C ADDITIONAL EXPERIMENTAL ANALYSIS

C.1 Analysis of annealing scheme and hyper-parameter choices

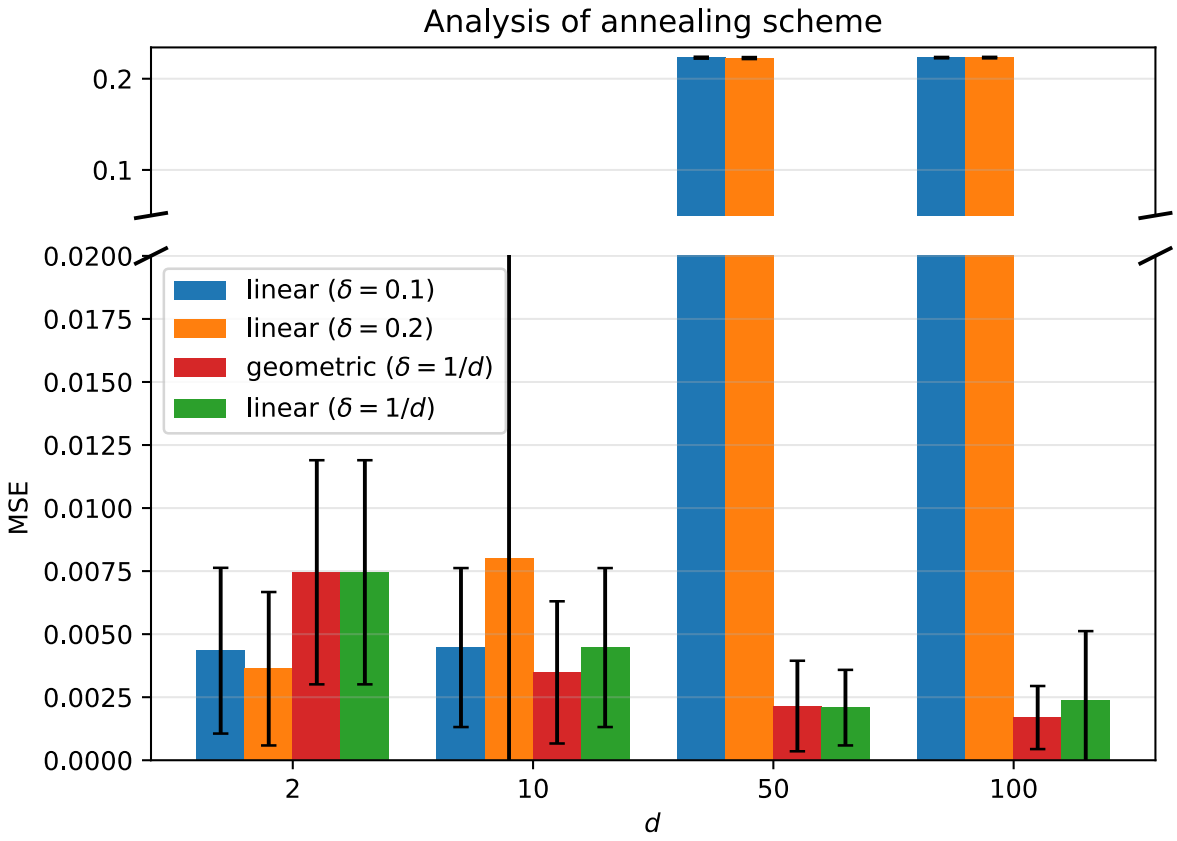

Figure 4: Analysis of the mean absolute error (MAE) for different annealing schemes and different values of δ. Lower MAE values are better.

In all the following we analyze FKL + PathG + anneal with different settings. In all plots of this section the very right (green bar) corresponds to the setting that is used in the main paper.

Figure 4 shows the comparison of different annealing schemes. "linear" denotes the linear annealing scheme of Algorithm 1, where δ is set to either a fixed value for all experiments, or set to $1/d$. As described in the main paper, for "linear $\delta = 1/d$", we set $t_{int} := 1000$. For the other values ($\delta = 0.1$ and $\delta = 0.2$), we increase t_{int} so that the total number of iterations (i.e. the computational budget) stays the same. "geometric" denotes the following geometric annealing scheme:

$$\beta_t = \delta^{1-(\lfloor t/t_{int} \rfloor/m)},$$

where β_t denotes the value of β at iteration t, and m denotes the total number of times β is increased. We set $m := 1/\delta$ to match the computational budget with "linear $\delta = 1/d$". The results show that for higher dimensions ($d \geq 50$) a lower value of δ becomes important.

Additionally, Figure 5 shows the median SNR (of each dimension with interquartile ranges) at the beginning of training.[7] From $d = 2$ to $d = 50$ all methods' SNR drop, but for $d = 100$, the choice of $\delta = 1/d$ leads to an increase in SNR.

[7] β_1 is the same for "linear $\delta = 1/d$" and "geometric $\delta = 1/d$", so SNR at the initial phase are the same.

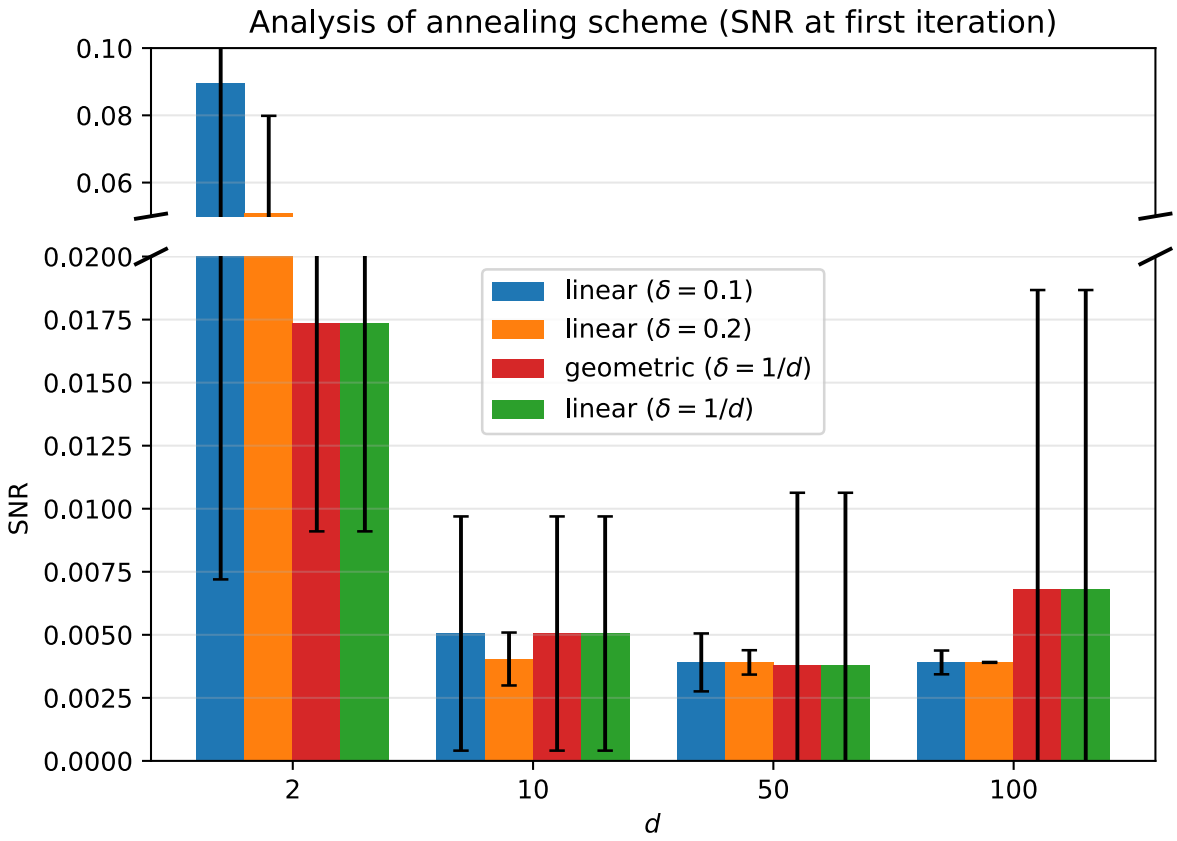

Figure 5: Analysis of the signal to noise ratio (SNR) at the first iteration for different annealing schemes and different values of δ. Higher SNR values are better.

Figure 6 shows the impact of different choices for the initial variance σ_0^2. We see that setting $\sigma_0^2 \leq 100$ provides stable results for all d.

Figure 7 shows the impact of different duration times t_{int}. Note that, different from the other experiments, the total computational budget for each value of t_{int} changes, in particular, the runtime of $t_{init} = 100$ and $t_{init} = 500$ is one tenth and one half of that of $t_{init} = 1000$, respectively. We see that reducing t_{int} from 1000 to 500 leads only to a marginal decrease in MSE.

C.2 Analysis of SNR at beginning and end of training

In Figure 8, we compare the SNR at the beginning and end of training of the proposed method and methods with or without path gradients/annealing. We find that both path gradients and annealing are effective to stabilize training (i.e. higher SNR compared to the other methods).

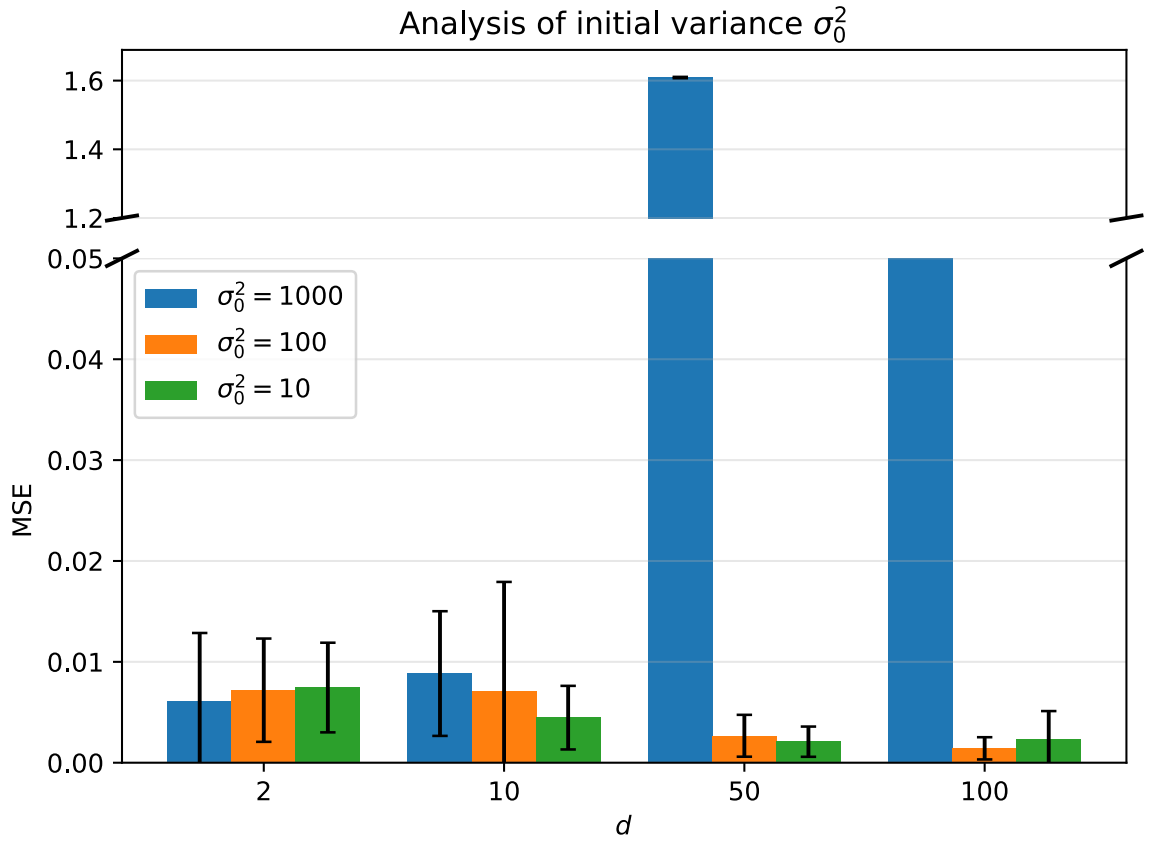

Figure 6: Analysis of the mean absolute error (MAE) for different initial variance values. Lower MAE values are better.

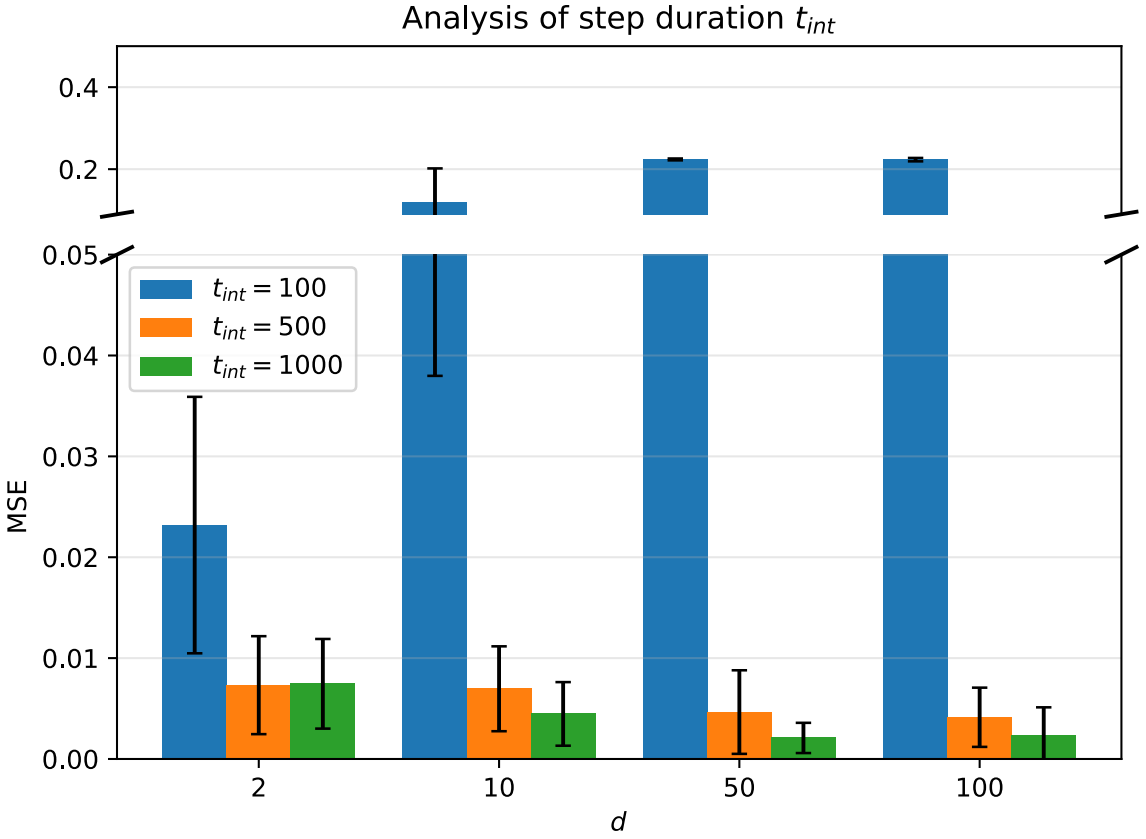

Figure 7: Analysis of the mean absolute error (MAE) for different duration times t_{int}. Lower MAE values are better.

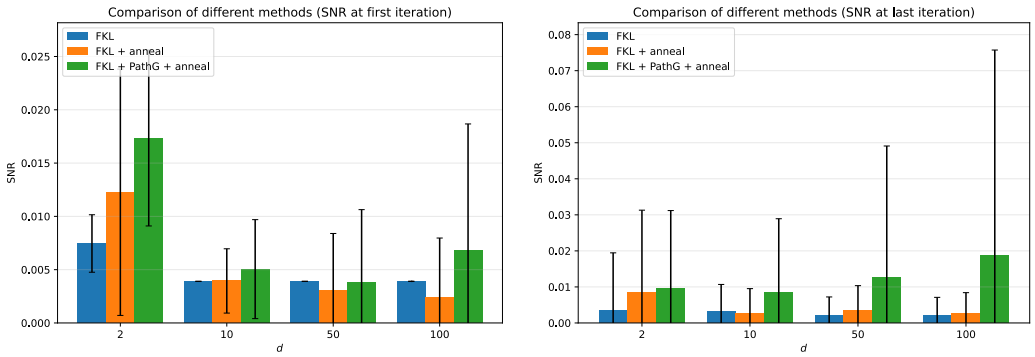

Figure 8: Analysis of the signal to noise ratio (SNR) at the first iteration (left) and last iteration (right) for different methods. Higher SNR values are better.