Simulation setup:

- Data generating processes:

$$y_i = g(x_i) + e_i,$$
$$g(x_i) = a \sin\left(b 2\pi x_i + \frac{\pi}{4}\right).$$

- $x_i \sim U[0,1]$ and $e_i \sim N(0, \sigma_i^2)$, where $\sigma_i = 1$ for the homoskedastic simulation and $\sigma_i = \sqrt{5}x_i^2$ for the heteroskedastic simulation.

- The parameter a is selected to control the population $R^2 = a^2/(2+a^2)$. We set $R^2 = 0.25, 0.5, 0.75, 0.9$. Also, b=1 in all simulations.

- Sample sizes: $n = 50, 75, 100, 125, 150, 200, 300, 400, 600, 800, 1000$.

- Estimation: LS estimator using a quadratic splines.

$$\hat{g}_r(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \sum_{j=1}^{r} \hat{\beta}_{2+j}(x - t_j)^2 1(x > t_j).$$

- Splines of order: $0, 1, 2, ..., K$ where $K = 4n^{0.15}$. That is, $K = 7, 8, 8, 8, 8, 9, 9, 10, 10, 11, 11$ for 11 sample sizes.

- We compute the IMSE $= \mathrm{E} \int_0^1 (\hat{g}(x) - g(x))^2 \, dx$. We approximate the IMSE by using a 1001 point grid on [0,1] and simulation using 10,000 draws. We normalize the IMSE by dividing by the IMSE of the LS averaging estimator with the infeasible optimal weights.