## A BENCHMARK PROBLEM SET

The problems considered in Section 5 are described in paper [2] and implemented in pure Python except for problems PB2.40 and PB2.41 where we used routines from the numpy library. Also note that in references [ $2,15,18$ ] there is a replication error defining the $g_{1}$ constraint function for the Himmelblau (HB) problem: the coefficient attached to the term $x_{1} x_{4}$ is given as 0.00026 whereas it should be 0.006262 [16] and we remind the correct description below:

Problem HB (Himmelblau's nonlinear optimization problem)

$$
\begin{array}{cl}
\underset{\boldsymbol{x}}{\operatorname{minimize}} & f(\boldsymbol{x})=5.3578547 x_{3}^{2}+0.8356891 x_{1} x_{5} \\
& +37.293239 x_{1}-40,792.141 \\
\text { subject to } \quad 0 \leq g_{1}(\boldsymbol{x}) \leq 92 \\
& 90 \leq g_{2}(\boldsymbol{x}) \leq 110 \\
& 20 \leq g_{3}(\boldsymbol{x}) \leq 25 \\
& 78 \leq x_{1} \leq 102 \\
& 33 \leq x_{2} \leq 45 \\
& 27 \leq x_{i} \leq 45(i=3,4,5) \\
g_{1}(\boldsymbol{x})=85.334407+0.0056858 x_{2} x_{5}+0.006262 x_{1} x_{4}-0.0022053 x_{3} x_{5} \\
g_{2}(\boldsymbol{x})=80.51249+0.0071317 x_{2} x_{5}+0.0029955 x_{1} x_{2}+0.0021813 x_{3}^{2} \\
g_{3}(\boldsymbol{x})=9.300961+0.0047026 x_{3} x_{5}+0.0012547 x_{1} x_{3}+0.0019085 x_{3} x_{4}
\end{array}
$$

## B FITTING A LOG-NORMAL DISTRIBUTION TO RUNTIMES DATA

For the active (1+1)-CMA-ES the paper [2] reports the $10-$ th, 50 -th and 90 -th quantiles, $\left\{m_{10}, m_{50}, m_{90}\right\}$. We propose to fit a model of the entire ECDF which is exact in these 3 quantiles. A convenient model for the runtime of an algorithm is a positive random variable which is the product of a high number of small factors, hence the choice of the Log-normal distribution.

A random variable $X$ follows the Log-normal distribution $X \sim$ $\log \mathcal{N}\left(\mu, \sigma^{2}\right)$ if $Y=\ln X$ follows the normal distribution $Y \sim$
$\mathcal{N}\left(\mu, \sigma^{2}\right)$. We are looking for the (estimated) parameters of the $\log$-normal $\hat{\mu}, \hat{\sigma}$.

First fit the median:

$$
\begin{equation*}
\hat{\mu}=\ln m_{50} \tag{9}
\end{equation*}
$$

Then solve in $\sigma$ for the repartition function at the $10-$ th and 90 -th quantiles:

$$
\begin{equation*}
\frac{1}{2}+\frac{1}{2} \operatorname{erf}\left(\frac{\ln x-\mu}{\sigma \sqrt{2}}\right)=c \tag{10}
\end{equation*}
$$

i.e. solve Equation (10) for $\mu=\hat{\mu}$ and $(x, c) \in\left\{\left(m_{10}, 0.1\right),\left(m_{90}, 0.9\right)\right\}$, where erf is the Gauss error function

The solution is:

$$
\begin{equation*}
\hat{\sigma}_{c}=\frac{1}{\sqrt{2}}(\ln x-\mu)\left[\operatorname{erf}^{-1}(2 c-1)\right]^{-1} \tag{11}
\end{equation*}
$$

where $\mathrm{erf}^{-1}$ is the inverse function of erf.
We obtain two values $\hat{\sigma}_{10}, \hat{\sigma}_{90}$ which are generally different because the 3 data points do not fit perfectly a log-normal distribution).

If we denote by $\Phi(x, \mu, \sigma)$ the cumulative distribution function of a log-normally distributed random variable with parameters $(\mu, \sigma)$, then our final estimate for the distribution of the runtimes, given $\left\{m_{10}, m_{50}, m_{90}\right\}$, would be:

$$
\begin{equation*}
\mathbb{1}_{\left\{x \leq m_{50}\right\}} \Phi\left(x, \hat{\mu}, \hat{\sigma}_{10}\right)+\mathbb{1}_{\left\{x>m_{50}\right\}} \Phi\left(x, \hat{\mu}, \hat{\sigma}_{90}\right) \tag{12}
\end{equation*}
$$

## C STATISTICAL COMPARISON OF ALGORITHMS DATA

Section 5 provides runtime data for many algorithm variants on several problems in the form of ECDF. Here, we test observed runtime differences for statistical significance. For each problem and each pair of algorithms, we compute the $p$-value from the two-sided Wilcoxon-Mann-Whitney test. If an algorithm fails, its runtime is replaced with $+\infty$, such that it's always worse than a successful run, and comparing two failing runs results in equality. We also apply a Bonferroni correction of $\beta=8$, that is the number of test problems. If algorithm $A$ is better than algorithm $B$ for a given problem with a $p$-value lower than $0.05 / \beta$, we present the item in the corresponding entry, see Table 2.


Table 2: Comparing the different algorithms' on the proposed benchmark test set. If the test is significant, meaning algorithm in row is better than algorithm in column $(p<0.05 / \beta)$ for a test problem, we display the problem name and $\log _{10} p$. If $p<10^{-3} / \beta$, it is displayed as bold text.

