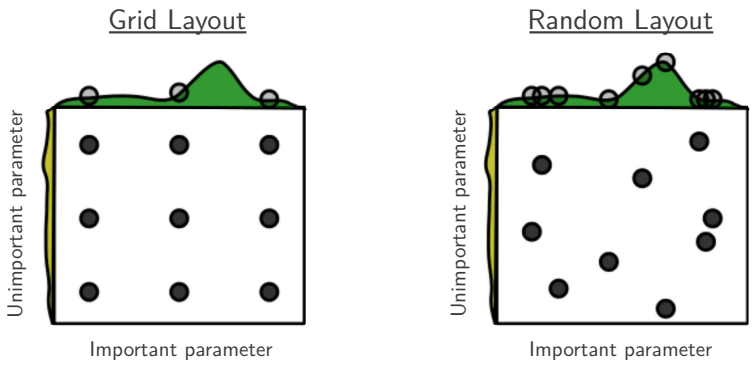

Figure 1: Grid and random search of nine trials for optimizing a function $f(x,y) = g(x) + h(y) \approx g(x)$ with low effective dimensionality. Above each square $g(x)$ is shown in green, and left of each square $h(y)$ is shown in yellow. With grid search, nine trials only test $g(x)$ in three distinct places. With random search, all nine trials explore distinct values of g. This failure of grid search is the rule rather than the exception in high dimensional hyper-parameter optimization.

given learning algorithm, looking at several relatively similar data sets (from different distributions) reveals that on different data sets, different subspaces are important, and to different degrees. A grid with sufficient granularity to optimizing hyper-parameters for all data sets must consequently be inefficient for each individual data set because of the curse of dimensionality: the number of wasted grid search trials is exponential in the number of search dimensions that turn out to be irrelevant for a particular data set. In contrast, random search thrives on low effective dimensionality. Random search has the same efficiency in the relevant subspace as if it had been used to search only the relevant dimensions.

This paper is organized as follows. Section 2 looks at the efficiency of random search in practice vs. grid search as a method for optimizing neural network hyper-parameters. We take the grid search experiments of Larochelle et al. (2007) as a point of comparison, and repeat similar experiments using random search. Section 3 uses Gaussian process regression (GPR) to analyze the results of the neural network trials. The GPR lets us characterize what Ψ looks like for various data sets, and establish an empirical link between the low effective dimensionality of Ψ and the efficiency of random search. Section 4 compares random search and grid search with more sophisticated point sets developed for Quasi Monte-Carlo numerical integration, and argues that in the regime of interest for hyper-parameter selection grid search is inappropriate and more sophisticated methods bring little advantage over random search. Section 5 compares random search with the expert-guided manual sequential optimization employed in Larochelle et al. (2007) to optimize Deep Belief Networks. Section 6 comments on the role of global optimization algorithms in future work. We conclude in Section 7 that random search is generally superior to grid search for optimizing hyper-parameters.