appendix B
Benchmarks and datasets

This appendix offers an overview of essential resources for optimization, including test functions, combinatorial optimization datasets, geospatial data, and machine learning datasets.

B.1 Optimization test functions

Optimization test functions, also known as *benchmark functions*, are mathematical functions used to evaluate the performance of optimization algorithms. Examples of these test functions include the following:

- *Ackley*—This is a widely used function for testing optimization algorithms. In its 2D form, it is characterized by a nearly flat outer region with a large hole at the center.
- *Bohachevsky*—This is a 2D unimodal function with a bowl shape. This function is known to be continuous, convex, separable, differentiable, nonmultimodal, nonrandom, and nonparametric, so derivative-based solvers can efficiently handle it. Note that a function whose variables can be separated is known as a *separable function*. *Nonrandom functions* contain no random variables. *Nonparametric functions* assume that the data distribution cannot be defined in terms of a finite set of parameters.
- *Bukin*—This function has many local minima, all of which lie in a ridge, and one global minimum $f(x_*) = 0$ at $x_* = f(-10,1)$. This function is continuous, convex, nonseparable, nondifferentiable, multimodal, nonrandom, and nonparametric. This raises the need to use a derivative-free solver (also known as a black-box solver) such as simulated annealing.

- *Gramacy & Lee*—This is a 1D function with multiple local minima and local and global trends. This function is continuous, nonconvex, separable, differentiable, nonmultimodal, nonrandom, and nonparametric.
- *Griewank 1D, 2D, and 3D functions*—These functions have many widespread local minima. These functions are continuous, nonconvex, separable, differentiable, multimodal, nonrandom, and nonparametric.

"Listing B.1_Optimization_test_functions.ipynb" in the book's GitHub repo shows examples of different test functions that can be implemented from scratch or retrieved from Python frameworks such as DEAP, pymoo, and PySwarms.

B.2 Combinatorial optimization benchmark datasets

"Listing B.2_CO_datasets.ipynb" in the book's GitHub repo provides examples of benchmark datasets for combinatorial optimization problems such as these:

- *Traveling salesman problem (TSP)*—Given a set of n nodes and distances for each pair of nodes, find a round trip of minimal total length visiting each node exactly once. Benchmark datasets are available at https://github.com/coin-or/jorlib/tree/master/jorlib-core/src/test/resources/tspLib/tsp .
- *Vehicle routing problem (VRP)*—Determine the optimal routes and schedules for a fleet of vehicles to serve a set of customers or locations. Benchmark datasets are available at https://github.com/coin-or/jorlib/tree/master/jorlib-core/src/test/resources/tspLib/vrp and http://neumann.hec.ca/chairedistributique/data/.
- *Job shop scheduling (JSS)*—JSS involves scheduling a set of jobs on a set of machines, where each job consists of multiple operations that must be processed on different machines in a specific order. The objective is to determine an optimal schedule that minimizes the makespan or total completion time of all jobs. Benchmark datasets are available at http://people.brunel.ac.uk/~mastjjb/jeb/orlib/files/jobshop1.txt and http://people.brunel.ac.uk/~mastjjb/jeb/orlib/files/jobshop2.txt.
- *Assembly line balancing problem (ALBP)*—ALBP addresses assigning tasks (work elements) to workstations to minimize the amount of the idle time on the line, while satisfying specific constraints. ALBP generally comprises all tasks and decisions related to equipping and aligning the productive units for a given production process before the actual assembly process can start. This encompasses setting the system capacity, which includes cycle time, number of stations, and station equipment, as well as assigning work content to productive units, which includes task assignment and determining the sequence of operations. Benchmark datasets are available at https://assembly-line-balancing.de/.
- *Quadratic assignment problem (QAP)*—QAP addresses determining the optimal assignment of a set of facilities to a set of locations. It is widely studied in operations research and has applications in various fields such as facility layout design, manufacturing, logistics, and telecommunications. Benchmark datasets are available at http://mistic.heig-vd.ch/taillard/problemes.dir/qap.dir/qap.html.

- Knapsack problem—Given a set of n items, each item i has a weight $w[i]$ and a value $v[i]$. You want to select a subset of these items such that the total weight of the selected items is less than or equal to a given weight limit W and the total value of the selected items is as large as possible. Benchmark datasets are available at http://people.brunel.ac.uk/~mastjjb/jeb/orlib/mknapinfo.html.
- *Set covering problem (SCP)*—Given a universe U of n elements and a collection S of m sets whose union equals the universe, the set covering problem is to find the smallest subcollection of S such that this subcollection still covers all elements of the universe U. Benchmark datasets are available at http://people.brunel.ac.uk/~mastjjb/jeb/orlib/scpinfo.html.
- Bin packing problem—Given a set of n items, each with a size $s[i]$, and a bin capacity C, the problem is to allocate each item to one bin so that the total size of the items in each bin does not exceed C and the number of bins used is minimized. Benchmark datasets are available at http://people.brunel.ac.uk/~mastjjb/jeb/orlib/binpackinfo.html.

B.3 Geospatial datasets

Spatial data is any type of data that directly or indirectly references specific geographic locations. Examples of this data include, but are not limited to

- Locations of people, businesses, assets, natural resources, new developments, services, and other built infrastructure
- Spatially distributed variables such as traffic, health statistics, demographics, and weather
- Data related to environmental change—ecology, sea level rise, pollution, temperature, etc.
- Data related to coordinating responses to emergencies and natural and man-made disasters—floods, epidemics, terrorism

If your optimization problems include geospatial data, you can retrieve this data from multiple online resources and open data repositories. Listings B.3 ("Listing B.3_Geospatial_data.ipynb") and B.4 ("Listing B.4_Geospatial_data_TBS.ipynb") in the book's GitHub repository show examples of how to fetch data from open sources such as OpenStreetMap (OSM), Overpass API, Open-Elevation API, etc.

B.4 Machine learning datasets

Neural combinatorial optimization has been employed on a variety of datasets. However, since this field is often concerned with solving classic optimization problems, benchmark datasets are often just standard instances of these problems. Besides the datasets included in listing B.2 ("Listing B.2_CO_datasets.ipynb"), listing B.5 ("Listing B.5_ML_datasets.ipynb") in the book's GitHub repository provides examples of datasets for neural combinatorial optimization, such as these:

- *Convex hull*—The problem of finding or computing a convex hull, given a set of points. A convex hull is a geometric shape, specifically a polygon, that fully encompasses a given set of points. It achieves this by optimizing two distinct parameters: it maximizes the area that the shape covers while simultaneously minimizing the boundary or circumference of the shape. Data is available in the book's GitHub repo (in the appendix B data folder).
- *TSP*—Dataset for training and testing TSP using pointer networks. Data is available in the book's GitHub repo (in the appendix B data folder).
- *TLC trip record data*—Yellow and green taxi trip records include fields capturing pick-up and drop-off dates and times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. Data is available at www.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

B.5 Data folder

The data folder included in the GitHub repo of the book (https://github.com/Optimization-Algorithms-Book/Code-Listings/tree/main/Appendix B/data) includes the following sample data:

- *ALBP*—Dataset for the assembly line balancing problem in chapter 6
- *AdministrativeBoundaries*—Administrative boundaries of different regions of interest around the world in geoJSON format, which can be used for map visualization
- *BikeShare*—Bike Share Toronto (TBS) ridership data containing anonymized trip data, including trip start day and time, trip end day and time, trip duration, trip start station, trip end station, and user type (see listing B.4: Listing B.4_Geospatial_data_TBS.ipynb)
- *CanadaTraffic*—Data from Statistics Canada's dataset for motor vehicle collisions in 2018, such as fatalities per 100,000 population, fatalities per billion vehicles-kilometers, injuries per billion vehicles-kilometers, fatalities per 100,000 licensed drivers, and injuries per 100,000 licensed drivers (see listing A.2: Listing A.2_Graph_libraries.ipynb)
- *OntarioHealth*—Health regions of Ontario, Canada (see listing A.2: Listing A.2_Graph_libraries.ipynb)
- *Police*—Toronto Police Service public safety data (see listing A.2: Listing A.2_Graph_libraries.ipynb)
- *PoliticalDistricting*—Political districting data used in chapter 8
- *PtrNets*—Convex hull and TSP data used by pointer networks (chapter 11)
- *TSP*—Traveling salesman problem instances