Stochastic Batch Acquisition for Deep Active Learning

Andreas Kirsch [*1] **Sebastian Farquhar** [*1] **Parmida Atighehchian** [2] **Andrew Jesson** [1]
Frederic Branchaud-Charron [2] **Yarin Gal** [1]

Abstract

We provide a stochastic strategy for adapting well-known acquisition functions to allow batch active learning. In deep active learning, labels are often acquired in batches for efficiency. However, many acquisition functions are designed for single-sample acquisition and fail when naively used to construct batches. In contrast, state-of-the-art batch acquisition functions are costly to compute. We show how to extend single-sample acquisition functions to the batch setting. Instead of acquiring the top-K points from the pool set, we account for the fact that acquisition scores are expected to change as new points are acquired. This motivates simple stochastic acquisition strategies using score-based or rank-based distributions. Our strategies outperform the standard top-K acquisition with virtually no computational overhead and can be used as a drop-in replacement. In fact, they are even competitive with much more expensive methods despite their linear computational complexity. We conclude that there is no reason to use top-K batch acquisition in practice.

1. Introduction

Active learning is a widely used strategy for efficient learning (Atlas et al., 1990; Settles, 2010). Often, unlabelled data is plentiful but labels are expensive. For example, labels for medical image data may require highly trained annotators and when labels are the results of scientific experiments each one can require months of work. Active learning uses information about the unlabelled data as well as the current state of the model to select labels that are most likely to be informative. In this way, as few labels as possible are sought in order to reach a given level of performance.

Most acquisition schemes are designed to acquire labels one at a time (e.g., Houlsby et al. (2011); Gal et al. (2017). Although they are highly effective at their intended goal, most existing active learning methods perform very poorly for batch selection. This makes it hard to parallelize labelling. For example, we might want to hire hundreds of annotators to work in parallel, or run more than one experiment at the same time. Single-point selection also greatly increases the cost of retraining the model for every new datapoint.

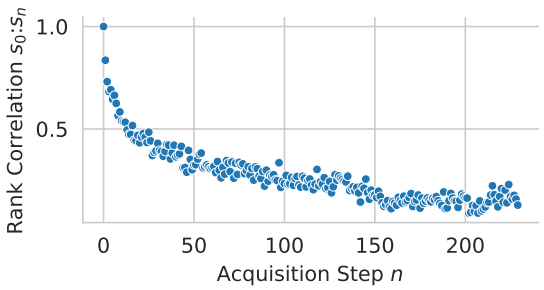


Figure 1. The current acquisition scores are only a loose proxy for later scores. Specifically, the Spearman rank-correlation between acquisition scores on the first and n'th time-step falls with n. Top-K acquisition implicitly assumes their rank-correlation remains 1, which we see is false. Our stochastic acquisition accounts for changing scores. Using neural network trained on MNIST initially with 20 points and 73% initial accuracy, rank over test set.

While there are precise and principled schemes which are specifically designed to acquire batches of points (Kirsch et al., 2019; Ash et al., 2020), they are expensive to compute and scale poorly to large batches because of combinatorial costs. Several recent works (Ash et al., 2020; 2021) trade off a principled motivation with various approximations to remain tractable.

A commonly used heuristic is to take the top-K highest scoring points from an acquisition scheme designed to select a single point. This suffers from a lack of diversity within batches because the scores ignore the *joint* informativeness for the model: they do not take redundancies between the highest scorers into account (Kirsch et al., 2019).

In this paper, we provide another perspective for the failure of top-K acquisition and use this to motivate a cheap and effective alternative to using top-K acquisition. Specifically, selecting the top-K points at acquisition step t amounts to

[*]Equal contribution [1]OATML, Department of Computer Science, University of Oxford [2]ServiceNow. Correspondence to: Andreas Kirsch <andreas.kirsch@cs.ox.ac.uk>, Sebastian Farquhar <sebastian.farquhar@cs.ox.ac.uk>.

Preliminary work.

Table 1. Acquisition runtime (seconds). Our stochastic acquisition methods are roughly as fast as top-K, and **orders of magnitude** faster than BADGE or BatchBALD. VGG-16 with $N = 10,000$ pool points with 10 classes. Times are acquisition only, no training. BatchBALD and BALD use 20 MC dropout samples.

K	Top-K	**Ours**	BADGE	BatchBALD
10	0.2 ± 0.0	0.2 ± 0.0	9.2 ± 0.3	566.0 ± 17.4
100	0.2 ± 0.0	0.2 ± 0.0	82.1 ± 2.5	$5,363.6 \pm 95.4$
500	0.2 ± 0.0	0.2 ± 0.0	409.3 ± 3.7	$29,984.1 \pm 598.7$

an assumption that the informativeness of these points is independent of each other. Imagine adding the top-K points at a given acquisition step t to the training set one by one. Each time, you retrain the model. Of course, the acquisition scores for the models trained with these additional samples will be different from the first set of scores. After all, the whole purpose of active learning is to add the *most informative* points—those that will update the model the most. Yet selecting a top-K batch all at once implicitly assumes that the score ranking will not be changed by the data. This is clearly wrong. We provide empirical confirmation that, in fact, the ranking of acquisition scores at step t and $t + K$ is decreasingly correlated as K grows (Figure 1). Moreover, this effect is especially strong for the most informative points (see §6.1 for more details).

Instead, we propose treating the current score ranking as a noisy approximation to future rankings. This can be implemented by sampling acquisitions from a distribution based on the current scores. We show empirically that this can result in better acquisitions than top-K and is competitive with more complicated custom algorithms like the clustering-based method BADGE (Ash et al., 2020) or the Bayesian information-based method BatchBALD (Kirsch et al., 2019) at a tiny fraction of the cost. Moreover, the cost scales linearly in K, unlike most batch acquisition schemes, and can be implemented with one line of code. We also emphasise that our approach is generally applicable to existing single-acquisition active learning acquisition functions and does not require Bayesian active learning.

In §2, we present active learning notation and commonly used acquisition functions. We propose our stochastic extensions in §3, relate them to previous works in §4, and validate them empirically in §5 on various datasets, showing that our method is competitive with much more complex ones despite being orders of magnitude computationally cheaper. Finally, we validate some of the underlying theoretical motivation in §6 and discuss limitations in §7.

2. Problem Setting

Our method applies to batch acquisition for active learning in a pool-based setting (Settles, 2010) where we have access to a large unlabelled *pool* set, but we can only label a small subset of the points. The challenge of active learning is to use what we already know in order to pick which points to label in the most efficient way. Generally, we want to avoid labelling points that are very similar to points that have already been labelled.

Notation. Following Farquhar et al. (2021), we formulate active learning over *indices* instead over datapoints. This simplifies the notation. The large, initially fully unlabelled, pool dataset containing M input points is

$$\mathcal{D}^{\text{pool}} = \{x_i\}_{i \in \mathcal{I}^{\text{pool}}}, \tag{1}$$

where $\mathcal{I}^{\text{pool}} = \{1, \dots, M\}$ is the initial full index set. We initialize a training dataset with N_0 randomly selected points from $\mathcal{D}^{\text{pool}}$ by acquiring their labels, y_i,

$$\mathcal{D}^{\text{train}} = \{(x_i, y_i)\}_{i \in \mathcal{I}^{\text{train}}}, \tag{2}$$

where $\mathcal{I}^{\text{train}}$ is the index set of $\mathcal{D}^{\text{train}}$ containing N_0 indices between 1 and M. A model of the predictive distribution, $\mathrm{p}(\hat{y} \mid x)$, can then be trained on $\mathcal{D}^{\text{train}}$.

Probabilistic Model. We assume classification with inputs X, true labels Y. The predicted labels are $\hat{Y}$, modelled using a discriminative classifier $\mathrm{p}(\hat{y} \mid x)$. In the case of Bayesian models we further assume a subjective probability distribution over the parameters, $\mathrm{p}(\omega)$, and have $\mathrm{p}(\hat{y} \mid x) = \mathbb{E}_{\mathrm{p}(\omega)}[\mathrm{p}(\hat{y} \mid x, \omega)]$.

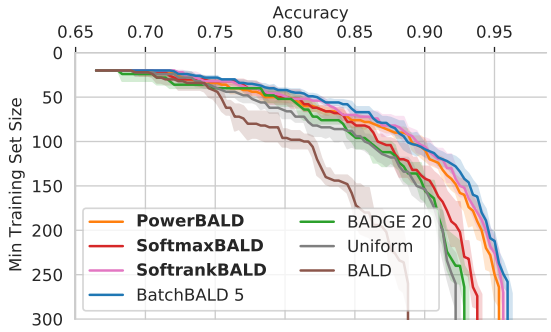
Active Learning. At each acquisition step, we select further points for which to acquire labels. While many methods acquire one point at a time (Houlsby et al., 2011; Gal et al., 2017), in general one can acquire a whole batch of K points. An acquisition function a takes $\mathcal{I}^{\text{train}}$ and $\mathcal{I}^{\text{pool}}$ and returns K indices from $\mathcal{I}^{\text{pool}}$ to be added to $\mathcal{I}^{\text{train}}$. We then label those K datapoints and add them to $\mathcal{I}^{\text{train}}$ while making them unavailable from the pool set. That is

$$\mathcal{I}^{\text{train}} := \mathcal{I}^{\text{train}} \cup a(\mathcal{I}^{\text{train}}, \mathcal{I}^{\text{pool}}), \tag{3}$$

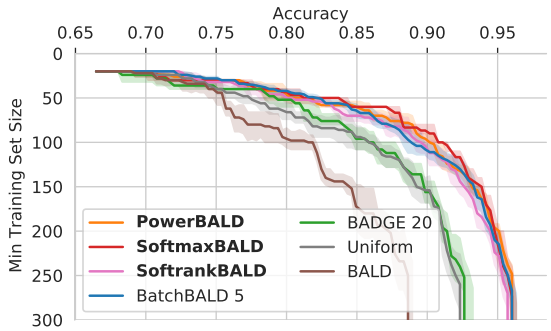
$$\mathcal{I}^{\text{pool}} := \mathcal{I}^{\text{pool}} \setminus \mathcal{I}^{\text{train}}. \tag{4}$$

A common way to construct the acquisition function is to define some scoring function, s, and to select the point(s) which score most highly.

BALD. One popular example of a scoring function is *BALD* (Houlsby et al., 2011) which uses a Bayesian model and



(a) Default $\beta = 1$



(b) PowerBALD, SoftmaxBALD: $\beta = 8$, Softrank: $\beta = 1$

Figure 2. Accuracy vs min. training set size on Repeated-MNIST with 4 repetitions (5 trials). **Up and to the right is better.** (a) Our stochastic acquisition functions outperform BALD and BADGE. Power- and SoftrankBALD are almost on par with BatchBALD. (b) Tuning β, we outperform SoTA BatchBALD. We plot accuracy versus the smallest training set size to reach said accuracy. We use acquisition batch sizes: BatchBALD–5, BALD and ours–10, BADGE–20. Figure 12 and 13 in the appendix show temperature ablations for the stochastic acquisition functions and ablations for BADGE's acquisition batch size.

computes the expected information gain between the predictive distribution and parameter distribution, $\Omega \mid \mathcal{D}^{\text{train}}$. For each candidate pool index, i, with mutual information, I, and entropy, H, the score is

$$\begin{aligned} s_{\text{BALD}}(i; \mathcal{I}^{\text{train}}) &\coloneqq \mathrm{I}[\hat{Y}; \Omega \mid X = x_i, \mathcal{D}^{\text{train}}] \\ &= \mathrm{H}[\hat{Y} \mid X = x_i, \mathcal{D}^{\text{train}}] \\ &\quad - \mathbb{E}_{\mathrm{p}(\omega \mid \mathcal{D}^{\text{train}})}[\mathrm{H}[\hat{Y} \mid X = x_i, \omega, \mathcal{D}^{\text{train}}]]. \end{aligned} \tag{5}$$

Entropy. Another common scoring function is the *(predictive) entropy* (Gal et al., 2017). It does not require Bayesian models, unlike BALD, but performs worse for data with high observation noise. It is identical to the first term of the BALD score

$$s_{\text{entropy}}(i; \mathcal{I}^{\text{train}}) \coloneqq \mathrm{H}[\hat{Y} \mid X = x_i, \mathcal{D}^{\text{train}}]. \tag{6}$$

Acquisition Functions. These scoring functions were introduced to be used for single-point acquisition:

$$a_s(\mathcal{I}^{\text{train}}) \coloneqq \operatorname*{arg\,max}_{i \in \mathcal{I}^{\text{pool}}} s(i; \mathcal{I}^{\text{train}}). \tag{7}$$

For deep learning in particular, single-point acquisition is computationally expensive, and it was assumed deep learning models rarely change much after adding a single additional point to the training set (but cf. Figure 1). Thus, single-point acquisition functions were trivially expanded to acquisition batches: the most commonly-used batch acquisition function naively selects the highest K scoring points

$$a_s^{\text{batch}}(\mathcal{I}^{\text{train}}; K) \coloneqq \operatorname*{arg\,max}_{I \subseteq \mathcal{I}^{\text{pool}}, |I| = K} \sum_{i \in I} s(i; \mathcal{I}^{\text{train}}). \tag{8}$$

Some acquisition functions are explicitly designed for batch acquisition (Kirsch et al., 2019; Ash et al., 2020). These often take into account the interaction between points, which can improve performance relative to simply selecting the top-K scoring points. However, existing methods are computationally expensive. For example, BatchBALD rarely scales to batch-sizes of more than around 5–10 points (Kirsch et al., 2019), see Table 1.

3. Method

Figure 1 shows how the initial acquisition scores quickly decorrelate from future scores. When we pick the top-K points at the current time step, we are not picking the points that are the most informative given the other selected points.

Instead, our work uses stochastic sampling to account for uncertainty using a simple model of the noise process governing how scores change. We examine three simple stochastic extensions of single-sample scoring functions $s(i; \mathcal{I}^{\text{train}})$ that make slightly different assumptions. Our methods are compatible with conventional active learning frameworks that typically take the top-K highest scoring samples. For example, predictive entropy and BALD are easily adapted.

Our stochastic acquisition distributions are built around the assumption that future scores are a perturbation of the current score by a noise distribution. In all cases, we model the noise distribution as the addition of a Gumbel-distributed random variable, which is frequently used for modelling extrema. The perturbation is applied to three basic quantities in the three sampling schemes. In particular, we consider the scores themselves, the log-scores, and the rank of the scores. Perturbing the log-scores amounts to an assumption that low-informativeness points ought to be actively avoided and that scores are non-negative. Perturbing the ranks can be seen as a robustifying assumption which requires the relative scores to be reliable but allows the absolute scores to be unreliable.

Here, we present the three versions with their associated sampling distributions, summarized in Table 2.

Table 2. Summary of stochastic acquisition variants. Perturbing the scores s_i themselves with $\epsilon_i \sim \text{Gumbel}(0; \beta^{-1})$ i.i.d. yields a softmax distribution. Log-scores result in a power distribution, with assumptions that are reasonable for active learning. Using the score-ranking, r_i finally is a robustifying assumption.

Perturb	Distribution	Probability mass
$s_i + \epsilon_i$	Softmax	$\propto \exp \beta s_i$
$\log s_i + \epsilon_i$	Power	$\propto s_i^{\beta}$
$-\log r_i + \epsilon_i$	Soft-rank	$\propto r_i^{-\beta}$

Softmax Acquisition. The most naive variant assumes that the scores are perturbed by a Gumbel-distributed random variable $\epsilon_i \sim \text{Gumbel}(0; \beta^{-1})$

$$s^{\text{softmax}}(i; \mathcal{I}^{\text{train}}) := s(i; \mathcal{I}^{\text{train}}) + \epsilon_i. \tag{9}$$

In fact, taking the highest-scoring points from this perturbed distribution is equivalent to sampling from a softmax/Boltzmann/Gibbs distribution without replacement with a 'coldness' parameter $\beta \geq 0$ which represents the rate at which the scores are expected to change as more data is acquired. This fact follows from the Gumbel-Max trick (Gumbel, 1954; Maddison et al., 2014) and more specifically the Gumbel-Top-K trick (Kool et al., 2019). Expanding on Maddison et al. (2014), we have

Proposition 3.1. *For scores s_i, $i \in \{1, \dots, n\}$, and $k \leq n$ and $\beta > 0$, if we draw $\epsilon_i \sim$ Gumbel$(0; \beta^{-1})$ independently, then $\arg top_k\{s_i + \epsilon_i\}_i$ is an (ordered) sample without replacement from the categorical distribution Categorical$(\frac{\exp(\beta s_i)}{\sum_j \exp(\beta s_j)}, i \in \{1, \dots, n\})$.*

As $\beta \to \infty$, this distribution will converge towards top-K acquisition, and for $\beta \to 0$ towards uniform acquisition. We provide a short proof in appendix A.

Power Acquisition. We extend this with a variant that assumes that scores are positive and scores that tend to zero guarantee that a sample is not informative, i.e. will not improve the model and we want to avoid sampling it. This is the case with commonly used acquisition scores like BALD and predictive entropy but not necessarily so with other scoring functions. In this case, we model the future log-scores as perturbations of the current log-score with Gumbel-distributed noise

$$s^{\text{power}}(i; \mathcal{I}^{\text{train}}) := \log s(i; \mathcal{I}^{\text{train}}) + \epsilon_i. \tag{10}$$

By Proposition 3.1, selecting the top-K s^{power} scores is equivalent to sampling from a power distribution

$$\mathrm{p}_{power}(i) \propto \left(\frac{1}{s(i; \mathcal{I}^{\text{train}})}\right)^{-\beta}. \tag{11}$$

This may be seen by noting that $\exp(\beta \log s(i; \mathcal{I}^{\text{train}})) = s(i; \mathcal{I}^{\text{train}})^{\beta}$.

Soft-Rank Acquisition. A final variant relaxes the assumption that the absolute scores are meaningful and relies on their rank order instead. This is potentially valuable in the common cases where the scores are estimated with high variance and where the *absolute* scores are unreliable but their *relative order* is reliable. However, if the absolute scores are accurate we would expect this method to perform worse than the others as it throws away the values of the actual scores.

Ranking the scores $s(i; \mathcal{I}^{\text{train}})$ with descending ranks $\{r_i\}_{i \in \mathcal{I}^{\text{pool}}}$ such that $s(r_i; \mathcal{I}^{\text{train}}) \geq s(r_j; \mathcal{I}^{\text{train}})$ for $r_i \leq r_j$ and smallest rank being 1, we sample index i with probability $\mathrm{p}_{\text{softrank}}(i) \propto r_i^{-\beta}$ with coldness β. This extension is invariant to the actual scores. Again, we can draw $\epsilon_i \sim \text{Gumbel}(0; \beta^{-1})$ once and set a perturbed rank

$$s^{\text{softrank}}(i; \mathcal{I}^{\text{train}}) := -\log r_i + \epsilon_i. \tag{12}$$

Taking the top-K samples is now equivalent to sampling without replacement from the rank distribution $\mathrm{p}_{\text{softrank}}(i)$.

When using BALD or entropy as underlying scoring function, power acquisition is generally the most sensible. The choice of a Gumbel distribution for the noise is largely one of mathematical convenience, as the maximum of sets of most distributions is not tractable. However, our methods work fairly well in practice, suggesting that the precise choice of perturbation is not critical for its effectiveness.

4. Related Work

Researchers in active learning (Atlas et al., 1990; Settles, 2010) have identified the importance of *batch* acquisition as well as the failures of top-K acquisition using straightforward extensions of single-sample methods in a range of settings including support-vector machines (Campbell et al., 2000; Schohn & Cohn, 2000; Brinker, 2003; Guo & Schuurmans, 2008), GMMs (Azimi et al., 2012), and neural networks (Sener & Savarese, 2018; Kirsch et al., 2019; Ash et al., 2020; Baykal et al., 2021). Many of these methods aim to introduce structured diversity to batch acquisition that accounts for the *interaction* of the acquired labels on the learning process. In most cases, the computational complexity scales poorly with the batch-size (K) or pool-size (M), for example because of the estimation of joint mutual information (Kirsch et al., 2019), the $\mathcal{O}(KM)$ complexity of using a k-means++ initialisation scheme (Ash et al., 2020), or the $\mathcal{O}(M^2 \log M)$ complexity of methods based on K-centre coresets (Sener & Savarese, 2018) (although heuristics and continuous relaxations can improve this somewhat). In contrast, our method is linear in M and has complexity $\mathcal{O}(\log K)$ in batch-size. In exchange, our method does not directly enforce diversity in the batch by modelling distances.

Sampling stochastically has not been extensively explored

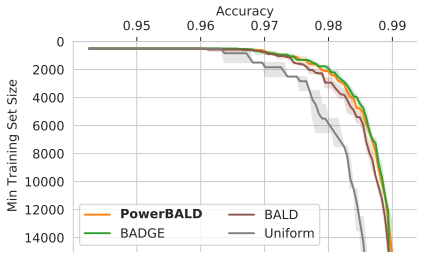


Figure 3. Accuracy vs min. training set size on MIO-TCD (3 trials). **Up and to the right is better.** PowerBALD outperforms BALD and performs on par with BADGE—each with acquisition batch size 100. So does SoftmaxBALD (c.f. appendix C.2). See the appendix for more details.

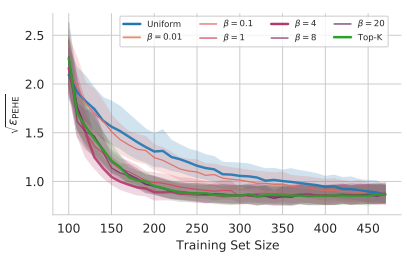


Figure 4. CausalBALD: IHDP dataset using power acquisition (400 trials). **Down and left is better.** At high temperature ($\beta = 0.1$), power acquisition is like random acquisition. As the temperature decreases, the performance improves (lower $\sqrt{\epsilon_{\text{PEHE}}}$), surpassing top-$K$ acquisition.

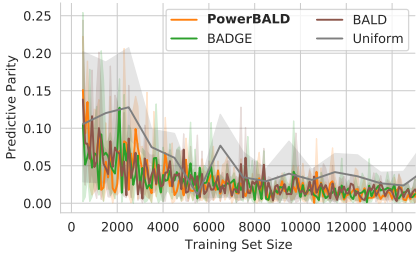


Figure 5. Spurious correlation experiments (3 trials). **Down and left is better.** Stochastic acquisition matches BADGE and BALD's predictive parity and performance (see §C.2.2 for the latter), which is reassuring as stochastic acquisition functions might be affected by spurious correlations.

for acquisition in active learning. It has been used as a step in clustering (Ash et al., 2020; Citovsky et al., 2021). Farquhar et al. (2021) propose stochastic acquisition as part of de-biasing actively learned estimators. Kirsch et al. (2019) note empirically that additional noise in scores can benefit batch acquisition, without further investigation. To our knowledge, this work is the first to compare stochastic acquisition methods as alternatives to naive top-K acquisition.

Stochastic prioritization has, however, been employed in reinforcement learning as *prioritized replay* (Schaul et al., 2016) which may be effective for reasons which are analogous to those motivating our approach.

5. Experiments

In this section, we empirically verify that our stochastic acquisition methods: outperform top-K acquisition; are generally competitive with specially-designed batch acquisition schemes like BADGE (Ash et al., 2020) and BatchBALD (Kirsch et al., 2019); and are vastly cheaper than these more complicated methods.

We demonstrate this with a range of experiments including computer vision and causal inference. We show that stochastic acquisition helps avoid selecting redundant samples on Repeated-MNIST (Kirsch et al., 2019), examine performance on CIFAR-10 and MIO-TCD (Luo et al., 2018), which is closer to a real-world dataset, and investigate edges cases on Synbols using different types of biases, class distributions and aleatoric uncertainty distributions.

We consider both BALD and predictive entropy as single-acquisition baselines and compare to our three variant methods. We focus on Power Acquisition in the main body of the paper as it fits BALD and entropy best: both scores are non-negative and zero scores imply uninformative samples. We ablate the performance and compare the three suggested stochastic acquisition types in §6 as well as providing performance results for all variants in appendix C.2 and C.1.

Experimental Setup. We document the experimental setup and model architectures in detail in appendix B.

5.1. Runtime Measurements

We emphasise that our method is computationally extremely efficient compared to specialized batch-acquisition approaches like BADGE and BatchBALD. Runtimes, shown in Table 1, are essentially identical for top-K and our stochastic acquisition. Both are orders of magnitude faster than BADGE and BatchBALD even for small batches. Unlike those methods, ours scales *linearly* in both pool size and batch size. Runtime numbers exclude the cost of training itself, since this is the same for all methods. The runtimes for top-K and stochastic acquisition appear constant over K simply because the execution time is dominated by fixed-cost memory operations up to the tenth-of-a-second precision displayed. The synthetic dataset used for benchmarking has 4,096 features, 10 classes, and 10,000 pool points. VGG-16 models were used for sampling predictions and latent embeddings.

5.2. Repeated-MNIST

Repeated-MNIST was introduced by Kirsch et al. (2019) to demonstrate pathologies in batch acquisition which are caused by redundancies in datasets. Redundant data are incredibly common in industry applications, but are artificially removed from standard benchmarks. Repeated-MNIST duplicates MNIST several times (specified using a dataset parameter) and adds Gaussian noise to prevent identical duplicates. We use an acquisition batch size of 10 and use 4 dataset repetitions. We use $\beta = 1$ as default for all stochastic acquisition functions. We also report results with tuned $\beta = 8$ for power and softmax acquisition and $\beta = 1$ for softrank acquisition.

Figure 2, shows the advantage of using stochastic acquisition for BALD. All three sampling schemes clearly outperform top-K BALD. They perform comparably to BatchBALD—BatchBALD is SOTA for small batch sizes—but our methods are much cheaper. For BatchBALD we have an acquisition batch size 5 because it becomes computationally infeasible for larger batches.

Our methods also outperform BADGE, despite being considerably cheaper again. We use an acquisition batch size of 20 for BADGE in order to *strengthen* the baseline—BADGE performs much better with larger batch sizes as we demonstrate Figure 13 in the appendix. Lines on figures interpolate linearly between available points. We mark the 95% confidence intervals.

5.3. Computer Vision: MIO-TCD

The Miovision Traffic Camera Dataset (MIO-TCD) (Luo et al., 2018) is a vehicle classificiation and localization dataset with 500,000 images designed to have 'realistic' data problems like class imbalance, duplicate data, compression artefacts and uninformative examples. For added complexity, the image scales vary widely, like natural data, with image widths between 100 and 2,000 pixels (see Figure 11 in the appendix).

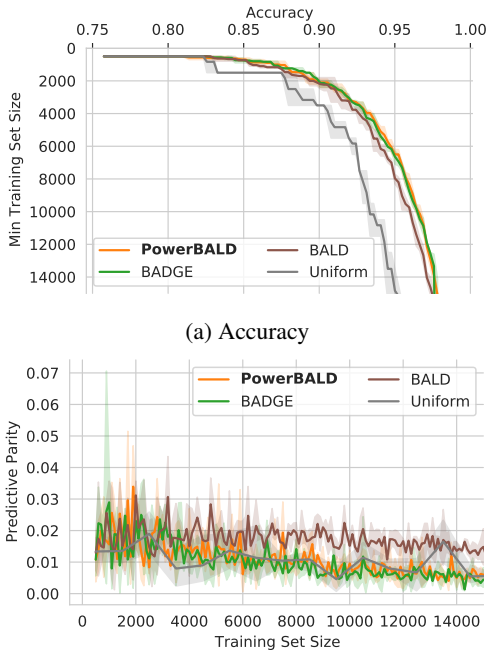
As we show in Figure 3, PowerBALD performs almost identically to BADGE despite three orders of magnitude lower computational cost, and both perform slightly better than BALD and much better than uniform acquisition.

For all methods, the model is a VGG-16 trained for 10 epochs using Monte Carlo dropout for acquisition (Gal et al., 2017) with 20 dropout samples. We acquire examples with batchsizes of 100 for all methods and $\beta = 1$. In appendix §C.2, we show the softmax and softrank methods perform comparably with BALD on this dataset and discuss hyperparameter choices.

5.4. Causal Treatment Effects: Semi-Synthetic Data and Infant Health Development Programme

To assess our methods beyond computer vision, we examine active learning for Conditional Average Treatment Effect (CATE) estimation (Heckman et al., 1997; 1998; Hahn, 1998; Abrevaya et al., 2015) on data from the Infant Health and Development Program (IHDP) estimating the causal effect of treatments on infant's health from observational data. CATE can be estimated probabilistically from observational data under certain assumptions and Jesson et al. (2021) show how to actively acquire data for label-efficient estimation. Among other subtleties, this includes prioritizing data for which matched treated/untreated pairs are available.

We follow the experiments of Jesson et al. (2021) on both synthetic data and the semi-synthetic IHDP dataset (Hill, 2011) which is commonly used in the causal effects estimation literature. In Figure 4 we show that power acquisition performs significantly better than either top-K or uniform acquisition, using a acquisition batch size of 10 in all cases with further ablations on synthetic data in appendix C.3. Note that methods like BADGE and BatchBALD are not well-defined in the causal effects estimation context, while our approach remains effective.



(a) Accuracy

(b) Predictive parity

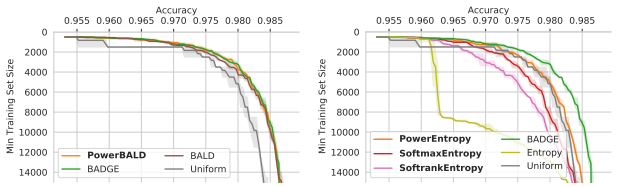
Figure 6. Under-represented groups (3 trials). PowerBALD slightly outperforms BALD and matches BADGE for both accuracy and predictive parity on an unbalanced Synbols dataset.

Performance on these tasks is measured using the expected Precision in Estimation of Heterogeneous Effect (PEHE (Hill, 2011)) such that $\sqrt{\epsilon_{\text{PEHE}}} = \sqrt{\mathbb{E}[(\tilde{\tau}(\mathbf{X}) - \tau(\mathbf{X}))^2]}$ (Shalit et al., 2017) where $\tilde{\tau}$ is the estimated CATE and τ is CATE (i.e., a form of RMSE).

5.5. Edge Cases: Synbols

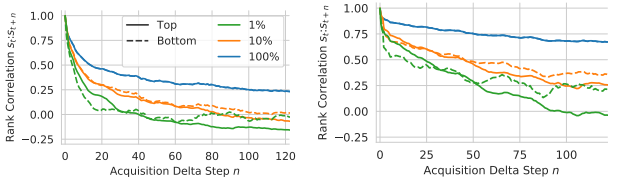
We use Synbols (Lacoste et al., 2020) to demonstrate the behaviour of batch active learning in artificially constructed edge cases. Synbols is a character dataset generator for classification where a user can specify the type and proportion of bias and insert artefacts, backgrounds, masking shapes, and so on. We selected three datasets with strong biases supplied by Lacoste et al. (2020); Branchaud-Charron et al. (2021) to evaluate our method. Experimental settings are similar to §5.3 with details in the appendix B. We use the default $\beta = 1$.

For these tasks, performance evaluation includes 'predic-



(a) BALD accuracy (b) Entropy accuracy

Figure 7. Missing Symbols (3 trials). In this dataset with high aleatoric uncertainty, PowerBALD matches BADGE and BALD performance. PowerEntropy significantly outperforms Entropy which confounds aleatoric and epistemic uncertainty.



(a) $t = 0$ (b) $t = 100$

Figure 8. Rank correlations for BALD scores on MNIST. Rank-orders decorrelate faster for the most informative samples and in the early stages of training. We consider the top or bottom 1%, 10% and 100% of points, and see to what extent the initial scores correlate with future scores in their rankings. **(a)** Initial correlations rapidly disappear, especially for the most informative points. The top-1% scorers' ranks in fact *anti-correlate* after roughly 50 acquisitions. This shows that initially informative samples become the least informative later—after similar samples were acquired. The bottom scorers tend towards towards being uncorrelated and scores are nearly 0 throughout training. **(b)** Later in training, $t = 100$, the acquisition scores stay more strongly correlated, suggesting that *the acquisition batch size could be increased later in training*. Rank correlations were smoothed with a size 10 Parzen window.

tive parity', also known as 'accuracy difference', which is the maximum difference in accuracy between subgroups—which are, in this case, different coloured characters. This measure is most widely used in domain adaptation and ethics (Verma & Rubin, 2018). We want to maximise the accuracy while minimising the predictive parity.

Spurious Correlations. This dataset includes spurious correlations between the colour of the character and its class. As shown in Branchaud-Charron et al. (2021), active learning is especially strong here as characters that do not follow the correlation will be informative and thus selected.

We compare the predictive parity between methods in Fig. 5. We do not see any significant difference between our method and BADGE or BALD. This is encouraging as stochastic approaches might select more examples following the spurious correlation and thus have higher predictive parity, but this is not the case.

Under-Represented Groups. This dataset includes a subgroup of the data is under-represented, specifically most characters are red while few are blue. As Branchaud-Charron et al. (2021) show, active learning can improve accuracy for these groups.

Our stochastic approach lets batch acquisition better capture under-represented subgroups. In Figure 6 we show that PowerBALD has almost identical accuracy to BADGE, despite being much cheaper, and outperforms BALD. At the same time (Figure 6b), PowerBALD has much lower predictive parity than BALD, demonstrating a fairer predictive distribution given the unbalanced dataset.

Missing Symbols. This dataset has high aleatoric uncertainty. Some images are missing information required to make high-probability predictions—these images have shapes randomly occluding the character—so even a perfect model would remain uncertain. Lacoste et al. (2020) demonstrated that entropy is ineffective on this data as it cannot distinguish between aleatoric and epistemic uncertainty, while BALD can do so. As a consequence, entropy will unfortunately preferably select samples with occluded characters, resulting in degraded active learning performance.

For BALD, we show in Figure 7a that as before our stochastic method performs on par with BADGE and comparable to BALD (perhaps marginally better). In contrast, for predictive entropy stochastic acquisition largely corrects the failure of entropy acquisition to account for missing data (Figure 7b) although PowerEntropy still underperforms BADGE here.

6. Investigation

In this section, we validate our assumptions of the underlying score dynamics by examining the score rank correlations across acquisitions and hypothesize about when top-K acquisition is the most detrimental to active learning.

6.1. Rank Correlations Across Acquisitions

Our method is based on assuming: (1) the acquisition scores s_t at step t are a proxy for scores $s_{t'}$ at step $t' > t$; (2) the larger $t' - t$ is, the worse a proxy s_t is for s'_t; (3) this effect is biggest for the most informative points.

We demonstrate these empirically by examining the Spearman rank correlation between scores during acquisition. Specifically, we actively train a model for n steps using greedy single-point acquisition functions (BALD). For each step, we compare the rank-order at that step to the starting rank order at step t.

Figure 1 shows that acquisition scores become less correlated as more points are acquired. Figure 8a shows this in more detail for the top and bottom 1%, 10% or 100% of scorers of the test set across acquisitions starting at step $t = 0$ for

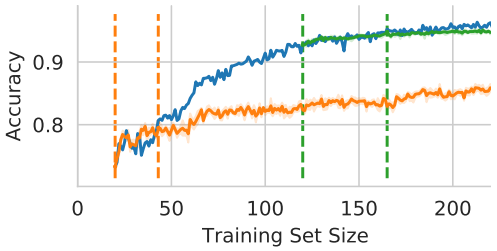


Figure 9. Top-K acquisition is less detrimental later in training. At $t \in \{20, 100\}$ of single run with individual BALD acquisition on MNIST (blue), we instead keep acquiring samples using the BALD scores at two those steps. Starting at training set size 20 (orange), the model performs well up to an acquisition batch size of 20 before the training trajectory visibly diverges; at training set size 120 (green), up to an acquisition batch size of 50. See §6.2.

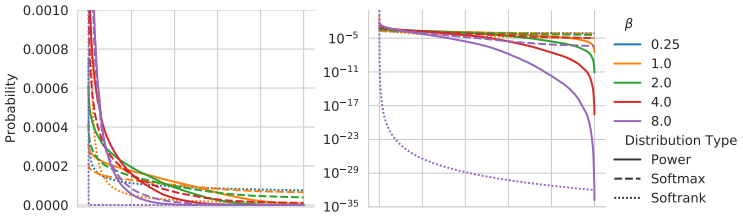


Figure 10. Score distribution for power and softmax acquisition of BALD scores on MNIST for varying Coldness β at $t = 0$. Linear and log plot over samples sorted by their BALD score. At $\beta = 8$ both softmax and power acquisition have essentially the same distribution for high scoring points (closely followed by the power distribution for $\beta = 4$). This might explain why the coldness ablation shows that these β to have very similar AL trajectories on MNIST. Yet, while softmax and power acquisition seem transfer to RMNIST, this is not the case for softrank which is much more sensitive to β. At the same time, power acquisition avoids low-scoring points more than softmax acquisition.

a model initialized with 20 points. The ranks of the top-10% scoring points (solid green) become quickly uncorrelated with future scores, and actually become *anti-correlated*. In contrast, the points overall (solid blue) correlate fairly well over time (though they have a much weaker training signal on average). This supports all three of our hypotheses.

6.2. Increasing Top-K Analysis

At the same time, we see that as training progresses and we converge towards the best model, the order of scores becomes more stable across acquisitions. In Figure 8b the model begins with 120 points, rather than 20. In the latter case, the most informative points are less likely to completely change the ordering—even the top-1% ranks do not become *anti-correlated*, only de-correlated.

Another way to investigate the effect of top-K selection is to freeze the acquisition scores during training and then continue 'active learning' as though those were the correct scores. We perform this toy experiment, showing that freezing scores early on harms performance greatly while doing it later has only a small effect (Figure 9). For frozen scores from a training set size of 20 (73% accuracy, $t = 0$), accuracy matches single-acquisition BALD until a training set size of roughly 40 (dashed orange lines) before diverging to a lower level. But freezing scores for a more accurate model, at a training set size of 120 labels (93% accuracy, $t = 100$), just selecting the next fifty points according to those frozen scores performs indistinguishably from step-by-step acquisition (dashed green lines). This shows that top-K acquisition hurts less later during training but can negatively affect performance massively at the start of training.

This raises the question of whether we ought to dynamically change the acquisition batch size during training: with smaller acquisition batches at the beginning and larger ones towards the end. We leave exploring this for future work.

6.3. Comparing Power, Softmax and Soft-Rank

To examine the three stochastic acquisition variants, we plot their score distributions, extracted from the same MNIST toy example, in Figure 10. Power and softmax acquisition distributions are similar for $\beta = 8$ (power, softmax) and $\beta = 4$ (softmax). This might explain why active learning with these β shows similar accuracy trajectories.

We find that power and softmax acquisition are quite insensitive to β and selecting $\beta = 1$ generally works well. Except where noted otherwise, we therefore use $\beta = 1$, since it is often not practical to tune this hyperparameter in the real world. In contrast, softrank acquisition is much more sensitive to its β parameter. This is also evidenced in the temperature ablations in §C.1 and §C.2.1 in the appendix.

7. Discussion & Conclusion

We have provided an alternative approach to batch acquisition for active learning. Our stochastic method is dramatically faster than sophisticated batch-acquisition strategies like BADGE and BatchBALD, while retaining comparable performance in a wide range of settings. At the same time, it is sometimes better and never worse than the naive top-K acquisition heuristic which is commonly used, though flawed. We see no reason to continue using top-K acquisition.

At the same time, our analysis opens doors for future research. Although our stochastic model was chosen for computational and mathematical simplicity, future work could explore more sophisticated modelling of the predicted changes in scores that take both the current model and dataset into account. In its simplest form, this might mean adapting the temperature of the acquisition distribution to the dataset or estimating it online. Our experiments also highlight that acquisition batch size could to be dynamic, with larger batch sizes acceptable later in training.

References

Abrevaya, J., Hsu, Y.-C., and Lieli, R. P. Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, 33(4):485–505, 2015.

Ash, J. T., Zhang, C., Krishnamurthy, A., Langford, J., and Agarwal, A. Deep batch active learning by diverse, uncertain gradient lower bounds, 2020.

Ash, J. T., Goel, S., Krishnamurthy, A., and Kakade, S. Gone fishing: Neural active learning with fisher embeddings, 2021.

Atighehchian, P., Branchaud-Charron, F., and Lacoste, A. Bayesian active learning for production, a systematic study and a reusable library. In *ICML Workshop on uncertainty and robustness in deep learning*, 2020.

Atlas, L., Cohn, D., and Ladner, R. Training Connectionist Networks with Queries and Selective Sampling. *Neural Information Processing Systems*, 1990.

Azimi, J., Fern, A., Zhang-Fern, X., Borradaile, G., and Heeringa, B. Batch Active Learning via Coordinated Matching. *International Conference on Machine Learning*, 2012.

Baykal, C., Liebenwein, L., Feldman, D., and Rus, D. Low-regret active learning, 2021.

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., and Walker, K. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft, May 2020.

Branchaud-Charron, F., Atighehchian, P., Rodríguez, P., Abuhamad, G., and Lacoste, A. Can active learning preemptively mitigate fairness issues? *ICLR Workshop on Responsable AI*, 2021.

Brinker, K. Incorporating Diversity in Active Learning with Support Vector Machines. *International Conference on Machine Learning*, 2003.

Campbell, C., Cristianini, N., and Smola, A. Query Learning with Large Margin Classifiers. *International Conference on Machine Learning*, 2000.

Citovsky, G., DeSalvo, G., Gentile, C., Karydas, L., Rajagopalan, A., Rostamizadeh, A., and Kumar, S. Batch active learning at scale, 2021.

Farquhar, S., Gal, Y., and Rainforth, T. On statistical bias in active learning: How and when to fix it. *International Conference on Learning Representations*, 2021.

Gal, Y., Islam, R., and Ghahramani, Z. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pp. 1183–1192. PMLR, 2017.

Gumbel, E. J. *Statistical theory of extreme values and some practical applications: a series of lectures*, volume 33. US Government Printing Office, 1954.

Guo, Y. and Schuurmans, D. Discriminative Batch Mode Active Learning. In *Advances in Neural Information Processing Systems*, 2008.

Hahn, J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pp. 315–331, 1998.

Heckman, J. J., Ichimura, H., and Todd, P. E. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The review of economic studies*, 64(4):605–654, 1997.

Heckman, J. J., Ichimura, H., and Todd, P. Matching as an econometric evaluation estimator. *The review of economic studies*, 65(2):261–294, 1998.

Hill, J. L. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Houlsby, N., Huszár, F., Ghahramani, Z., and Lengyel, M. Bayesian active learning for classification and preference learning. *arXiv:1112.5745*, 2011.

Jesson, A., Tigas, P., van Amersfoort, J., Kirsch, A., Shalit, U., and Gal, Y. Causal-bald: Deep bayesian active learning of outcomes to infer treatment-effects from observational data. *Advances in Neural Information Processing Systems*, 34, 2021.

Kirsch, A., van Amersfoort, J., and Gal, Y. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In *Advances in Neural Information Processing Systems*, pp. 7024–7035, 2019.

Kool, W., Van Hoof, H., and Welling, M. Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement. In *International Conference on Machine Learning*, pp. 3499–3508. PMLR, 2019.

Lacoste, A., Rodríguez, P., Branchaud-Charron, F., Atighehchian, P., Caccia, M., Laradji, I., Drouin, A., Craddock, M., Charlin, L., and Vázquez, D. Synbols: Probing learning algorithms with synthetic datasets. *NeurIPS*, 2020.

Luo, Z., Branchaud-Charron, F., Lemaire, C., Konrad, J., Li, S., Mishra, A., Achkar, A., Eichel, J., and Jodoin, P.-M. Mio-tcd: A new benchmark dataset for vehicle classification and localization. *IEEE Transactions on Image Processing*, 27(10):5129–5141, 2018.

Maddison, C. J., Tarlow, D., and Minka, T. A* sampling. In *NIPS*, 2014.

Neyman, J. edited and translated by dorota m. dabrowska and terrence p. speed (1990). on the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472, 1923.

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Rubin, D. B. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980.

Schaul, T., Quan, J., Antonoglou, I., and Silver, D. Prioritized experience replay, 2016.

Schohn, G. and Cohn, D. Less is more: Active learning with support vector machines. pp. 839–846. Morgan Kaufmann, 2000.

Sekhon, J. S. The neyman-rubin model of causal inference and estimation via matching methods. *The Oxford handbook of political methodology*, 2:1–32, 2008.

Sener, O. and Savarese, S. Active Learning for Convolutional Neural Networks: A Core-Set Approach. 2018.

Settles, B. Active Learning Literature Survey. *Machine Learning*, 2010.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pp. 3076–3085. PMLR, 2017.

van Amersfoort, J., Smith, L., Jesson, A., Key, O., and Gal, Y. Improving deterministic uncertainty estimation in deep learning for classification and regression. *arXiv*, 2021.

Verma, S. and Rubin, J. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pp. 1–7. IEEE, 2018.

A. Proof of Proposition 3.1

First, we remind the reader that a random variable G is Gumble distributed $G \sim \text{Gumbel}(\mu; \beta)$ when its cumulative distribution function follows $\text{p}(G \leq g) = \exp(-\exp(-\frac{g-\mu}{\beta}))$.

Furthermore, the Gumbel distribution is closed under translation and positive scaling:

Lemma A.1. *Let $G \sim$ Gumbel$(\mu; \beta)$ be a Gumbel distributed random variable, then:*

$$\alpha G + d \sim \textit{Gumbel}(d + \alpha\mu; \alpha\beta). \tag{13}$$

Proof. We have $\text{p}(\alpha G + d \leq x) = \text{p}(G \leq \frac{x-d}{\alpha})$. Thus, we have:

$$\text{p}(\alpha G + d \leq x) = \exp(-\exp(-\frac{\frac{x-d}{\alpha} - \mu}{\beta})) \tag{14}$$

$$= \exp(-\exp(-\frac{x - (d + \alpha\mu)}{\alpha\beta})) \tag{15}$$

$$\Leftrightarrow \alpha G + d \sim \text{Gumbel}(d + \alpha\mu; \alpha\beta). \tag{16}$$

□

We can then easily prove Proposition 3.1 using Theorem 1 from Kool et al. (2019), which we present it here slightly reformulated to fit our notation:

Lemma A.2. *For $k \leq n$, let $I_1^*, \ldots, I_k^* = \arg\text{top}_k\{s_i + \epsilon_i\}_i$ with $\epsilon_i \sim$ Gumbel$(0; 1)$, i.i.d.. Then $I_1^*, \ldots, I_k^*$ is an (ordered) sample without replacement from the Categorical $\left(\frac{\exp s_i}{\sum_{j \in n} \exp s_j}, i \in \{1, \ldots, n\}\right)$ distribution, e.g. for a realization $i_1^*, \ldots, i_k^*$ it holds that*

$$P\left(I_1^* = i_1^*, \ldots, I_k^* = i_k^*\right) = \prod_{j=1}^{k} \frac{\exp s_{i_j^*}}{\sum_{\ell \in N_j^*} \exp s_\ell}$$

where $N_j^* = N \backslash \left\{i_1^*, \ldots, i_{j-1}^*\right\}$ is the domain (without replacement) for the j-th sampled element.

Now, it is easy to prove the proposition:

Proposition 3.1. *For scores s_i, $i \in \{1, \ldots, n\}$, and $k \leq n$ and $\beta > 0$, if we draw $\epsilon_i \sim$ Gumbel$(0; \beta^{-1})$ independently, then $\arg top_k\{s_i + \epsilon_i\}_i$ is an (ordered) sample without replacement from the categorical distribution Categorical$(\frac{\exp(\beta\, s_i)}{\sum_j \exp(\beta\, s_j)}, i \in \{1, \ldots, n\})$.*

Proof. As $\epsilon_i \sim \text{Gumbel}(0; \beta^{-1})$, define $\epsilon_i' \coloneqq \beta\epsilon_i \sim \text{Gumbel}(0; 1)$. Further, let $s_i' \coloneqq \beta s_i$. Applying Lemma A.2 on s_i' and ϵ_i', $\arg\text{top}_k\{s_i' + \epsilon_i'\}_i$ yields (ordered) samples without replacement from the categorical distribution Categorical$(\frac{\exp(\beta\, s_i)}{\sum_j \exp(\beta\, s_j)}, i \in \{1, \ldots, n\})$. However, multiplication by β does not change the resulting indices of $\arg\text{top}_k$:

$$\arg\text{top}_k\{s_i' + \epsilon_i'\}_i = \arg\text{top}_k\{s_i + \epsilon_i\}_i, \tag{17}$$

concluding the proof. □

B. Experimental setup

Full code for all experiments will be available at `anonymized_github_repo`.

B.1. Repeated-MNIST.

We used the same setup as Kirsch et al. (2019); a LeNet-5 is trained with early stopping using the Adam optimizer and a learning rate of 0.001. We sample predictions using 100 MC-Dropout samples for BALD. The weights are reinitialized after each acquisition step.

The Repeated-MNIST dataset is constructed as in Kirsch et al. (2019) with duplicated examples from MNIST with isotropic Gaussian noise with standard deviation 0.1 added to the input features.

B.2. Synbols & MIO-TCD.

The full list of hyperparameters for the Synbols and MIO-TCD experiments is presented in Table 3. Our experiments are built using the BaaL library (Atighehchian et al., 2020). We compute predictive parity using FairLearn (Bird et al., 2020). Results shown in Table 1 were run inside Docker containers with 8 CPUs (2.2Ghz) and 32 Gb of RAM.

In Figure 11, we show a set of images with common issues we can find in MIO-TCD.

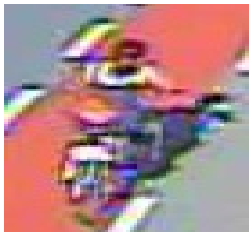
B.3. CausalBALD

Using the Neyman-Rubin framework (Neyman, 1923; Rubin, 1974; Sekhon, 2008), the CATE is formulated in terms of the potential outcomes, Y_t, of treatment levels $\text{t} \in \{0, 1\}$. Given observable covariates, $\mathbf{X}$, the CATE is defined as the expected difference between potential outcomes at measured value $\mathbf{X} = \mathbf{x}$: $\tau(\mathbf{x}) = \mathbb{E}[\text{Y}_1 - \text{Y}_0 \mid \mathbf{X} = \mathbf{x}]$. This causal quantity is fundamentally unidentifiable from observational data without further assumptions, because it is not possible to observe both Y_1 and Y_0 for a given unit. However, under the assumptions of consistency, non-interference, ignorability, and positivity, the CATE is identifiable as the statistical quantity $\widetilde{\tau}(\mathbf{x}) = \mathbb{E}[\text{Y} \mid \text{T} = 1, \mathbf{X} = \mathbf{x}] - \mathbb{E}[\text{Y} \mid \text{T} = 0, \mathbf{X} = \mathbf{x}]$ (Rubin, 1980).

Jesson et al. (2021) define BALD acquisition functions for active learning CATE functions from observational data

(a) A good example in MIOTCD dataset.

(b) An example of duplicated samples in the dataset.



(c) An example of class confusion between motorcycle and bicycle.

(d) An example of heavy compression artefact.

(e) An example of low resolution samples.

Figure 11. MIO-TCD Dataset is designed to include common artefacts from production data. The size and quality of the images vary greatly between crops; from high-quality cameras on sunny days to low-quality cameras at night. (a) shows an example of clean samples that can be clearly assigned to a class. (b)(c)(d) and (e) show the different categories of noise. (b) shows an example of many near-duplicates that exist in the dataset. (c) is a good example where the assigned class is subject to interpretation (d) is a sample with heavy compression artefacts and (e) is an example of samples with low resolution which again is considered a hard example to learn for the model.

Table 3. Hyper-parameters used in Section 5.3 and 5.5

Hyperparameter	Value
Learning rate	0.001
Optimizer	SGD
Weight decay	0
Momentum	0.9
Loss function	Crossentropy
Training duration	10
Batch size	32
Dropout p	0.5
MC iterations	20
Query size	100
Initial set	500

when the cost of acquiring an outcome, y, for a given covariate and treatment pair, $(\mathbf{x}, \mathrm{t})$, is high. Because we do not have labels for Y_1 and Y_0 for each $(\mathbf{x}, \mathrm{t})$ pair in the dataset, their acquisition function focuses on acquiring data points $(\mathbf{x}, \mathrm{t})$ for which it is likely that a matched pair $(\mathbf{x}, 1 - \mathrm{t})$ exists in the pool data or has already been acquired at a previous step. We follow their experiments on their synthetic dataset with limited positivity, and the semi-synthetic IHDP dataset (Hill, 2011). Details of the experimental setup are given in (Jesson et al., 2021), we use their provided code, and implement the power acquisition function.

The settings for causal inference experiments are identical to those used in Jesson et al. (2021), using the IHDP dataset (Hill, 2011). Like them, we use a Deterministic Uncertainty Estimation model (van Amersfoort et al., 2021) which are initialized with 100 datapoints and acquire 10 datapoints per acquisition batch for 38 steps. The dataset has 471 pool points and a 201 point validation set.

C. Further experimental analysis

C.1. Repeated-MNIST

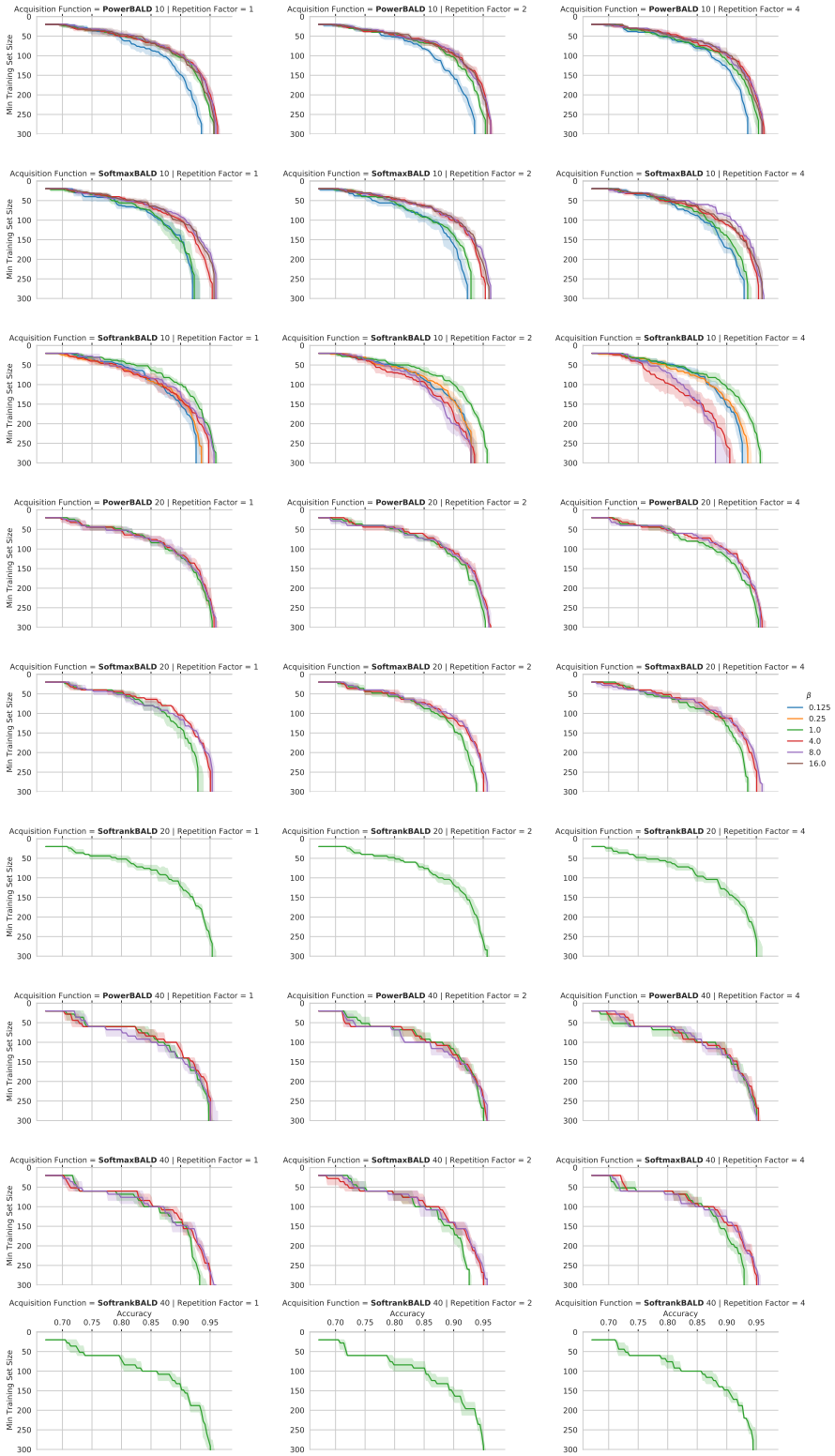


Figure 12. Test accuracy temperature ablation over Repeated-MNIST for different stochastic acquisition functions. Generally $\beta = 1$ works very well. For power and softmax acquisition $\beta = 8$ seems to work well across batch sizes and Repeated-MNIST repetition ranges.

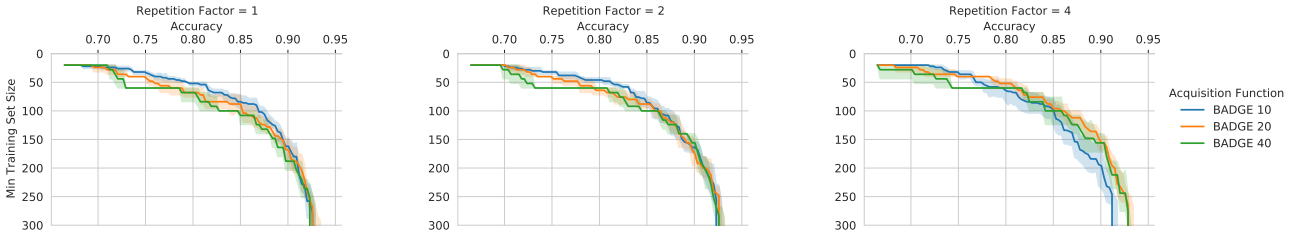


Figure 13. Test accuracy acquisition batch size ablation for BADGE. For Repeated-MNIST with 4 repetitions BADGE with acquisition batch size 20 performs best. Hence, we use that for Figure 2.

C.2. MIO-TCD and Synbols

C.2.1. Temperature Ablations

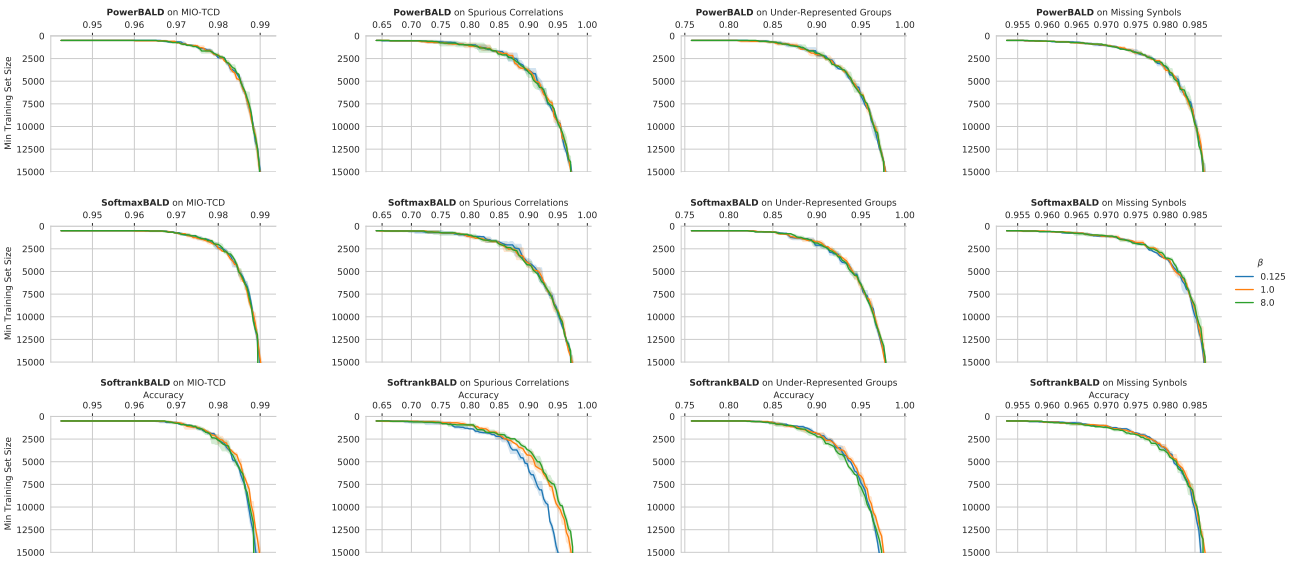


Figure 14. Accuracy temperature ablation for different stochastic acquisition functions using BALD on MIO-TCD and Synbols. $\beta = 1$ seems to be optimal almost everywhere, except SoftrankBALD on the spurious correlation dataset for which $\beta = 8$ is better. This is good because $\beta = 1$ is the default.

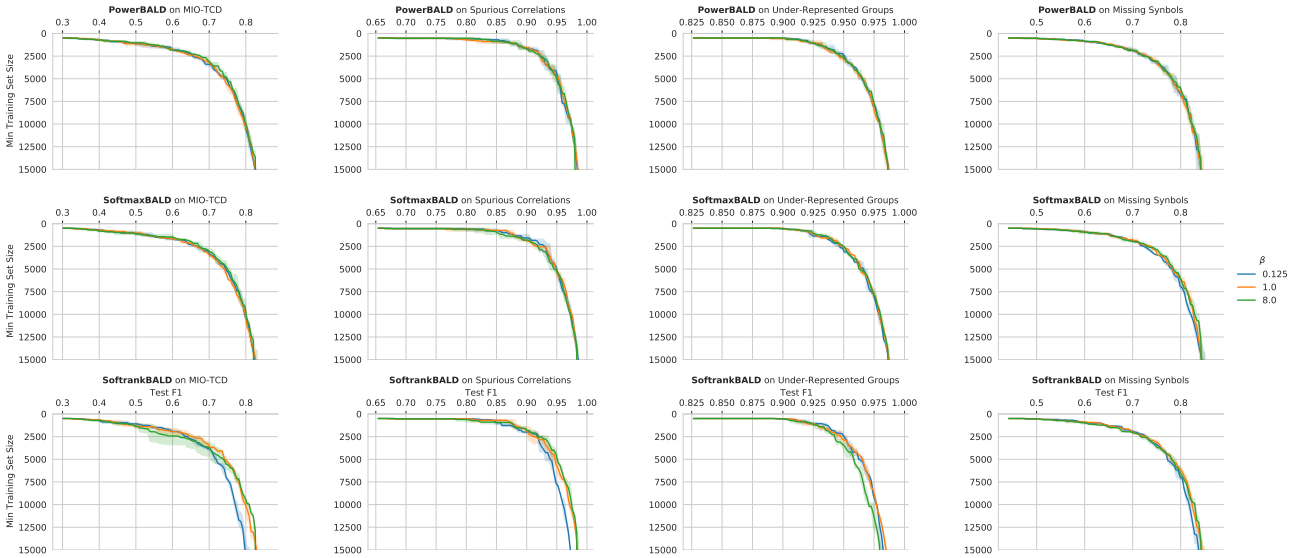


Figure 15. F1 score temperature ablation for different stochastic acquisition functions using BALD on MIO-TCD and Synbols. $\beta = 1$ seems to be optimal almost everywhere, except SoftrankBALD on the spurious correlation dataset for which $\beta = 8$ is better. This is good because $\beta = 1$ is the default.

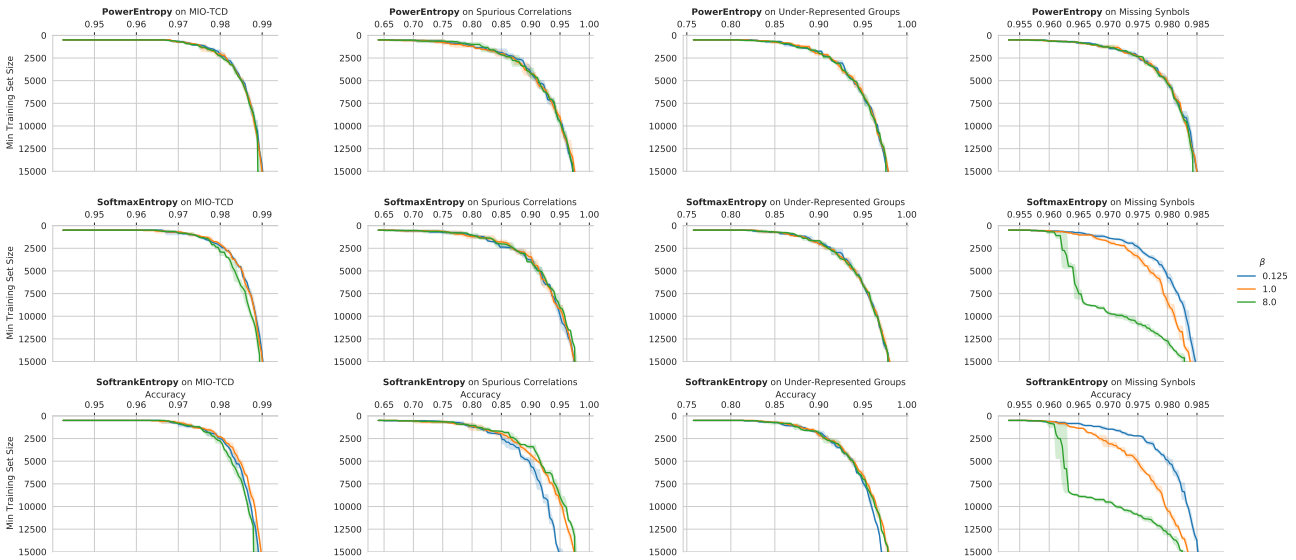


Figure 16. Accuracy temperature ablation for different stochastic acquisition functions using entropy on MIO-TCD and Synbols. $\beta = 1$ seems to be optimal almost everywhere, except SoftrankBALD on the spurious correlation dataset for which $\beta = 8$ is better and on the missing data variant of Synbols. Here $\beta = 8$ performs badly. This is good because $\beta = 1$ is the default and remains the best choice.

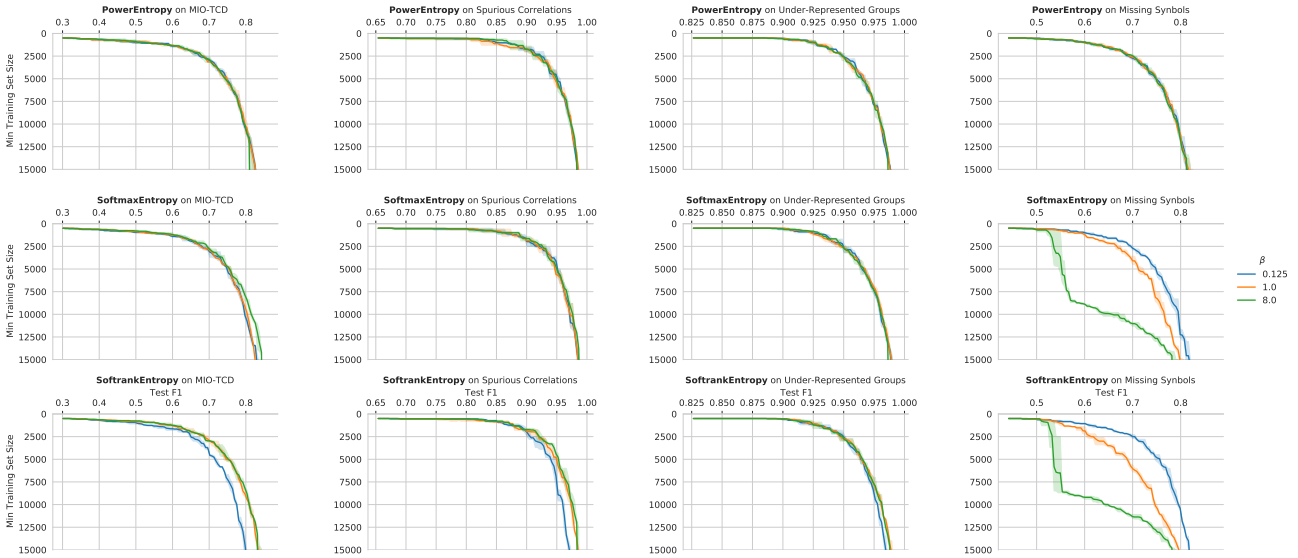


Figure 17. F1 score temperature ablation for different stochastic acquisition functions using entropy on MIO-TCD and Symbols. $\beta = 1$ seems to be optimal almost everywhere, except SoftrankBALD on the spurious correlation dataset for which $\beta = 8$ is better. This is good because $\beta = 1$ is the default.

C.2.2. Spurious correlation

In Figure 5, we noted that our stochastic methods were matches performance of BADGE and BALD. In Figure 18, we show the same performance.

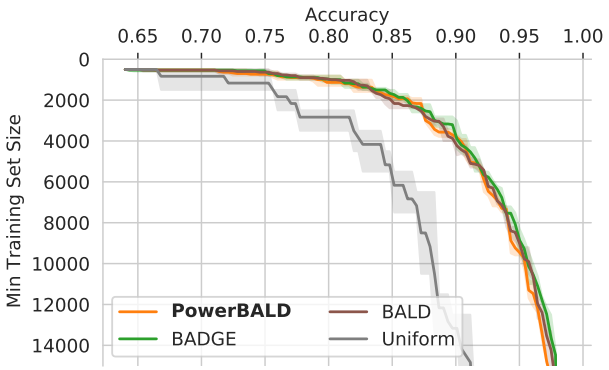


Figure 18. Test accuracy for BALD on Synbols Spurious Correlations. Averaged over 3 runs.

C.2.3. Entropy Baseline for MIO-TCD and Synbols

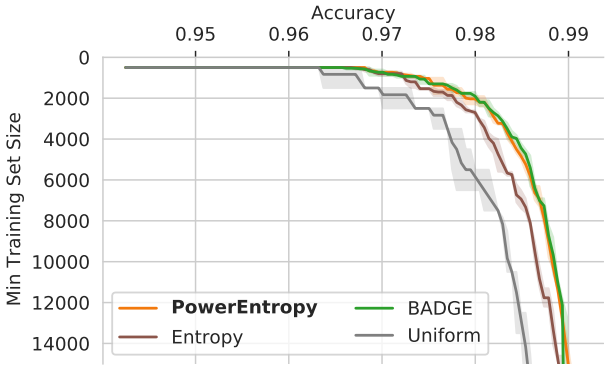


Figure 19. Test accuracy on MIO-TCD for entropy variants for $\beta = 1$.

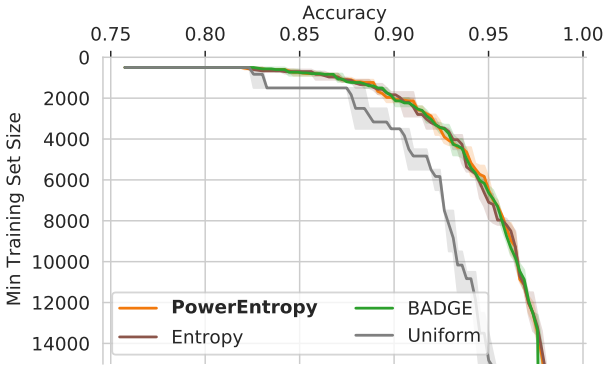


Figure 20. Test accuracy on Synbols Minority Groups for entropy variants for $\beta = 1$.

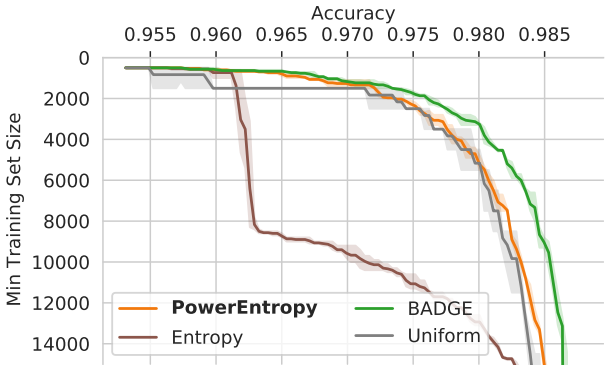


Figure 21. Test accuracy on Synbols Missing Characters for entropy variants for $\beta = 1$.

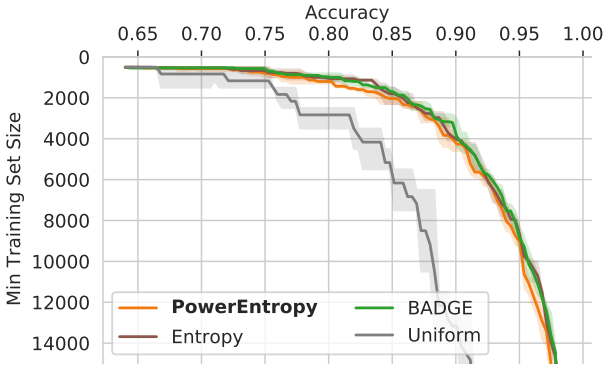


Figure 22. Test accuracy on Synbols Spurious Correlations for entropy variants for $\beta = 1$.

C.2.4. Predictive Parity for Synbols

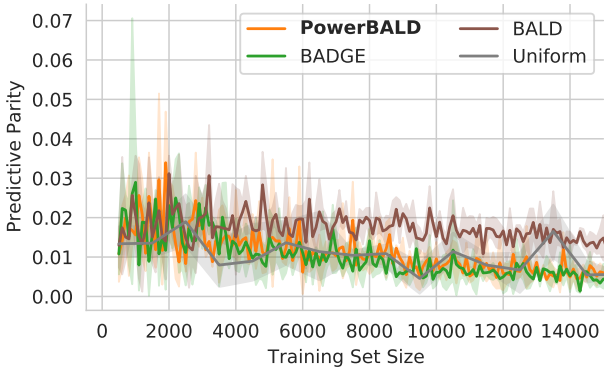


Figure 23. Test predictive parity on Synbols Minority Groups for BALD variants for $\beta = 1$.

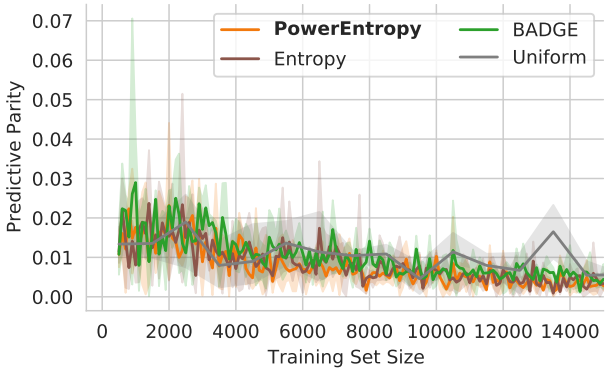


Figure 24. Test predictive parity on Synbols Minority Groups for entropy variants for $\beta = 1$.

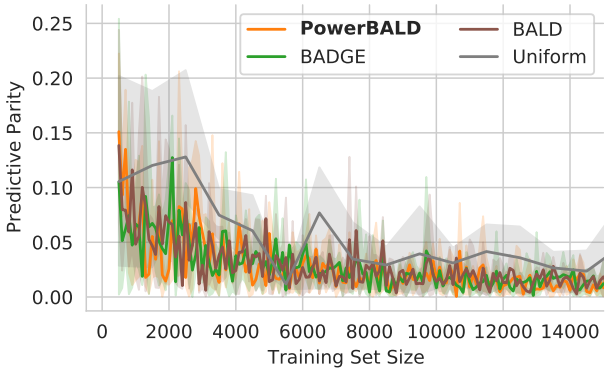


Figure 25. Test predictive parity on Synbols Spurious Correlations for BALD variants for $\beta = 1$.

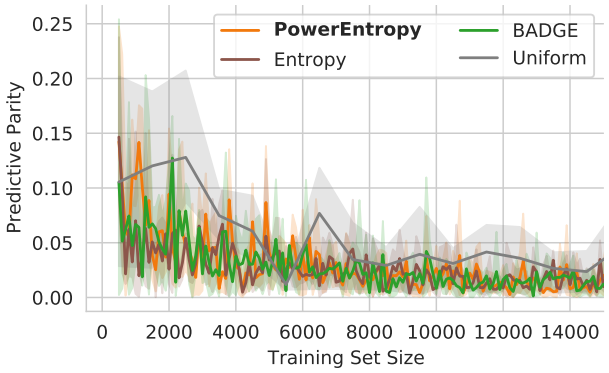


Figure 26. Test predictive parity on Synbols Spurious Correlations for entropy variants for $\beta = 1$.

C.2.5. F1 Scores for MIO-TCD

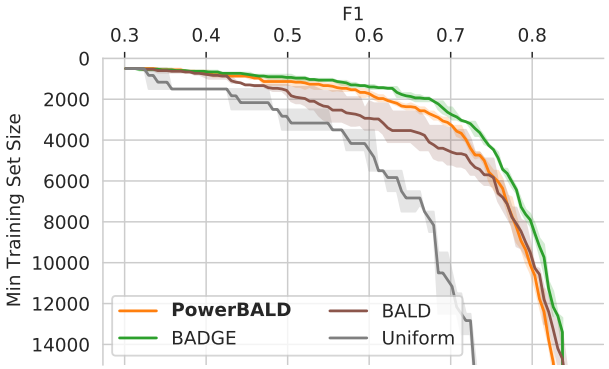


Figure 27. Test F1 score on MIO-TCD for BALD variants for $\beta = 1$.

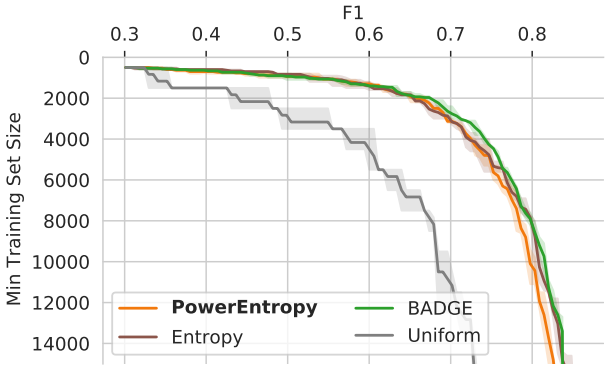


Figure 28. Test F1 score on MIO-TCD for entropy variants for $\beta = 1$.

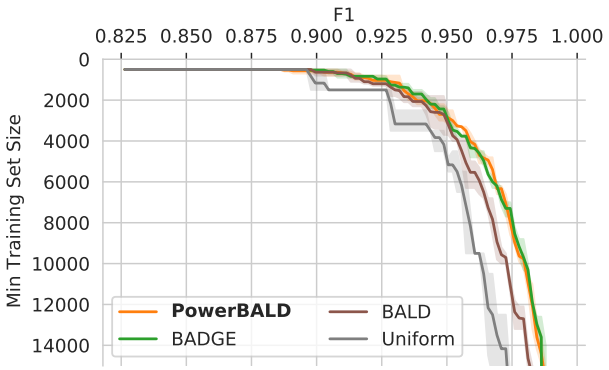


Figure 29. Test F1 score on Synbols Minority Groups for BALD variants for $\beta = 1$.

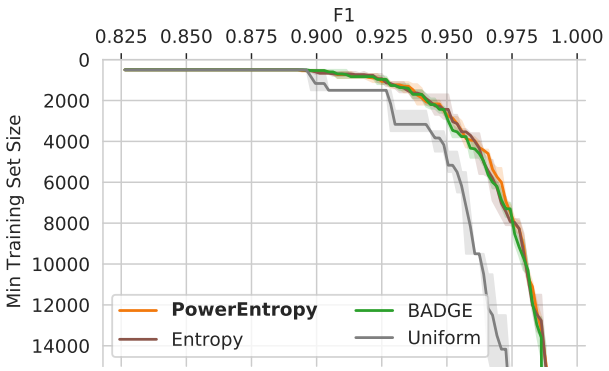


Figure 30. Test F1 score on Synbols Minority Groups for entropy variants for $\beta = 1$.

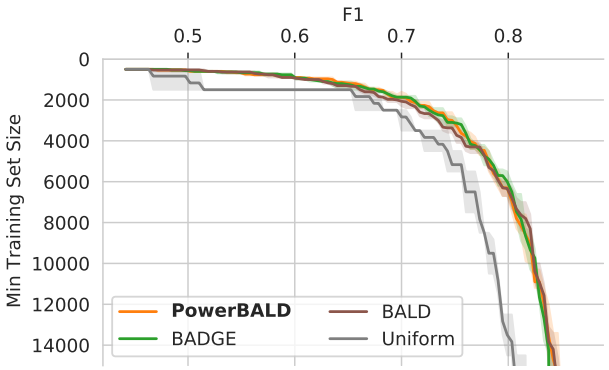


Figure 31. Test F1 score on Synbols Missing Characters for BALD variants for $\beta = 1$.

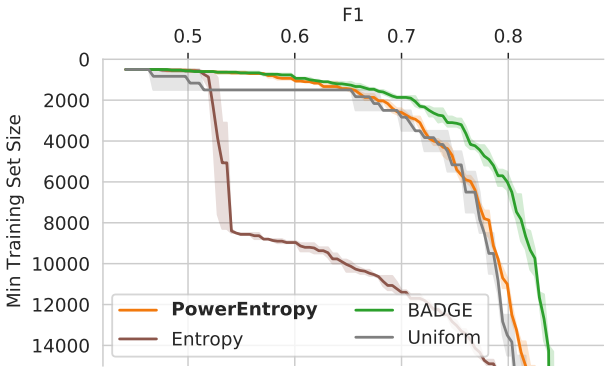


Figure 32. Test F1 score on Synbols Missing Characters for entropy variants for $\beta = 1$.

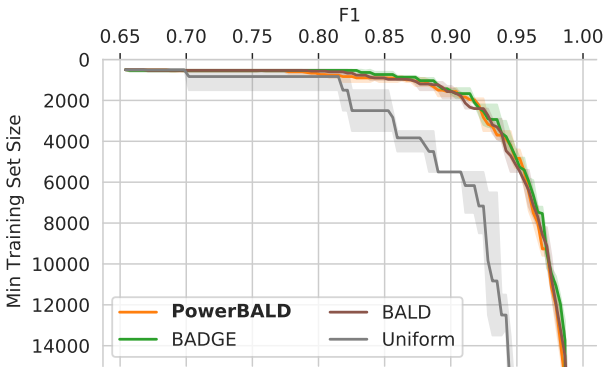


Figure 33. Test F1 score on Synbols Spurious Correlations for BALD variants for $\beta = 1$.

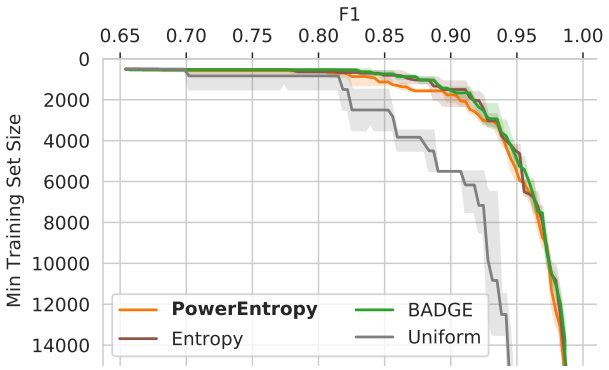


Figure 34. Test F1 score on Synbols Spurious Correlations for entropy variants for $\beta = 1$.

C.2.6. Comparison of Stochastic Acquisition Functions (BALD & entropy variants) on MIO-TCD and Synbols

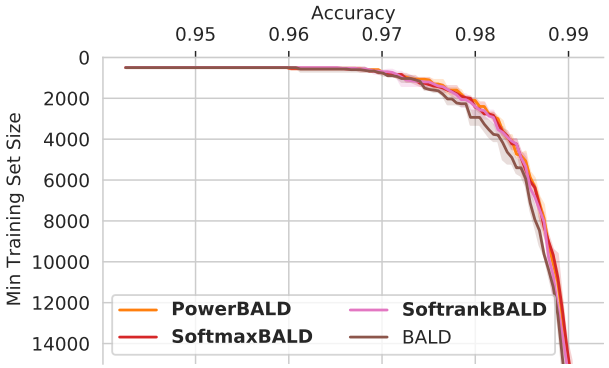


Figure 35. Test accuracy on MIO-TCD for BALD variants for $\beta = 1$.

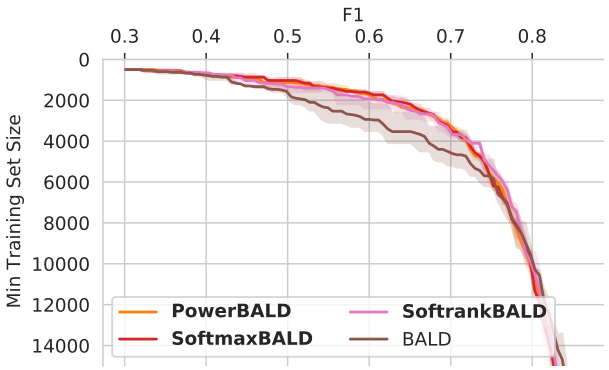


Figure 36. Test F1 score on MIO-TCD for BALD variants for $\beta = 1$.

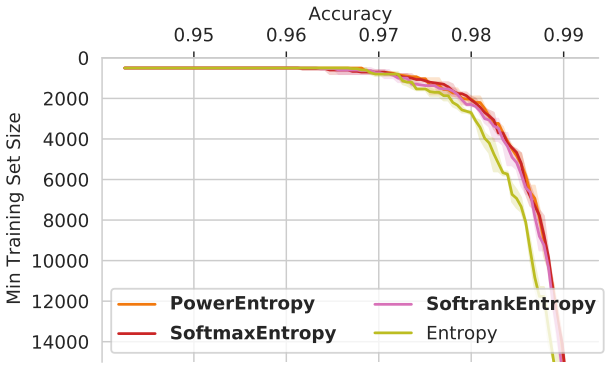


Figure 37. Test accuracy on MIO-TCD for entropy variants for $\beta = 1$.

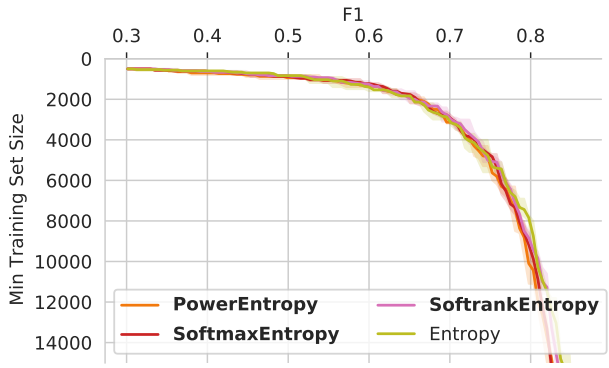


Figure 38. Test F1 score on MIO-TCD for entropy variants for $\beta = 1$.

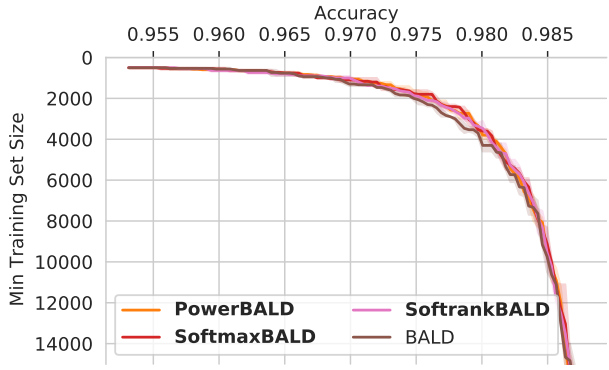


Figure 39. Test accuracy on Synbols Missing Characters for BALD variants for $\beta = 1$.

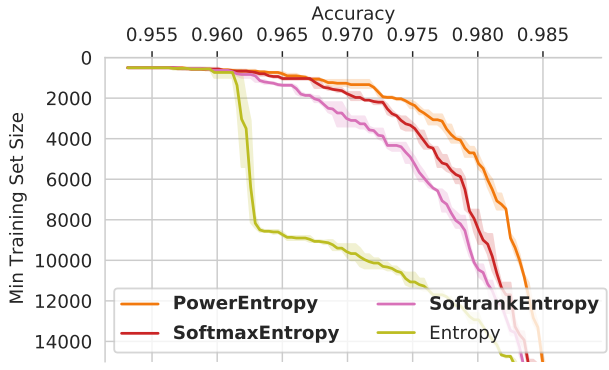


Figure 40. Test accuracy on Synbols Missing Characters for BALD variants for $\beta = 1$.

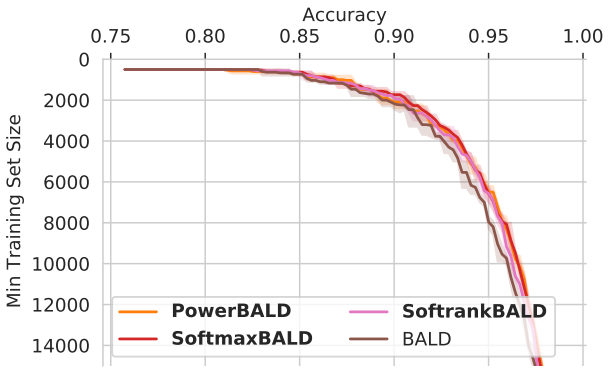


Figure 41. Test accuracy on Synbols Minority Groups for BALD variants for $\beta = 1$.

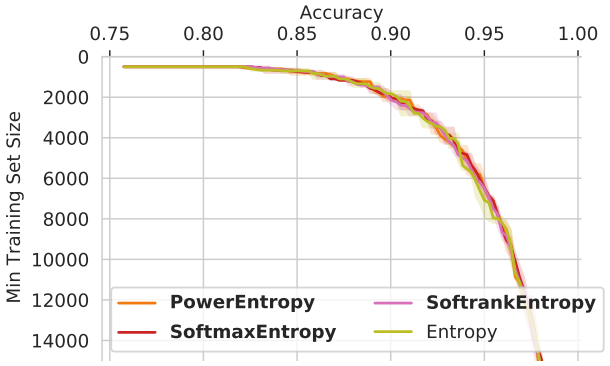


Figure 42. Test accuracy on Synbols Minority Groups for BALD variants for $\beta = 1$.

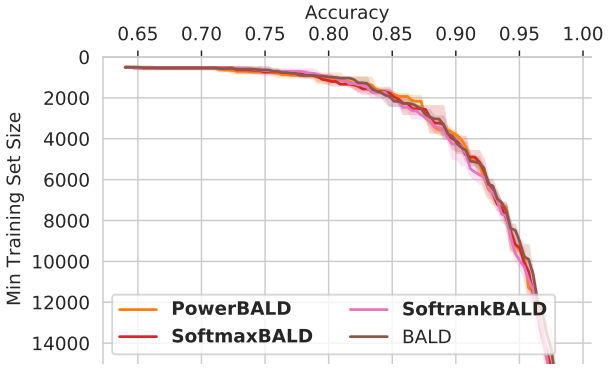


Figure 43. Test accuracy on Synbols Spurious Correlations for BALD variants for $\beta = 1$.

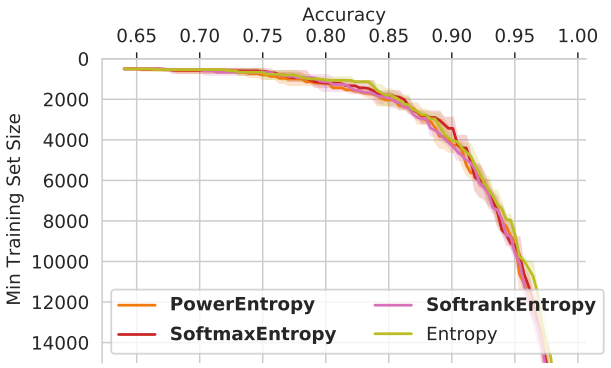
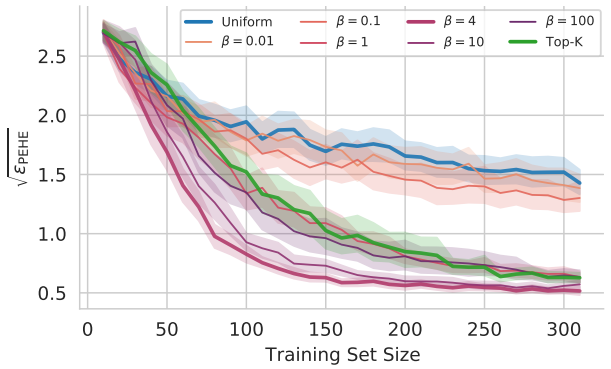


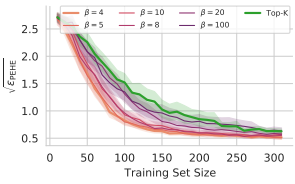
Figure 44. Test accuracy on Synbols Spurious Correlations for entropy variants for $\beta = 1$.

C.3. Further CausalBALD Ablations

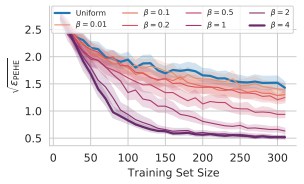
We provide further temperature investigation for CausalBALD on the entirely synthetic dataset which is used by Jesson et al. (2021). This demonstrates the ways in which the temperature can be chosen to interpolate between uniform and top-K acquisition.



(a) Overall Ablation (Subset)



(b) Low Temperature Only



(c) High Temperature Only

Figure 45. CausalBALD: Synthetic Dataset. (a) At a very high temperature ($\beta = 0.1$), PowerBALD behaves very much like random acquisition, and as the temperature decreases the performance of the acquistion function improves (lower $\sqrt{\epsilon_{\text{PEHE}}}$). (b) Eventually, the performance reaches an inflection point ($\beta = 5.0$) and any further decrease in temperature results in the acquisition strategy performing more like top-K. We see that under the optimal temperature, power acquisition significantly outperforms both random acquisition and top-K over a wide range of temperature settings.