Bayesian Optimization with Improved Tree-structured Parzen Estimator

Vipul Chalotra[1, 2] **Inneke Van Nieuwenhuyse**[1, 2] **Sasan Amini**[1, 2]

1. Computational Mathematics Research Group & Data Science Institute, Hasselt University, Belgium

2. FlandersMake@UHasselt

Tree-structured Parzen Estimator (TPE) is a density-based surrogate that models **the distribution of configurations given their performance,** rather than the distribution of performance given configurations. It separates promising from non-promising regions using two density estimators and scales efficiently to **high-dimensional, mixed, and conditional search spaces.**

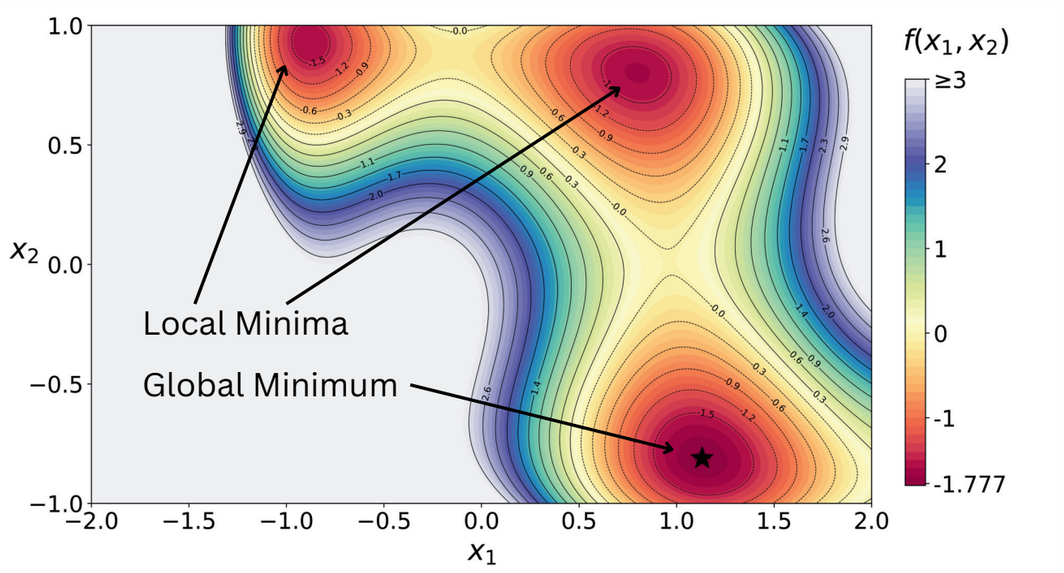
TPE is widely used for hyperparameter optimization (HPO) due to its versatility. We develop **Adaptive TPE (ATPE),** an improvement to the state-of-the-art TPE implementation in a leading HPO framework Optuna, by enhancing the Parzen Estimator (Kernel Density Estimator) with **an adaptive bandwidth approach** for continuous decision variables.

ATPE improves sampling under fixed budgets by achieving:

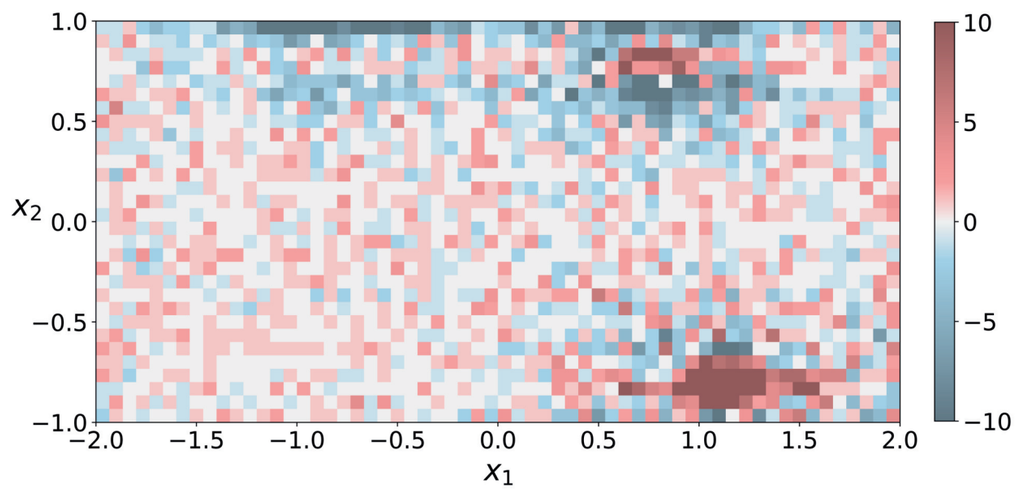
1. Better exploration of under-sampled regions → wider red spread than blue on the heatmap
2. Higher likelihood of reaching the global optimum → dark red bins around the global minimum
3. Faster convergence to an optimum → clustered red bins on and around minima

Illustrative 2D Benchmark: Modified/Harder 3-Hump Camelback

ATPE maintains broader & more balanced coverage than Optuna



ATPE globally converges in **80% of runs**, vs Optuna's 55%

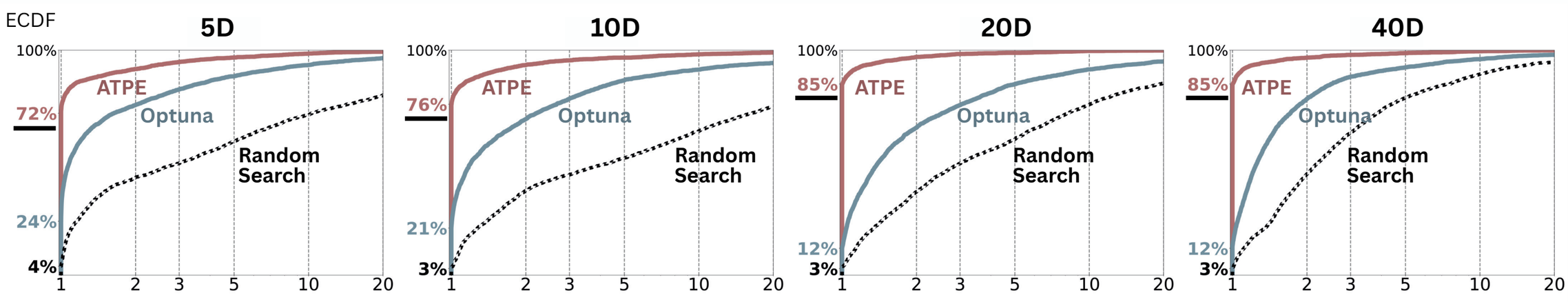


Red - ATPE sampled more **Blue - Optuna sampled more**

Combined sampling difference - 50 runs for 100 trials for both

Adaptive TPE consistently surpasses Optuna and Random Search, showing the strongest gains in higher dimensions.

Curves that rise earlier and more steeply indicate better and more consistent optimization performance.



The empirical cumulative distribution functions (ECDFs) are aggregated over 24 COCO BBOB functions, 5 instances, and 25 seeds per dimension.
The 3 y-labels (when x=1 for each dimension) represent the proportion of wins each algorithm achieved after 200 trials.