Supplementary Material - Efficient Bayesian Optimization for Target Vector Estimation

A Derivation of the approximate $NC\chi^2$ distribution

Let $\mathbf{y} = [y_1, y_2, \cdots, y_K]^T$ be independent normally distributed variables:

$$\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}^2)),$$

for means $\boldsymbol{\mu} = [\mu_1, \mu_2, \cdots, \mu_K]^T$ and standard deviations $\boldsymbol{\sigma} = [\sigma_1, \sigma_2, \cdots, \sigma_K]^T$. Furthermore define $d = \|\mathbf{y} - \mathbf{y}^*\|_2^2$ for (non-random) target vector $\mathbf{y}^* = [y_1^*, y_2^*, \cdots, y_K^*]^T$. For the individual terms in

$$d = \sum_{k=1}^{K} (y_k - y_k^*)^2 = \sum_{k=1}^{K} d_k^2 \,,$$

we then have $d_k \sim \mathcal{N}\left(\mu_k - y_k^*, \sigma_k^2\right)$. When $\sigma_k = 1$ for all k then d will, by definition, follow the noncentral χ^2 distribution:

$$d \sim NC\chi^2(K, \lambda')$$
$$\lambda' = \sum_{k=1}^{K} (\mu_k - y_k^*)^2 \,,$$

with K degrees of freedom and noncentrality parameter λ' defined as the sum of squared means. We approximate this case by re-scaling each term d_k as follows:

$$p\left(d_k\gamma^{-1}\right) \approx \mathcal{N}(d_k\gamma^{-1} \mid (\mu_k - y_k^*)\gamma^{-1}, 1)$$
$$\gamma = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\sigma_k^2}.$$

It then follows that

$$p(d\gamma^{-2}) \approx NC\chi^2(d\gamma^{-2} \mid K, \lambda)$$
$$\lambda = \gamma^{-2}\sum_{k=1}^{K} (\mu_k - y_k^*)^2 \,.$$

Using that for random variable z and differentiable transformation g we have $p(z) = p(g(z)) \cdot |g'(z)|$ we get

$$p(d) \approx NC\chi^2(d\gamma^{-2} \mid K, \lambda)\gamma^{-2} \,.$$

Next, we show that the above approximation is unbiased. The mean of a noncentral χ^2 distribution is conveniently given by $K + \lambda$ so taking the expectation of d under the approximate distribution yields

$$\mathbb{E}[d] = \mathbb{E}[d\gamma^{-2}]\gamma^2 = (K + \lambda)\gamma^2.$$

Taking instead the expectation of d as the sum of squared differences we get

$$\begin{aligned}
\mathbb{E}[d] &= \sum_{k=1}^{K} \mathbb{E}\left[(y_k - y_k^*)^2\right] \\
&= \sum_{k=1}^{K} \mathbb{E}\left[\left(\frac{y_k - y_k^*}{\sigma_k}\right)^2\right]\sigma_k^2 \\
&= \sum_{k=1}^{K} \mathbb{E}\left[z_k^2\right]\sigma_k^2,
\end{aligned}$$

for $z_k \sim \mathcal{N}\left((y_k - y_k^*)\sigma^{-1}, 1\right)$. Since

$$z_k^2 \sim NC\chi^2(1, \lambda_z), \qquad \lambda_z = (y_k - y_k^*)^2\sigma_k^{-2},$$

we continue the above derivation by

$$\begin{aligned}
\cdots &= \sum_{k=1}^{K} \left(1 + (y_k - y_k^*)^2\sigma_k^{-2}\right)\sigma_k^2 \\
&= \sum_{k=1}^{K} \left(\sigma_k^2 + (y_k - y_k^*)^2\right) \\
&= \sum_{k=1}^{K} \sigma_k^2 + \sum_{k=1}^{K} (y_k - y_k^*)^2 \\
&= (K + \lambda)\gamma^2 \,.
\end{aligned}$$

Note that for $K = 1$ the above is no longer an approximation.

B Bayesian optimization with warped Gaussian processes

For the warped GP's we used a summed hyperbolic tangent transformation as originally proposed in [14] with 2 summation terms, i.e. $g(d) = d + \sum_{\ell=1}^{2} \tanh(a_\ell(b_\ell + d))$ thus introducing 4 new hyper-parameters. The EI utility for unseen input $\mathbf{x}_j$ derived by first sampling 10,000 points in the lower 0.999 confidence interval in the observed space.

Objective		**2N EI**	**2N LCB**	**GP EI**	**GP LCB**	**WGP EI**	**WGP LCB**
BNH	Mean	1.24e-04	**7.45e-05**	2.06e-04	2.55e-04	1.20e-02	5.26e-03
	Std	2.07e-04	5.71e-05	2.61e-04	2.65e-04	1.27e-02	4.58e-03
SRN	Mean	**2.88e-05**	4.32e-05	4.84e-04	2.37e-03	8.46e-03	2.74e-03
	Std	2.27e-05	4.61e-05	7.24e-04	3.09e-03	9.66e-03	3.41e-03
OSY	Mean	2.62e-03	**1.28e-03**	3.54e-02	2.20e-02	5.43e-02	2.64e-02
	Std	4.06e-03	1.48e-03	4.85e-02	1.65e-02	5.22e-02	2.23e-02
TwoBarTrussDesign	Mean	**2.75e-03**	1.32e-01	6.62e-03	1.44e-02	2.28e-02	7.35e-02
	Std	2.55e-03	1.82e-01	6.05e-03	2.90e-02	2.16e-02	8.33e-02
WeldedBeamDesign	Mean	1.68e-02	**1.51e-03**	1.87e-02	9.14e-03	1.19e-02	6.58e-03
	Std	1.56e-02	1.56e-03	2.89e-02	1.01e-02	1.41e-02	5.07e-03
Rosenbrock	Mean	8.71e-05	**3.49e-05**	3.06e-04	2.57e-04	7.73e-04	1.63e-03
	Std	7.84e-05	3.68e-05	6.00e-04	3.11e-04	8.45e-04	2.13e-03
Ackley	Mean	4.87e-04	6.90e-04	3.48e-03	**2.95e-04**	3.70e-03	1.42e-03
	Std	8.72e-04	9.74e-04	7.30e-03	5.28e-04	5.05e-03	1.45e-03
Bohachevsky	Mean	6.05e-05	**2.04e-05**	1.38e-03	3.64e-04	1.02e-02	1.99e-02
	Std	9.31e-05	2.26e-05	2.93e-03	7.69e-04	1.97e-02	4.35e-02
Griewank	Mean	**1.36e-04**	1.71e-04	8.45e-03	3.14e-03	1.25e-02	6.80e-03
	Std	1.83e-04	3.94e-04	1.41e-02	5.32e-03	1.47e-02	1.46e-02
H1	Mean	4.87e-02	2.19e-02	**1.05e-02**	5.53e-02	2.75e-02	3.81e-02
	Std	8.16e-02	2.82e-02	1.08e-02	1.13e-01	4.70e-02	5.60e-02
Himmelblau	Mean	3.67e-05	**1.08e-05**	1.59e-04	4.57e-04	9.91e-04	1.06e-03
	Std	5.31e-05	1.11e-05	3.14e-04	8.59e-04	1.76e-03	1.69e-03
Rastrigin	Mean	**3.96e-05**	2.79e-03	1.41e-03	2.55e-03	1.22e-02	2.61e-03
	Std	4.90e-05	7.89e-03	1.96e-03	3.45e-03	1.64e-02	3.28e-03
Schaffer	Mean	2.01e-04	**5.43e-05**	1.54e-03	3.50e-04	5.31e-03	8.60e-03
	Std	1.74e-04	8.93e-05	2.80e-03	3.53e-04	9.31e-03	1.51e-02
Schwefel	Mean	1.46e-04	**9.01e-06**	6.86e-04	2.53e-03	5.53e-03	1.82e-03
	Std	2.89e-04	1.17e-05	7.79e-04	3.50e-03	7.07e-03	2.84e-03

Table 2: Distance after 30 iterations for each of the optimization setups. We abbreviate the surrogate models as 2-norm: **2N** (proposed model), standard GP: **GP**, and warped GP: **WGP**. The two rows for each function list resp. mean and standard deviation for 8 repetitions with the same target vector. Lowest mean values are highlighted.

We then used the warped distribution to empirically calculate the expected improvement over the lowest objective evaluation found thus far. For LCB we simply minimized the bound in the latent space since g^{-1} is monotonically increasing meaning that for fixed q

$$F^{-1}(q;f(\mathbf{x})) < F^{-1}(q \mid f(\mathbf{x}')) \\ \Leftrightarrow F^{-1}(q; g(f(\mathbf{x}))) < F^{-1}(q \mid g(f(\mathbf{x}')))$$

The synthetic results including the warped GP is listed in Table 2.