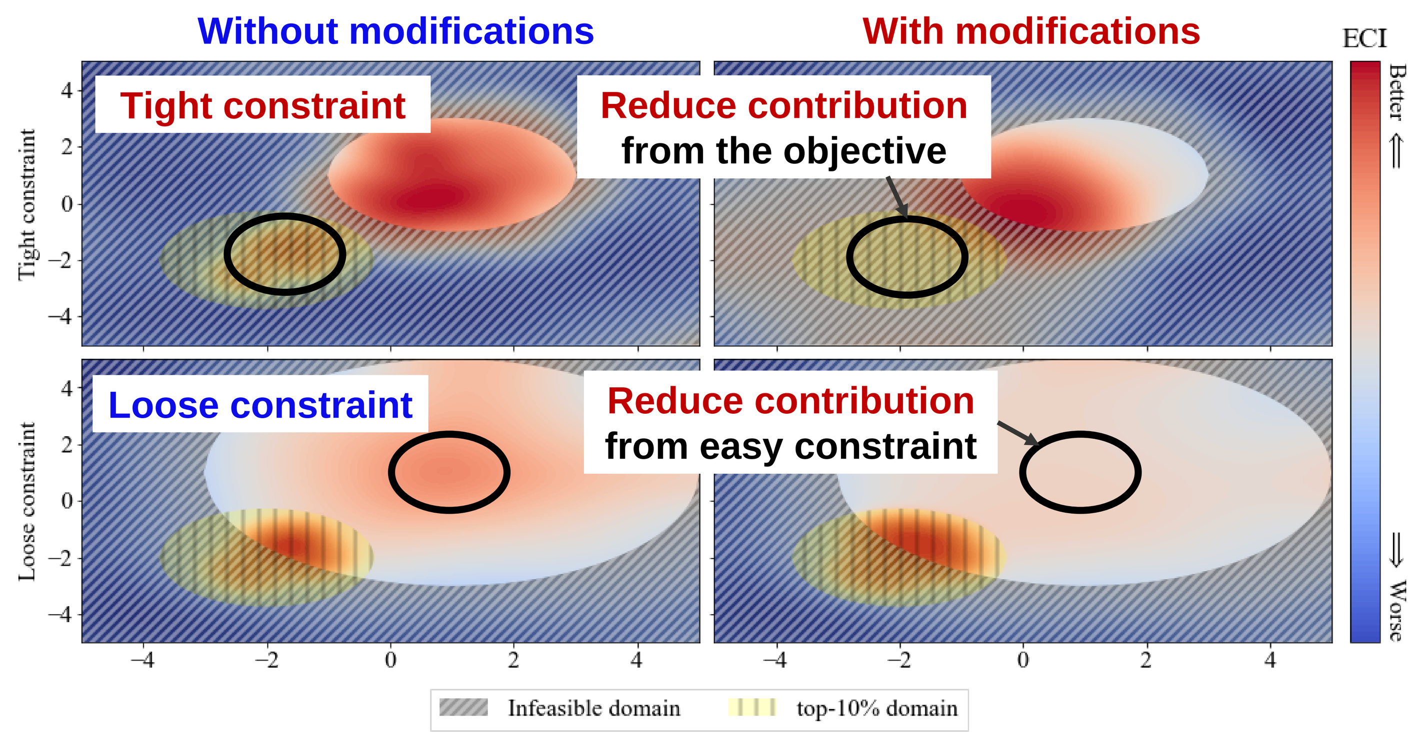
$$\prod_{i=0}^{C} r_i(\boldsymbol{x}|\mathcal{D})$$

- $r_0(\boldsymbol{x}|\mathcal{D})$ is the density ratio for f and $r_i(\boldsymbol{x}|\mathcal{D})$ ($i \in \{1, \dots, C\}$) is that for constraints
- Here is an example for the objective with one constraint $c(x) \le c^\star$
- Compute $r_0(x|\mathcal{D})$ by $p\left(x|\mathcal{D}_f^{(l)}\right)/p\left(x|\mathcal{D}_f^{(g)}\right)$ and $r_1(x|\mathcal{D})$ by $p\left(x|\mathcal{D}_c^{(l)}\right)/p\left(x|\mathcal{D}_c^{(g)}\right)$



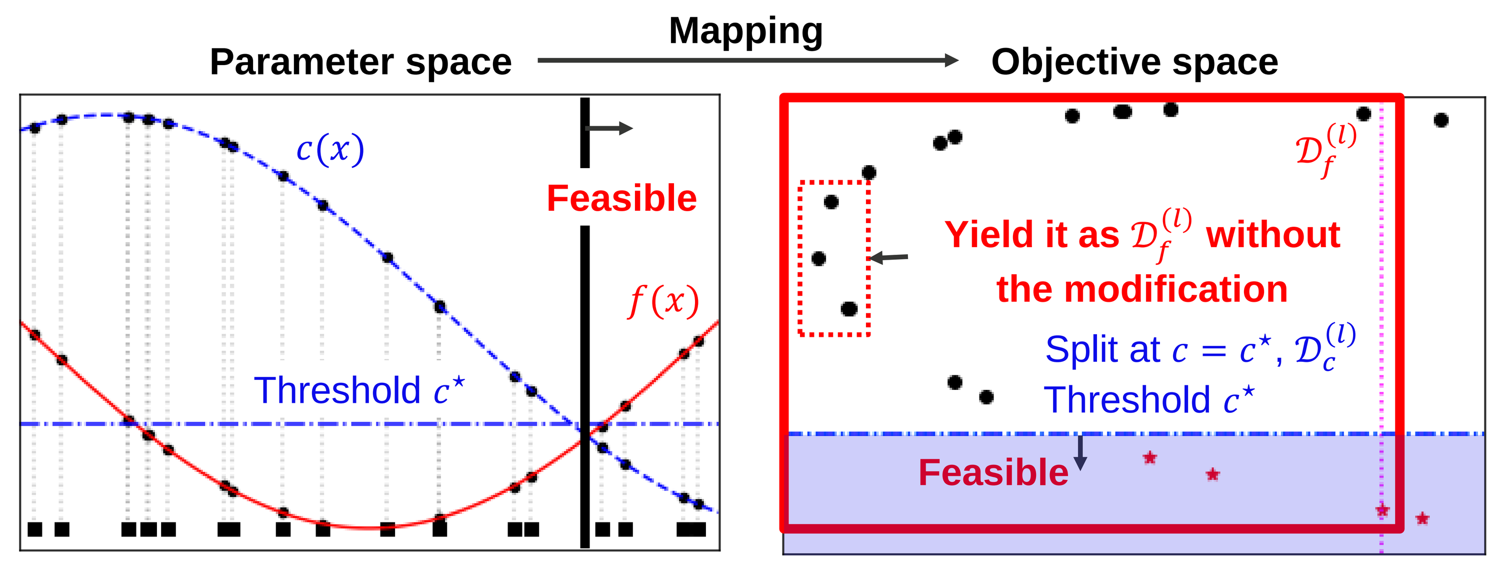
Modification I: Relative density ratio

- Use relative density ratio $r^{\mathrm{rel}}(\boldsymbol{x}|\mathcal{D}) = \frac{p(\boldsymbol{x}|\mathcal{D}^{(l)})}{\gamma p(\boldsymbol{x}|\mathcal{D}^{(l)}) + (1-\gamma) p(\boldsymbol{x}|\mathcal{D}^{(g)})}$ instead of density ratio
- $r^{\mathrm{rel}}(\boldsymbol{x}|\mathcal{D}) = r(\boldsymbol{x}|\mathcal{D})$ at $\gamma = 1$, so **generalize with TPE** when the whole domain is feasible
- For **tight constraint**, reduce the contribution from the objective
- For **loose constraint**, reduce the contribution from the constraint

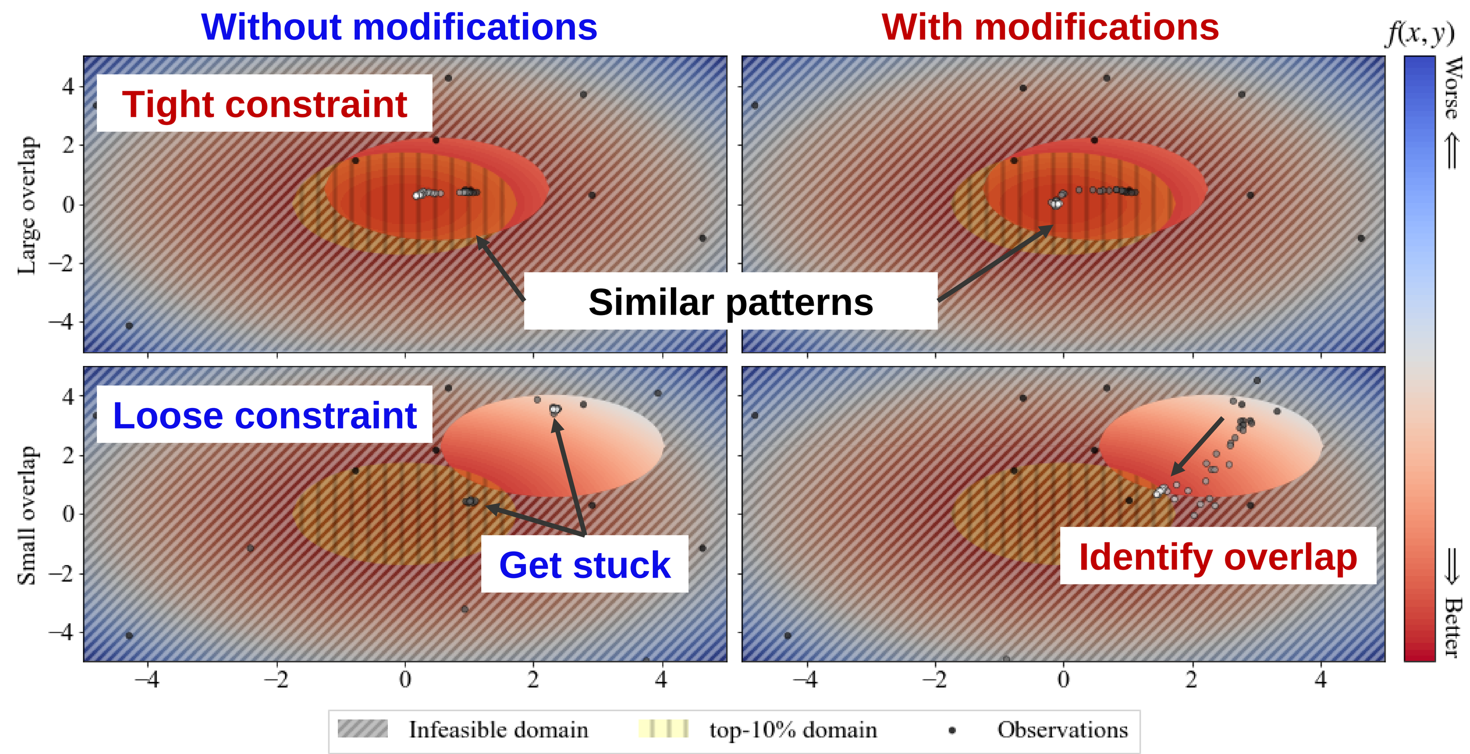


Modification II: Split algorithm

- Take until the top-γ quantile **feasible solutions** as $\mathcal{D}_f^{(l)}$ instead of the top-γ quantile solutions
- Guarantee $\mathcal{D}_f^{(l)}$ to have at least one feasible solution and thus c-TPE recognizes promising regions with feasible solutions and thus more robust

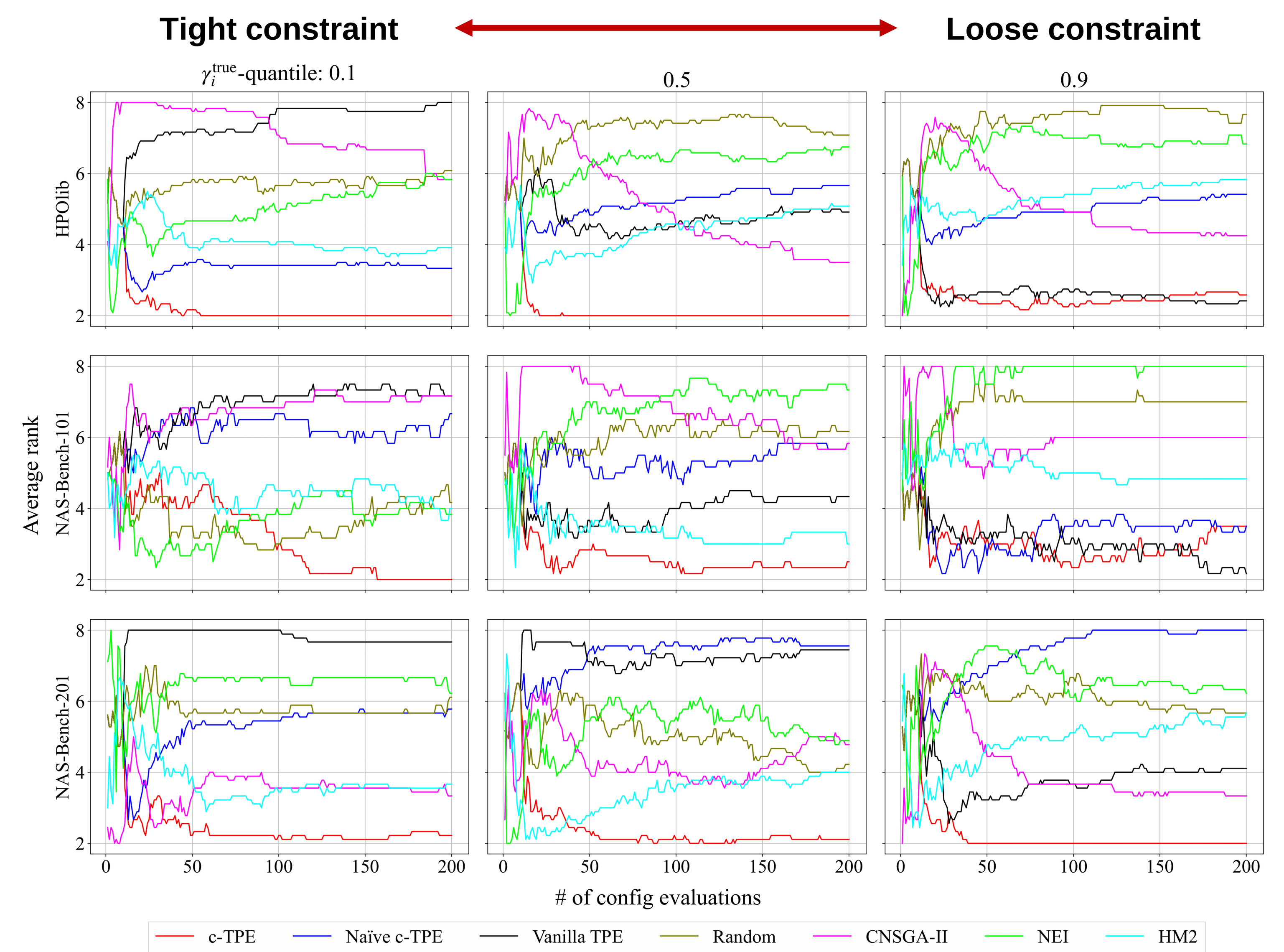


- For **large overlap** of promising regions and feasible domain is large, not a big problem
- For **small overlap**, guide to the overlap eventually



Experiments on tabular benchmarks

- **Summary of our modifications**
  - **Modification I** (relative density ratio)
    1. allow stable performance over various constraint levels
    2. generalize c-TPE with TPE when the whole domain is feasible
  - **Modification II** (new split algorithm)
    3. promote the exploration in feasible domain
    4. recover the original split when the whole domain is feasible
- **Setup**
  - 9 benchmarks: HPOlib (4 datasets), NAS-Bench-101 (2 search spaces), NAS-Bench-201 (3 datasets)
  - 3 constraints: 1. runtime, 2. network size, 3. runtime and network size
  - 9 different level of thresholds (10% is the tightest, 90% is the loosest constraint)
  - 50 different random seeds to test by the Wilcoxon signed-rank test
- **Results**
  1. Exhibit the best average rank with statistical significance over 81 settings
  2. Show stable performance (average rank) over various constraint levels (**Modification I**)
  3. Maintain the performance of the vanilla TPE, which optimizes as if there is no constraint, when the constraint level is small (**Modification I**)
  4. Demonstrate good performance on tight constraints on NAS-Bench-201, which we check it has the small overlap (**Modification II**)
  5. For high dimensions (26 dimensions in NAS-Bench-101), c-TPE did not show the distinctive performance and it might be better to search more greedily especially in loose constraint settings (90% in our case)



https://github.com/nabenabe0928/constrained-tpe