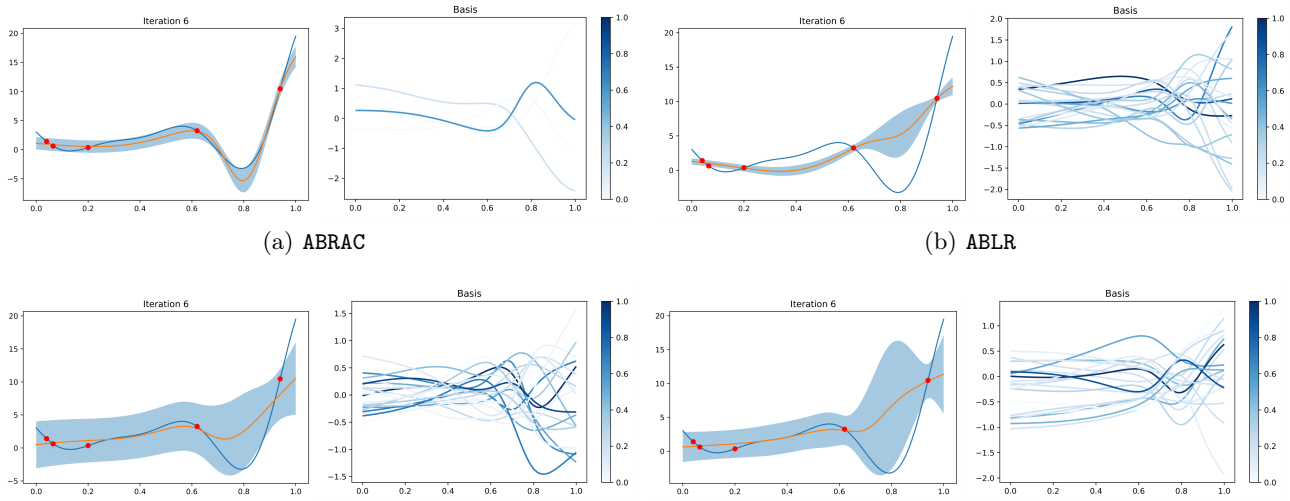

(a) `ABRAC` (b) `ABLR`

(c) `ABLR SGD fixed` (d) `ABLR L-BFGS fixed`

Figure 9: Comparison of the surrogate objective along with its basis functions after 5 total evaluations obtained by 4 different methods. In the first and the third panel the target function is shown in blue, the evaluations are the red points (which includes the same 5 random points for all methods), and the surrogate predictive mean is shown in orange with two standard deviations in transparent light blue. The second and fourth panel show the basis functions used to fit the surrogate model for given algorithm with colour representing its relative weight.

A Bayesian Linear Regression: extra material for Section 4.2

For the details omitted in Section 4.2, we continue with classical process of fitting Bayesian Linear regression (BLR) using Empirical Bayes, where one fits this probabilistic model by firstly integrating out weights $\mathbf{w}_{new}$, which leads to the multivariate Gaussian model Normal$(\mu(x_{new}), \sigma^2(x_{new}))$ for the next input x_{new} of the new task, whose mean and variance can be computed analytically (Bishop, 2006)

$$\begin{aligned}\mu(x_{new}) &= \beta_{new} f_{new}^\top K_{new}^{-1} (\mathbf{\Phi}_{new,b\downarrow})^\top \mathbf{y}_{new},\\ \sigma^2(x_{new}) &= f_{new}^\top K_{new}^{-1} f_{new} + \frac{1}{\beta_{new}},\end{aligned}$$

where $f_{new} = \phi_{\mathbf{z},b\downarrow}(x_{new})$ and $K_{new} = \beta_{new}(\mathbf{\Phi}_{new,b\downarrow})^\top \mathbf{\Phi}_{new,b\downarrow} + \text{Diag}(\alpha_{new})$. We use this as the input to the acquisition function, which decides the next point to sample. Before this step, we first need to obtain the parameters. This model defines a Gaussian process, for which we can compute the covariance matrix in the closed-form solution.

$$\begin{aligned}\text{cov}(y_i, y_j) &= \mathbb{E}\left[(\phi_{\mathbf{z},b\downarrow}(x_i)^\top w + \epsilon_i)(\phi_{\mathbf{z},b\downarrow}(x_j)^\top w + \epsilon_j)^\top\right]\\ &= \phi_{\mathbf{z},b\downarrow}(x_i)^\top \mathbb{E}\left[ww^\top\right] \phi_{\mathbf{z},b\downarrow}(x_j) + \mathbb{E}\left[\epsilon_i \epsilon_j\right]\\ &= \phi_{\mathbf{z},b\downarrow}(x_i)^\top \text{Diag}(\alpha)^{-1} \phi_{\mathbf{z},b\downarrow}(x_j) + \mathbb{E}\left[\epsilon_i \epsilon_j\right],\end{aligned}$$

thus the covariance matrix has the following form $\Sigma_{new} = \mathbf{\Phi}_{new,b\downarrow} \text{Diag}(\alpha_{new})^{-1} (\mathbf{\Phi}_{new,b\downarrow})^\top + \beta_{new}^{-1} I_{N_{new}}$. We obtain $\{\alpha_{new,i}\}_{i=1}^r$ and β_{new} by minimizing the negative log likelihood of our model has, up to constant factors in the following form

$$\min_{\{\alpha_{new,i}\}_{i=1}^r, \beta_{new}} \frac{1}{2} \log|\Sigma_{new}| + \mathbf{y}_{new}^\top \Sigma_{new}^{-1} \mathbf{y}_{new},$$

for which we use L-BGFS algorithm as this method is parameter-free and does not require to tune step size, which is a desired property as this procedure is run in every step of BO. In order to speed up optimization and obtain linear scaling in terms of evaluation, one needs to include extra modifications, starting with decomposition of the matrix $(\mathbf{\Phi}_{new,b\downarrow})^\top (\mathbf{\Phi}_{new,b\downarrow}) + {}^1/_{\beta_{new}} \text{Diag}(\alpha_{new})$ as $L_{new} L_{new}^\top$ using Cholesky. This decomposition exists

since $(\mathbf{\Phi}_{new,b\downarrow})^\top(\mathbf{\Phi}_{new,b\downarrow}) + {}^1\!/_{\beta_{new}}\,\mathrm{Diag}(\alpha_{new})$ is always positive definite, due to semi-positive definiteness of $(\mathbf{\Phi}_{new,b\downarrow})^\top(\mathbf{\Phi}_{new,b\downarrow})$ and positive diagonal matrix ${}^1\!/_{\beta_{new}}\,\mathrm{Diag}(\alpha_{new})$. Final step is to use Weinstein–Aronszajn identity, and Woodbury matrix inversion identity, which implies that our objective can be rewritten to the the equivalent form

$$\min_{\{\alpha_{new,i}\}_{i=1}^r,\beta_{new}} -\frac{(N_{new}-r)}{2}\log(\beta_{new}) - \frac{1}{2}\sum_{i=1}^{r}\log(\alpha_{new,i})$$
$$+\sum_{i=1}^{r}\log([L]_{ii}) + \frac{\beta_{new}}{2}(\|\mathbf{y}_{new}\|^2 - \|L_{new}^{-1}(\mathbf{\Phi}_{new,b\downarrow})^\top\mathbf{y}_{new}\|^2),$$

which leads to overall complexity $\mathcal{O}(d^2\max\{N_{new},d\})$, which is linear in number of evaluations of the function f_{new}.