Sequential design of point-wise measurements based on a deep generative model

Hyungil Moon[1†*] **Dominic T. Lennon**[1*] **Leon C. Camenzind**[2]
Liuqi Yu[2] **Dominik M. Zumbühl**[2] **G. Andrew D. Briggs**[1]
Michael. A. Osborne[3] **Edward A. Laird**[4] **Natalia Ares**[1]
[1] Department of Materials, University of Oxford, UK
[2] Department of Physics, University of Basel, Switzerland
[3] Department of Engineering, University of Oxford, UK
[4] Department of Physics, Lancaster University, UK
[†]hyungil.moon@materials.ox.ac.uk

Abstract

We present a new method to perform 2D measurements based on planning point-wise measurements. This can replace the raster scan for many scientific experiments. The proposed method is based on a deep generative model to generate multiple possible reconstructions given an arbitrary number of scattered measurements, and information theory is used to choose the optimal next location for the point-wise measurement. The proposed method is not domain-specific and could be used for any type of 2D measurement as long as training data is available.

1 Introduction

In this paper, we are interested in scientific experiments, where only point-wise measurement is available at a time, such as measurements of electric current through a solid-state device. It is often required to gather a 2D measurement map to extract useful information for scientific analysis. The most conventional way of measuring a 2D map using point-wise measurements is a raster scan. This paper proposes an alternative way of completing a 2D map in a more efficient way than the raster scan. The overall framework of the proposed method is i) predicting multiple full resolution measurements, called reconstructions in this paper, based on already collected data and ii) determining where the most informative location is for a new measurement. Especially, we focus on problems, for which there is a set of expected patterns that is used as a training set. In our experiment, we measure electric current through a quantum dot device over a certain range of bias and gate voltages, where we expect a pattern called Coulomb diamonds. It is important to measure the regions where signal changes like at the edges of the diamonds, because the change of signal contains most of the information. We show that this can be achieved by reducing the uncertainty of reconstructions. As we have many actual measurements and simulation examples, a deep generative model can be trained. It is important to generate multiple possible full-resolution reconstructions as the multiple possibility is the source of the optimal decision to reduce the uncertainty.

The proposed method is closely related with active learning [1], and design of experiments [2], of which objective is to reduce the uncertainty of a predictor. The most common approach on sequential design of experiments for global optimisation or uncertainty reduction is based on Gaussian process [3, 4, 5, 6, 7, 8]. Given existing observations, a Gaussian process model generates a posterior distribution for any unseen location, and the next observation location is chosen by optimising a

[*] Equally contributed. Hyungil Moon developed the algorithm, and Dominic Lennon developed a network interface and setup the device and experiments.

Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019), Vancouver, Canada.

selection criterion. The biggest difference of the proposed method is replacing a Gaussian process regression model with a deep generative model and approximating the selection criterion with random samples of the model. When a training set is available, a deep generative model has several advantages. Firstly, the prediction of a deep generative model is not necessarily Gaussian, or uni-modal. Secondly, it is faster to generate a random sample of full-resolution reconstruction. Note that the aim of this paper is to introduce a recent paper [9], emphasizing the contribution and novelty in terms of machine learning.

2 Mathematical objective

2D scan based on a point-wise measurement is a popular method in scientific measurement to characterise underlying physical quantities. In our experiments on quantum-dot devices, a measurement output is electric current in amperes given bias and gate voltages. A 2D current map across the bias and gate voltages, called a stability diagram, contains physical information about a quantum dot [10]. The device of interest in this paper is a single-dot device, hence we expect to see a pattern called Coulomb diamonds, which is a sequence of aligned diamonds as in Figure 1. The stability diagrams basically consist of edges and flat regions, and the edges contains almost all physical information. Therefore, it is beneficial to scan edges first than flat regions.

This paper focuses on planning a measurement sequence to estimate the entire 2D map with least measurements, which is essentially the same objective with active learning and design of experiments for uncertainty reduction. This objective is not only more general, but also very closely related with the problem of measuring edges first, because flat regions are easily predictable, and most uncertainty is on the edges.

Let Y be a set containing all pixel values in full-resolution, and Y_n be a set of n pairs of location x_j and point measurement y_j: $Y_n = \{(x_j, y_j) \mid j = 1 \sim n\}$. The likelihood is defined as

$$p(Y_n \mid Y) \propto \exp\bigl(-\lambda \Sigma_{(x,y)\in Y_n} |y - Y(x)|\bigr) , \tag{1}$$

where $Y(x)$ is the pixel value of Y at x, and λ is a sensitivity parameter. The objective of the algorithm is to choose the next measurement location x_{n+1} to minimise the uncertainty of Y or equivalently the entropy $H(Y|Y_n, y_{n+1})$. Throughout this paper, we simply write y_{n+1} instead of (x_{n+1}, y_{n+1}) for the condition of random variables and distributions for brevity.

The dimensionality of Y is the number of pixels, but the effective dimensionality is much smaller, because most combinations of pixel values are extremely unlikely. Therefore, we use an embedding vector $\mathbf{z}$, which is trained by a Variational Auto-Encoder (VAE), so that $\mathbf{z}$ can be decoded to $Y(\mathbf{z})$, and Y can be encoded to $\mathbf{z}(Y)$. The objective in the embedding space is to minimise $H(\mathbf{z}|Y_n, y_{n+1})$ by choosing the location of x_{n+1}. It is well kown that minimising the entropy is equivalent to

- maximising the mutual information: $\mathrm{argmax}_{x_{n+1}} I(\mathbf{z}|Y_n; y_{n+1}|Y_n, x_{n+1})$.
- maximising the expected KL divergence: $\mathrm{argmax}_{x_{n+1}} \mathbb{E}_{y_{n+1}} \Bigl[\mathrm{KL}\bigl(p_n(\mathbf{z} \mid y_{n+1}) \| p_n(\mathbf{z})\bigr)\Bigr]$, where $p_n(\cdot) = p(\cdot|Y_n)$.

Detailed derivations can be found in [9].

3 Approximation

The expected KL divergence requires two integrals:

$$\mathbb{E}_{y_{n+1}} \Bigl[\mathrm{KL}\bigl(p_n(\mathbf{z} \mid y_{n+1}) \| p_n(\mathbf{z})\bigr)\Bigr] = \int_{y_{n+1}} p_n(y_{n+1}) \int_{\mathbf{z}} p_n(\mathbf{z} \mid y_{n+1}) \log \frac{p_n(\mathbf{z} \mid y_{n+1})}{p_n(\mathbf{z})} d\mathbf{z} dy_{n+1}. \tag{2}$$

Since this integral has no closed form expression, we propose an approximation using importance sampling [11]. Let $n_s < n$ denote the number of measurements that are used for sampling reconstructions $\hat{\mathbf{z}}_1, \ldots, \hat{\mathbf{z}}_M$, and define weights and normalised weights as $w_i = p_n(\hat{\mathbf{z}}_i)/p_{n_s}(\hat{\mathbf{z}}_i)$, $w_i' = p_n(\hat{\mathbf{z}}_i|y_{n+1})/p_{n_s}(\hat{\mathbf{z}}_i)$, $\tilde{w}_i = w_i / \sum_{j=1}^M w_j$, and $\tilde{w}_i' = w_i / \sum_{i=j}^M w_j'$. The normalised weights

Algorithm 1: Algorithm for the efficient measurement

```
Input: Batch schedule $\mathcal{N}_b$, the number of samples for the approximation $M$
Output: The measurement $Y_n$
Measure 8×8 initial scan $Y_{64}$;
$n \leftarrow 64$ ;
repeat
    Generate samples $\hat{z}_1, \dots, \hat{z}_M$ using a MCMC sampler ;
    Calculate the information gain map using the approximated criterion (3) ;
    $n_{\mathrm{next}} \leftarrow \min\{n' \in \mathcal{N}_{\mathrm{batch}} | n' > n\}$ ;
    $n_{\mathrm{batch}} \leftarrow n_{\mathrm{next}} - n$;
    The next batch $\leftarrow$ a set of $n_{\mathrm{batch}}$ best locations in the information gain map ;
    Measure the batch ;
    $n \leftarrow n_{\mathrm{next}}$
until Stopping criterion satisfied;
```

can be interpreted as probabilities, $P_n(i) = \tilde{w}_i$ and $P_{n+1}(i) = \tilde{w}'_i$, because they are non-negative and sum to one. The inner integral in (2) can be approximated with the probabilities:

$$\mathbb{E}_{y_{n+1}}\Big[\mathrm{KL}\big(p_n(\mathbf{z} \mid y_{n+1}) \| p_n(\mathbf{z})\big)\Big] \approx \int_{y_{n+1}} p_n(y_{n+1}) \mathrm{KL}(P_{n+1} \| P_n) dy_{n+1} + C,$$

where C is a constant. The posterior sample from $p_n(y_{n+1})$ can be drawn by i) generating samples $\hat{\mathbf{z}}_i$ ($i = 1 \sim M$) from $p_n(\mathbf{z})$; ii) decoding the samples to reconstructions $Y_{\hat{\mathbf{z}}_i}$; ii) adding a noise value at the location: $Y_{\hat{\mathbf{z}}_i}(x_{n+1}) + \epsilon$. The outer integration of (2) can be approximated by the same importance sampling method with setting $\epsilon = 0$ for the sake of fast computation, which yields:

$$\mathbb{E}_{y_{n+1}}\Big[\mathrm{KL}\big(p_n(\mathbf{z} \mid y_{n+1}) \| p_n(\mathbf{z})\big)\Big] \approx \sum_{i=1}^{M} P_n(i) \mathrm{KL}(P_{n+1} \| P_n) + C. \tag{3}$$

By using this approximation, M samples of $\mathbf{z}$ at time n_s can be used to approximate (2) for any $n \geq n_s$ for all possible locations x_{n+1}. Note that $P_n(i) = 1/M$ if $n = n_s$, which means each sample is a perfect sample at time n. The approximation becomes worse if n is too different from n_s, as the estimator variance cause by the weights becomes larger, i.e., the effective sample size becomes lower [12].

It can be easily verified that the approximation (3) only depends on $(x_{n_s+1}, y_{n_s+1}) \sim (x_n, y_n)$, $\hat{\mathbf{z}}_1 \sim \hat{\mathbf{z}}_M$, and x_{n+1}. We call (3) the information gain, and the 2D map of information gain for each possible x_{n+1} is called the information gain map. The optimal x_{n+1} is the location having the maximum value on the map. From our experiments, it takes approximately 50 ms to calculate an entire information gain map with NVIDIA GTX 1080 TI when the full resolution is 128×128 and $M = 100$.

4 Embedding

To create the embedding space, we use a VAE with some modifications. Basically, the VAE encodes images of resolution 128×128 to 100-dimensional $\mathbf{z}$ vectors and decode them to reconstructions, of which resolution is again 128×128. Let Y^t and $\hat{Y}$ denote an training example and a reconstruction, respectively. The reconstruction loss of a plain VAE is the negative log-likelihood (NLL) of $p(Y^t|\hat{Y})$ as in (1). Because the reconstructions of the plain VAE is blurry, we use a contextual loss using a discriminator network that distinguishes whether an image is from training data or generated data. For a set of some selected layers of the discriminator network, denoted by K, the contextual loss is $l = \sum_{k \in \{0\} \cup K} a_k n_k^{-1} \mathbf{1} \cdot |h_k(Y^t) - h_k(\hat{Y})|$, where $h_k(\cdot)$ is the vectorised output of kth layer given a corresponding input, $k = 0$ indicates the input layer, n_k is the number of elements of the kth layer output, a_k is a weight for each layer, and $\mathbf{1}$ is a one-vector with an appropriate length. The layer weight a_k is periodically adjusted during training to make the effect of each layer same. The contextual loss is same with the NLL, if $K = \emptyset$. For better reconstruction result, 8×8 low-resolution images are additionally fed to the decoder, which is called conditional VAE (CVAE) [13]. We found that the additional information helps to improve global consistency of the reconstructions.

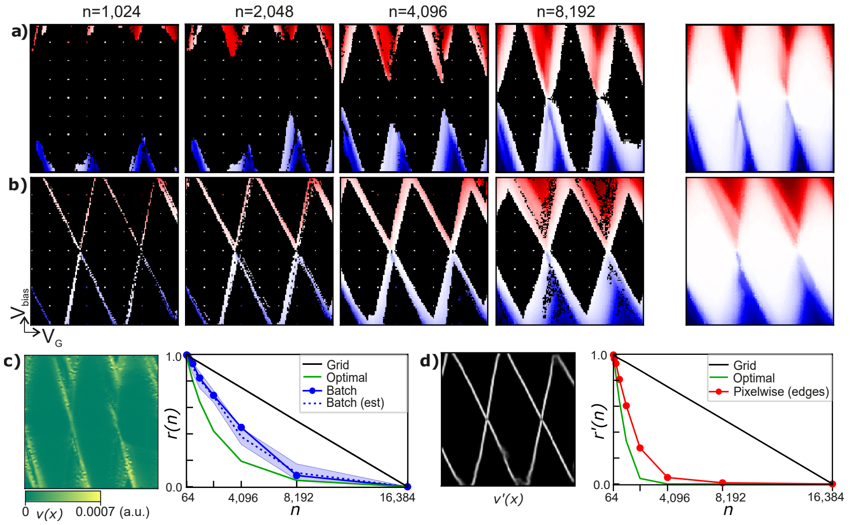

Figure 1: **a, c)** The batch algorithm with raw reconstructions. **b, d)** The pixel-wise algorithm with segmented maps of reconstructions.

5 Decision

After training the CVAE, Algorithm 1 can be used for the efficient measurement[2]. For the experiments, we use a Matropolis-Hastings sampler [14] with a Gaussian kernel for sampling $\hat{\mathbf{z}}_1, \ldots, \hat{\mathbf{z}}_M$. In terms of efficiency per pixel-count, choosing the single best location after each pixel observation is the best. However, changing the location to measure is expensive, because ramping voltages takes a long time. Therefore, we propose a batch selection method, which selects n_{batch} pixels at a time, not a single pixel. Then the path for the batch is optimised. We set $\mathcal{N}_b = \{2^i | i = 7, \ldots, 14\}$ for the experiments.

6 Result

Figure 1**a** shows the collected measurements for some selected n with the batch method. Black, red, blue, and white color in the images means no data, high current, low current, zero current, respectively. Figure 1c is the graph of proportional unmeasured gradient $r(n) = 1 - v(n)/v(N)$, where $N = 16,384$, v_x is the gradient magnitude at x on the full measurement Y_N, and $v(n) = \sum_{i=1}^{n} v_{x_i}$. The optimal line in Figure 1**c** is the theoretical optimum that no algorithm can exceed, computed by choosing the highest v_x with knowing the full measurement Y_N. The shaded region is a credible interval of real-time estimation of $r(n)$. The black line in Figure 1**c** is the performance of a multi-resolution grid method, which doubles the resolution of the scan on the given window, which results in low discrepancy points. In this paper, the performance of a Gaussian process regression based model is not compared, but the performance will be close to the black line, when integrated-mean-squared-error (IMSE) [2] is used as a selection criterion to reduce the uncertainty of a predictor, as it tends to spread points evenly. Details about time for computations and experiments, stopping criterion, and more examples can be found in [9].

Since the most important feature is the boundary between zero current regions and current-flowing regions, we developed a segmentation method that converts $\hat{Y}$ to a segmented map $f(\hat{Y})$ and applied the same criterion (3) to the segmented maps $f(\hat{Y}_1), \ldots, f(\hat{Y}_M)$. Figure 1**b** shows the measurement sequence, and the panel **d** shows the quantitative performance by setting v'_x the gradient magnitude at x on $f(\hat{Y}_N)$. We can see that the algorithm prioritise the boundary of interest.

[2] Code and an example available at `https://github.com/returnddd/CVAE_for_QE`

7 Conclusion

In this paper, we showed that information theoretic decision-making is useful for planning point-wise measurements with posterior samples of a deep generative model. It can be used to focus on where signals change, or more contextual information by transforming reconstructions to labels or quantities of interest. The limitation of this research is the availability of a training dataset. For some scientific experiments, a simple simulator is available, but the real-world measurement is far more complicated, which makes it impossible to use the simulator outcome as a training set. Transfer learning, domain adaptation, and style-transfer are promising candidates to connect the real world and a simple simulator.

References

[1] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian Active Learning for Classification and Preference Learning. 2011. Preprint at https://arxiv.org/abs/1112.5745.

[2] Jerome Sacks, William J. Welch, Toby J. Mitchell, and Henry P. Wynn. Design and Analysis of Computer Experiments. *Statistical Science*, 4(4):409–423, nov 1989.

[3] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[4] Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, December 1998.

[5] Peter Frazier, Warren Powell, and Savas Dayanik. The knowledge-gradient policy for correlated normal beliefs. *INFORMS Journal on Computing*, 21(4):599–613, 2009.

[6] Brochu Eric, Nando D. Freitas, and Abhijeet Ghosh. Active preference learning with discrete choice data. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 409–416. Curran Associates, Inc., 2008.

[7] Jose Miguel Hernandez-Lobato, Michael Gelbart, Matthew Hoffman, Ryan Adams, and Zoubin Ghahramani. Predictive entropy search for bayesian optimization with unknown constraints. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1699–1707, Lille, France, 07–09 Jul 2015. PMLR.

[8] Mark McLeod, Michael A. Osborne, and Stephen J. Roberts. Optimization, fast and slow: optimally switching between local and Bayesian optimization. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, May 2018. https://github.com/markm541374/gpbo.

[9] D. T. Lennon, H. Moon, L. C. Camenzind, Liuqi Yu, D. M. Zumbühl, G. A .D. Briggs, M. A. Osborne, E. A. Laird, and N. Ares. Efficiently measuring a quantum device using machine learning. *npj Quantum Information*, 5(1), sep 2019.

[10] R. Hanson, L. P. Kouwenhoven, J. R. Petta, S. Tarucha, and L. M.K. Vandersypen. Spins in few-electron quantum dots. *Reviews of Modern Physics*, 79(4):1217–1265, 2007.

[11] Arnaud Doucet, Nando de Freitas, and Neil Gordon. *Sequential Monte Carlo Methods in Practice*. Springer New York, 2001.

[12] Augustine Kong, Jun S. Liu, and Wing Hung Wong. Sequential imputations and bayesian missing data problems. *Journal of the American Statistical Association*, 89(425):pp. 278–288, 1994.

[13] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3483–3491. Curran Associates, Inc., 2015.

[14] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, apr 1970.