Task-Agnostic Sample Design for Machine Learning

Bhavya Kailkhura and **Jayaraman J. Thiagarajan** and **Qunwei Li** and **Jize Zhang** and **Yi Zhou** and **Peer-Timo Bremer**

Abstract

In this paper, we present a framework to understand the effect of sampling properties of training data on the generalization gap of machine learning (ML) algorithms. In particular, we express generalization gap in terms of the power spectra of the sample design and that of the function to be learned. Using this framework, we show that space-filling sample designs, such as blue noise and Poisson disk sampling, which optimize spectral properties, outperform random designs in terms of generalization gap. Our analysis also sheds light on design principles for constructing optimal task-agnostic sample designs that minimize the generalization gap. We corroborate our findings using regression experiments with neural networks on: a) synthetic functions, and b) a complex scientific simulator for inertial confinement fusion (ICF).

Keywords: Design of experiments, generalization gap, scientific machine learning

1. Introduction

Machine learning (ML) techniques have led to incredible advances in a wide variety of commercial applications, and similar approaches are rapidly being adopted in several scientific and engineering problems. Traditionally, ML research has focused on developing modeling techniques and training algorithms to learn generalizable models from historic labeled data. However, in several applications, we encounter a key challenge even before building the model – determining the input samples for which the responses should be collected (referred to as *task-agnostic sample design* problem). This is particularly true for emerging applications in physical sciences and engineering where curated datasets are not available *a priori*. For example, in inertial confinement fusion (ICF) (Anirudh et al., 2020), one needs to build a high-fidelity mapping from the process inputs, say target and laser settings, to process outputs, such as ICF implosion neutron yield and X-ray diagnostics. In such scenarios, the properties of collected data directly control the generalization error of ML models. However, determining the right samples to use for learning hinges on understanding the intricate interplay between sampling properties and the ML generalization error. Unfortunately, our theoretical understanding is very limited in this regard, and hence existing sample design approaches rely upon a variety of heuristics, e.g., generating so called space-filling sample designs (Joseph, 2016) to cover the input space as uniformly as possible.

Most existing theoretical frameworks (Boucheron et al., 2005; Bousquet and Elisseeff, 2002) only study the generalization properties of random i.i.d. designs. Intuitively, this assumption ignores the dependency of generalization gap on data properties except the sample size (data-independent bounds). While some efforts exist to obtain data-dependent bounds (Koltchinskii and Panchenko, 2000; Herbrich and Williamson, 2002; Xu and Mannor, 2012), they still focus on studying model design related questions while ignoring sample design

aspects. To the best of our knowledge, there does not exist a framework in the literature that can help study the generalization error of generic sample designs (e.g., space-filling). This paper proposes to study generalization error from the viewpoint of the sampler generating the training data. We fill a crucial gap by developing a framework[1] capable of characterizing the generalization performance of generic sample designs based on metrics expressive enough to quantify a broad range of sample distributions.

2. Risk Minimization using Monte Carlo Estimates

We consider the following general supervised learning setup: We have two spaces of objects $X \in \mathbb{T}^d$ (toroidal unit cube $[0,1]^d$) and $Y \in \mathbb{R}$ where $Y = \mathcal{F}(X)$. The goal of a learning algorithm is to learn a function $h : X \to Y$ (often called *hypothesis*) which approximates the true (but unknown) function $\mathcal{F}$. We assume access to training data comprised of N samples $S = \{(x_1, y_1), \cdots, (x_N, y_N)\}$ drawn from an unknown distribution $P(x, y)$ and denote loss function by $l(\cdot, \cdot)$. We infer a hypothesis $h(\cdot)$ by minimizing the population risk:

$$R_P(h) \triangleq \mathbb{E}_{P(x,y)}[l(h(x), y)] = \int l(h(x), y) dP(x, y). \tag{1}$$

Empirical Risk Minimization: In general, the joint distribution $P(x, y)$ is unknown and hence the risk $R_P(h)$ cannot be computed. Hence, an approximation referred as *empirical risk* is often used which is obtained by averaging the loss function on the training data:

$$R_S(h) \triangleq 1/N \int_{\mathbb{T}} S(x) l(h(x), y) dx, \tag{2}$$

where $\mathbb{T}$ is the sampling domain and $S(x)$ is a sample design rewritten as a random signal S composed of N Dirac functions at positions $S(x) = \sum \delta(x - x_i)$ for $i = 1, \cdots, N$.

Generalization Gap: In ML, the performance of a supervised learning algorithm is measured by the generalization gap, which is the expected difference between the population risk and the empirical risk:

$$\text{gen}(h) \triangleq \mathbb{E}_S[(R_P(h) - R_S(h))^2] = bias^2 + var(R_S(h)), \tag{3}$$

which is the expected difference between the population risk and its empirical risk on the training data for a fixed hypothesis $h(\cdot)$.

3. Connecting Generalization Gap with Sample Design

3.1 Monte Carlo Estimator of Risk in the Spectral domain

The MC estimator for risk as given in Eq. 2 can be transformed to the Fourier domain ϕ using the fact that the dot-product of functions (the integral of the product) is equivalent to the dot-product of their Fourier coefficients (Pilleboue et al., 2015). This allows us to pose the MC estimator for empirical risk as follows:

$$R_S(h) \triangleq 1/N \int_{\phi} \mathcal{F}_S(\mathbf{k}) \mathcal{F}_l(\mathbf{k})^* d\mathbf{k}, \tag{4}$$

1. A longer version of this paper is available online (Kailkhura et al., 2019).

where $\mathcal{F}_S$, $\mathcal{F}_l$ denote the Fourier transforms of the sampling function S and the loss function l and $(*)$ denotes complex conjugate.

3.2 Spectral Analysis of the Generalization Gap

We now use the spectral domain version of empirical risk to define generalization gap:

$$\text{gen}(h) = \left(\mathbb{E}(R_S(h)) - R_P(h)\right)^2 + \frac{1}{N^2}\int_{\phi\times\phi} \mathbb{E}(\mathcal{F}_{S,l}(\mathbf{k},\mathbf{k}'))d\mathbf{k}d\mathbf{k}' - \left(\mathbb{E}(R_S(h))\right)^2, \quad (5)$$

where $\mathcal{F}_{S,l}(\mathbf{k},\mathbf{k}') \triangleq \mathcal{F}_S(\mathbf{k}) \cdot \mathcal{F}_l(\mathbf{k})^* \cdot \mathcal{F}_S(\mathbf{k}')^* \cdot \mathcal{F}_l(\mathbf{k}')$. Using this definition, we derive an explicit closed-form relation of the generalization gap with the power spectra of both S and l. We simplify Eq. 5 by restricting our analysis to homogeneous (or unbiased) designs.

Theorem 1: *The generalization gap for homogeneous sample designs in terms of the power spectra of both the sampling pattern $\mathcal{P}_S$ and the loss function $\mathcal{P}_l$ can be obtained as:*

$$gen(h) \triangleq \frac{1}{N}\int_{\Theta} \mathbb{E}(\mathcal{P}_S(\mathbf{k}))\mathcal{P}_l(\mathbf{k})d\mathbf{k}, \quad (6)$$

where Θ is the Fourier domain ϕ without DC frequency.

When the design is isotropic (i.e., the power spectrum is radially symmetric), the generalization gap can be directly computed from the radial mean power spectra of the loss with $\rho = |\mathbf{k}|$, i.e., $\hat{\mathcal{P}}_l$ and the sample design $\hat{\mathcal{P}}_S$.

Proposition 1: *The generalization gap for isotropic homogeneous sample designs is*

$$gen(h) \triangleq \frac{\mu(\mathcal{S}^{d-1})}{N}\int_0^{\infty} \rho^{d-1}\mathbb{E}(\hat{\mathcal{P}}_S(\rho))\hat{\mathcal{P}}_l(\rho)d\rho, \quad (7)$$

where $\mu(\mathcal{S}^d)$ is the Lebesgue measure of a d-dimensional unit sphere in $\mathbb{R}^d$ given by $2\sqrt{\pi^d}/\Gamma(d/2)$.

It can be seem from these theoretical results that the shape of the spectrum of sample design plays an important role. An ideal sample design power spectrum should have a large *zero-region*, i.e., low frequency band with zero power to minimize the generalization gap. These results also indicate that sample designs with optimized spectral properties (e.g., space-filling sample designs (Kailkhura et al., 2018)) will result in models with superior generalization. Next, we corroborate this conclusion via experiments.

4. Experiments

In this section, we corroborate our theoretical findings via regression experiments.

4.1 Experimental Setup

In our experiments, we consider three state-of-the-art sample design, namely random, blue noise and Poisson disk sample (PDS) designs. To generate space-filling sample designs (i.e., blue noise and PDS), we use gradient descent based approach as proposed in (Muniraju et al., 2018). Figure 1 illustrates the sample design along with their spectral properties for $d = 2$ and $N = 1000$. Note that blue noise and Poisson disk designs have optimized spectral properties, i.e., a large *zero-region*. We vary training sample set size from 200 to 1000. For

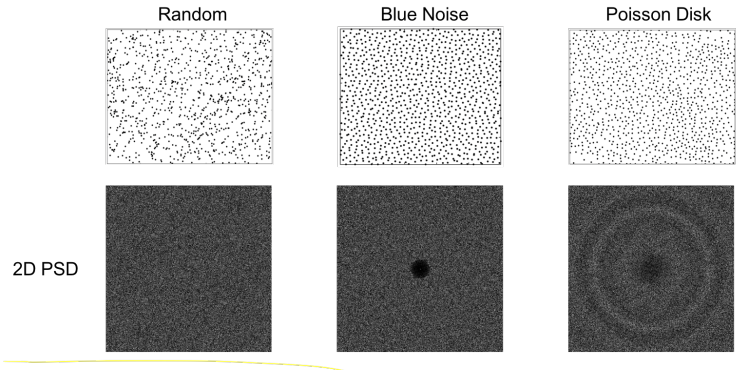
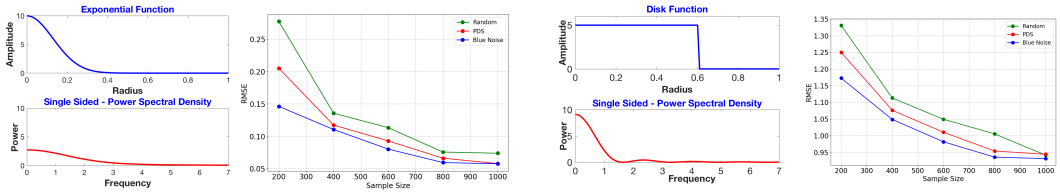


Figure 1: Sample design along with their spatial/spectral properties.



(a) (b)

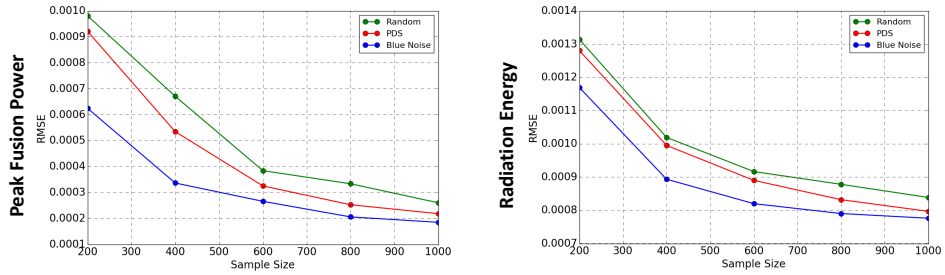
Figure 2: Generalization comparison on synthetic functions: (a) exponential, (b) disk.

both the experiments, we use neural network with two hidden layers, with 200 and 100 nodes respectively, each followed by a LeakyReLu activation function. For training algorithm, we use ADAM optimizer with learning rate and batch size to be 0.01 and 64, respectively. We evaluate the generalization performance of neural networks learnt using different sample designs based on root mean square error (RMSE) on 10^3 unseen regular grid based test samples. All the results are averaged over 20 independent realizations.

4.2 Results

4.2.1 Synthetic Functions

In this experiment, we consider regression problem of learning analytical functions and perform a comparative study of different sample designs, in terms of their generalization performance. We consider two synthetic functions with known but different spectral behavior: a) disk function: $y = 5$ if $|x| < 6$ ($y = 0$ otherwise), and b) exponential function: $y = 10 * exp(-30 * (|x|^2))$ where $x \in [0, 1]^3$. We use experimental setup as described in Section 4.1. In Figure 2, we show radial average of both the functions and their power spectral densities. For both of the functions, we see that models trained on blue noise and PDS sample designs generalize significantly better for all sampling budgets as compared to models trained on random sample design. Furthermore, the gain is significantly higher in low-sampling regime which makes spectral designs an attractive solution in small-data ML applications.



(a) (b)

Figure 3: Generalization comparison on ICF: (a) peak fusion power, (b) radiation energy.

4.2.2 Inertial Confinement Fusion (ICF) Simulator

Next, we consider a scientific machine learning problem of learning a regression model for an inertial confinement fusion (ICF) simulator developed at the National Ignition Facility (NIF). The NIF is aimed at demonstrating inertial confinement fusion (ICF), that is, thermonuclear ignition and energy gain in a laboratory setting. We use the NIF JAG simulator[2] with different input parameters, such as, laser power, pulse shape etc. For each simulation run, several output quantities, such as peak fusion power, yield, etc., are obtained. In this experiment, we vary three input parameters and study the problem of learning a model to regress peak fusion power and radiation energy. We use experimental setup as described in Section 4.1. Note that, the function and its spectral behavior is not known in this experiment and its may not comply with any of our assumptions. In Figure 3, we observe that regression error patterns are consistent with our observations in the previous experiment. Blue noise design performs the best followed by the PDS design. This shows that our finding that spectral designs are superior compared to random designs hold even in this real-world setting.

The performance gain with both synthetic function and ICF simulator can be credited to superior spectral properties of blue noise and PDS designs compared to random designs. These observations corroborate our theoretical results which show that the shape of the power spectra has a major impact on generalization gap and sampling designs with optimized spectral properties (i.e., blue noise and PDS) are superior to random designs.

5. Conclusions

We presented a framework to study effect of task-agnostics sample designs on the generalization gap of ML models. We showed that the generalization gap is related to the power spectra of a sample design and the function of interest. We provided design guidelines towards constructing optimal sample designs for a given problem. There are still many interesting questions that remain to be explored such as an analysis of the generalization gap for cases where input domain is a non-linear manifold. Other directions such as designing higher quality sample designs than currently possible and importance sampling can also be pursued.

2. `https://github.com/rushilanirudh/macc`

Acknowledgments

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

References

Rushil Anirudh, Jayaraman J Thiagarajan, Peer-Timo Bremer, and Brian K Spears. Improved surrogates in inertial confinement fusion with manifold and cycle consistencies. *Proceedings of the National Academy of Sciences*, 117(18):9741–9746, 2020.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

Ralf Herbrich and Robert C Williamson. Algorithmic luckiness. *Journal of Machine Learning Research*, 3(Sep):175–212, 2002.

V Roshan Joseph. Space-filling designs for computer experiments: A review. *Quality Engineering*, 28(1):28–35, 2016.

Bhavya Kailkhura, Jayaraman J Thiagarajan, Charvi Rastogi, Pramod K Varshney, and Peer-Timo Bremer. A spectral approach for the design of experiments: Design, analysis and algorithms. *The Journal of Machine Learning Research*, 19(1):1214–1259, 2018.

Bhavya Kailkhura, Jayaraman J Thiagarajan, Qunwei Li, and Peer-Timo Bremer. A look at the effect of sample design on generalization through the lens of spectral analysis. *arXiv preprint arXiv:1906.02732*, 2019.

Vladimir Koltchinskii and Dmitriy Panchenko. Rademacher processes and bounding the risk of function learning. In *High dimensional probability II*, pages 443–457. Springer, 2000.

Gowtham Muniraju, Bhavya Kailkhura, Jayaraman J Thiagarajan, Peer-Timo Bremer, Cihan Tepedelenlioglu, and Andreas Spanias. Coverage-based designs improve sample mining and hyper-parameter optimization. *arXiv preprint arXiv:1809.01712*, 2018.

Adrien Pilleboue, Gurprit Singh, David Coeurjolly, Michael Kazhdan, and Victor Ostromoukhov. Variance analysis for monte carlo integration. *ACM Transactions on Graphics (TOG)*, 34(4):124, 2015.

Huan Xu and Shie Mannor. Robustness and generalization. *Machine learning*, 86(3):391–423, 2012.