Active Deep Probabilistic Subsampling

Active Deep Probabilistic Subsampling

Hans van Gorp [1] **Iris A.M. Huijben** [1] **Bastiaan S. Veeling** [2] **Nicola Pezzotti** [3 4] **Ruud J.G. van Sloun** [1 4]

Abstract

Subsampling a signal of interest can reduce costly data transfer, battery drain, radiation exposure and acquisition time in a wide range of problems. The recently proposed Deep Probabilistic Subsampling (DPS) method effectively integrates subsampling in an end-to-end deep learning model, but learns a static pattern for all datapoints. We generalize DPS to a sequential method that actively picks the next sample based on the information acquired so far; dubbed *Active-DPS* (A-DPS). We validate that A-DPS improves over DPS for MNIST classification at high subsampling rates. Moreover, we demonstrate strong performance in active acquisition Magnetic Resonance Image (MRI) reconstruction, outperforming DPS and other deep learning methods.

1. Introduction

Present-day technologies produce and consume vast amounts of data, which is typically acquired using an analog-to-digital converter (ADC). The amount of data digitized by an ADC is determined not only by the temporal sampling rate, but also by the manner in which spatial acquisitions are taken, e.g., by using a specific design of sensor arrays.

Reducing the number of sample acquisitions needed, can lead to meaningful reductions in scanning time, e.g., in Magnetic Resonance Imaging (MRI), radiation exposure, e.g., in Computed Tomography (CT), battery drain, and bandwidth requirements. While the Nyquist theorem is traditionally used to provide theoretical bounds on the sampling rate, in recent years signal reconstruction from sub-Nyquist sampled data has been achieved through a framework called Compressive Sensing (CS).

[1]Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands [2]Department of Computer Science, University of Amsterdam, Amsterdam, The Netherlands [3]Department of Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands [4]Philips Research, Eindhoven, The Netherlands. Correspondence to: Hans van Gorp <h.v.gorp@tue.nl>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

First proposed by Donoho (2006), and later applied for MRI by Lustig et al. (2007), CS leverages structural signal priors, specifically sparsity under some known transform. By taking compressive measurements followed by iterative optimization of a linear system under said sparsity prior, reconstruction of the original signal is possible while sampling at sub-Nyquist rates. Researchers have employed CS with great success in a wide variety of applications, such as radar (Baraniuk & Steeghs, 2007; Ender, 2010), seismic surveying (Herrmann et al., 2012), spectroscopy (Sanders et al., 2012), and medical imaging (Han et al., 2016; Lai et al., 2016).

However, both the need to know the sparsifying basis of the data, and the iterative nature of the reconstruction algorithms, still hamper practical applicability of CS in many situations. These limitations can be overcome by the use of deep learning reconstruction models that make the sparsity assumption implicit, and facilitate non-iterative inference once trained. Moreover, the (typically random) nature of the measurement matrix in CS does, despite adhering to the given assumptions, not necessarily result in an optimal measurement given the underlying data statistics and the downstream system task. This has recently been tackled by algorithms that learn the sampling scheme from a data distribution.

In general, these data-driven sampling algorithms can be divided into two categories: algorithms that learn sampling schemes which are fixed once learned (Huijben et al., 2020a;b;c; Ravishankar & Bresler, 2011; Sanchez et al., 2020; Bahadir et al., 2019; Bahadir et al., 2020; Weiss et al., 2019), and algorithms that learn to actively sample (Ji et al., 2008; Zhang et al., 2019; Jin et al., 2019; Pineda et al., 2020; Bakker et al., 2020); selecting new samples based on sequential acquisition of the information. The former type of algorithms learn a sampling scheme that - on average - selects informative samples of all instances originating from the training distribution. However, when this distribution is multi-modal, using one globally optimized sampling scheme, can easily be sub-optimal on instance-level.

Active acquisition algorithms deal with such shifts in underlying data statistics by conditioning sampling behavior on previously acquired information from the instance (e.g. the image to be sampled). This results in a sampling sequence

that varies across test instances, i.e. sampling is *adapted* to the new data. This adaptation as a result of conditioning, promises lower achievable sampling rates, or better downstream task performance for the same rate, compared to sampling schemes that operate equivalently on all data.

In this work, we extend the Deep Probabilistic Subsampling (DPS) framework (Huijben et al., 2020a) to an active acquisition framework by making the sampling procedure iterative and conditional on the samples already acquired, see Fig. 1. We refer to our method as Active Deep Probabilistic Subsampling (A-DPS). We show how A-DPS clearly exploits the ten different modalities (i.e. the digits) present in the MNIST dataset to adopts instance-adaptive sampling sequences. Moreover, we demonstrate both on MNIST (LeCun et al., 1998) and the real-world fast MRI knee dataset (Zbontar et al., 2018), that A-DPS outperforms other state-of-the-art models for learned sub-Nyquist sampling. Our code is publicly available.[1]

2. Related work

Recently, several techniques for learning a fixed sampling pattern have been proposed, especially in the field of MR imaging, in which Ravishankar & Bresler (2011) were one of the firsts. In this work, the authors make use of non-overlapping cells in k-space, and move samples between these cells.During training Ravishankar & Bresler (2011) alternate between reconstruction and relocation of sampling positions. After a reconstruction step they sort the cells in terms of reconstructing error and an infinite-p norm. Selected samples from lower scoring cells are relocated to higher scoring cells in a greedy fashion.

Sanchez et al. (2020) also propose a greedy approach, in which samples are not relocated between cells, but greedily chosen to optimize a reconstruction loss on a batch of examples. Both of the types of greedy optimization do however not allow for joint learning of sampling together with a downstream reconstruction/task model, as the reconstruction has to either be parameter-free or pretrained to work well with a variety of sampling schemes.

Bahadir et al. (2019) on the other hand propose to learn the sampling pattern by thresholding pixel-based i.i.d. samples drawn from a uniform distribution, dubbed Learning-based Optimization of the Under-sampling PattErn (LOUPE). The sample rate of LOUPE is indirectly controlled by promoting sparsity through the use of an ℓ_1 penalty on the thresholds.

One of the first active sampling schemes was proposed by Ji et al. (2008), who leverage CS reconstruction techniques that also give a measure of uncertainty of the reconstruction using Bayesian modeling. Ji et al. (2008) leveraged this uncertainty in the reconstruction to adaptivly select the next measurement that will reduce this uncertainty by the largest amount. However, this method - and other similar works from (Carson et al., 2012; Li et al., 2013) - rely on linearly combined measurements, rather than discrete sampling, with which we concern ourselves here.

In the field of MRI, Zhang et al. (2019) propose an active acquisition scheme by leveraging a reconstruction and adversarial neural network. Whereas the reconstruction network is trained to reconstruct MR images from the subsampled Fourier space (k-space), the adversarial network is trained to distinguish between already sampled, and omitted lines in this space. The k-space line that is most believed to be 'fake' (i.e. filled in by the reconstruction network) by the adversarial network, is sampled next. However, This framework only works for undersampled Fourier to image reconstruction tasks, as the discriminator requires mappings of the image in k-space. Jin et al. (2019) put forth an active acquisition scheme for MRI by leveraging reinforcement learning (RL). Two neural networks, one for sampling and one for reconstruction are trained jointly using a Monte-Carlo tree search, resulting in a sampling policy that is dependent on the current reconstruction of the image.

More recently, both Pineda et al. (2020) and Bakker et al. (2020) proposed RL-based active acquisition techniques. Pineda et al. (2020) leverages a Double Deep Q-Network. The model is trained using a modified ϵ-greedy policy, in which the best action is taken with probability $1-\epsilon$, and an exploratory action is taken with probability ϵ. Bakker et al. (2020) compare greedy with non-greedy training, finding that the greedy method leads to a higher degree of adaptability, especially for tasks with a long horizon (i.e. more samples to be taken). Both of the frameworks proposed by Pineda et al. (2020) and Bakker et al. (2020) make use of a pretrained reconstruction network, which differs from the proposed A-DPS method that enables joint training of both the reconstruction (task) network and sampling network.

Even though subsampling is an extreme form of data compression, we differentiate from typical data compression architectures like deep encoder-decoder structures (Theis et al., 2017; Ballé et al., 2017), as these methods do not reduce data rates at the measurement stage. The feedback recurrent autoencoder proposed by Yang et al. (2020) is however related to A-DPS through its use of a recurrent context. But whereas Yang et al. (2020) learn a context to inform the encoder stage of the network, A-DPS uses this to inform the sampling pattern.

[1] https://github.com/IamHuijben/Deep-Probabilistic-Subsampling

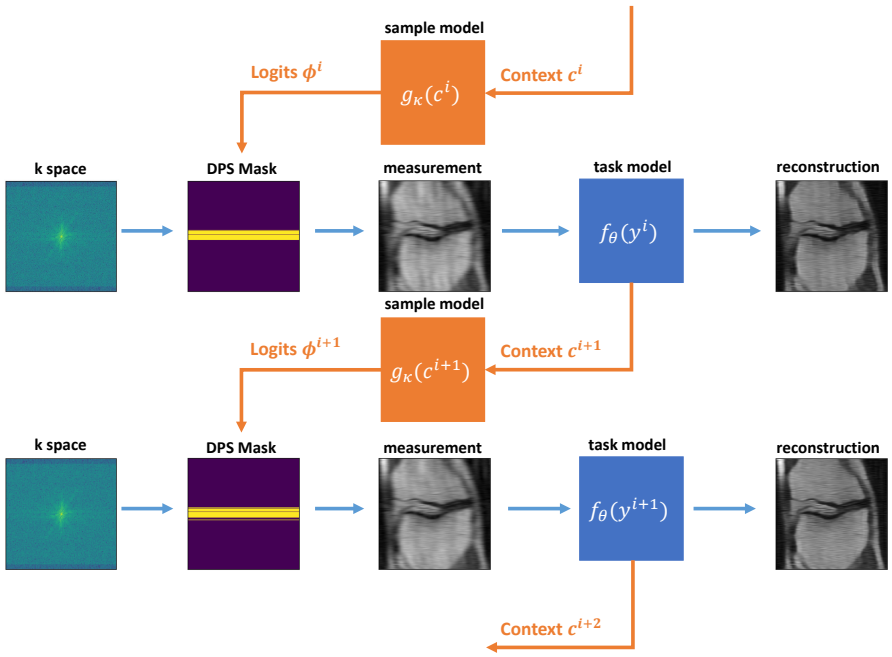

Figure 1. A-DPS learns to actively pick new samples in a sequential setup based on contextual information generated by the task model on previously acquired samples. Here we show an example of an MRI reconstruction task. Note however that A-DPS can be used for more tasks besides reconstruction, e.g., classification. The extension over DPS is shown in orange.

3. Method

3.1. General framework

Given a prediction task s we are interested in learning to predict an optimal subsampling scheme $\boldsymbol{A} \in \{0,1\}^{M\times N}$ (with $M \ll N$) on an input signal $\boldsymbol{x} \in R^N$, resulting in a measurement $\tilde{\boldsymbol{y}} \in R^M$:

$$\tilde{\boldsymbol{y}} = \boldsymbol{A}\boldsymbol{x}. \tag{1}$$

Each row in $\boldsymbol{A}$ is constrained to have ℓ_0-norm of 1, while each column in $\boldsymbol{A}$ is constrained to have an ℓ_0-norm of either 0 or 1, i.e. each of the N candidate samples is selected at most once. In the rest of this paper we will index these candidate samples with $n \in \{1,\dots,N\}$, and the selected samples with $m \in \{1,\dots,M\}$. The percentage of selected samples from the candidate samples is called the sampling ratio $r = M/N \cdot 100\%$.

We also introduce a non-compressed form of the measurement $\tilde{\boldsymbol{y}}$, called $\boldsymbol{y} \in R^N$, that contains $N-M$ zeros, and M non-zeros at the sampled indices specified by $\boldsymbol{A}$, i.e., the *masked* input. This way, the location of samples from $\boldsymbol{x}$ is preserved, which is especially useful when $\boldsymbol{A}$ changes during training. To acquire $\boldsymbol{y}$ from $\boldsymbol{x}$, one seeks a subsampling mask $\boldsymbol{d}$ that can be applied on $\boldsymbol{x}$ via:

$$\boldsymbol{y} = \boldsymbol{d} \quad \boldsymbol{x} = \boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}, \tag{2}$$

where denotes an element-wise multiplication. From the resulting measurement $\boldsymbol{y}$ we then aim at predicting the downstream task s through:

$$\hat{s} = f_\theta(\boldsymbol{y}), \tag{3}$$

where $f_\theta(.)$ is a function that is differentiable with respect to its input and parameters θ, e.g., a neural network. Normally, optimization of the task model $f_\theta(.)$ is achieved through backpropagation of some loss function $\mathcal{L}(s,\hat{s})$. However, calculating gradients on the sampling matrix is blocked by its combinatorial nature, inhibiting joint training of the task with the sampling operation. The DPS framework provides a solution to this problem, on which we will elaborate in the next section.

3.2. DPS: Deep Probabilistic Subsampling

To enable joint training of the sampling operation with the downstream task model, Huijben et al. (2020a) introduce DPS. Rather than optimizing $\boldsymbol{A}$ directly, they propose to optimize a generative sampling model $P(\boldsymbol{A}|\boldsymbol{\phi})$, where $\boldsymbol{\phi}$ are learned unnormalized logits of (possibly multiple) categorical distribution(s). Each distribution expresses the probabilities for sampling any of the elements x_n from $\boldsymbol{x}$ through sampling matrix $\boldsymbol{A}$. More specifically, $\phi_{m,n}$ is the log-probability for setting $a_{m,n} = 1$, and thus sampling x_n as m^{th} sample.

To generate a sampling pattern from these unnormalized logits, i.e. implementation of this conditional model, the Gumbel-max trick is leveraged (Gumbel, 1954). In the

Gumbel-max trick the unnormalized logits are perturbed with i.i.d. Gumbel noise samples $e_{m,n} \sim \text{Gumbel}(0, 1)$. By selecting the maximum of this perturbation a realization of the sampling mask can be found using:

$$\boldsymbol{A}_{m,:} = \text{one-hot}_N \left\{ \underset{n}{\text{argmax}} \left\{ w_{m-1,n} + \phi_{m,n} + e_{m,n} \right\} \right\}, \tag{4}$$

where $\boldsymbol{A}_{m,:}$ denotes the m-th row of $\boldsymbol{A}$ and one-hot$_N$ creates a one-hot vector of length N, with the one at the index specified by the argmax operator. Moreover, the cumulative mask $w_{m-1,n} \in \{-\infty, 0\}$ masks previously selected samples by adding minus infinity to those logits, thereby ensuring sampling without replacement.

During backpropagation, gradients are computed by relaxing this sampling procedure using the Gumbel-softmax trick (Jang et al., 2016; Maddison et al., 2017), resulting in:

$$\begin{aligned} &\nabla_{\phi_m} \boldsymbol{A}_{m,:} := \\ &\nabla_{\phi_m} E_{e_m} \left[\text{softmax}_\tau \left\{ w_{m-1,n} + \phi_{m,n} + e_{m,n} \right\} \right], \end{aligned} \tag{5}$$

where τ denotes the temperature parameter of the softmax operator. Setting $\tau > 0$ results in a smoothed sampling matrix $\boldsymbol{A}$ (i.e. elements can have values between 0 and 1 as well), allowing gradients to distribute over multiple logits during training. In the limit of $\tau \to 0$ the softmax operator approaches the one-hot argmax function of equation (4). Although this approach – also known as straight-through Gumbel-softmax – leads to biased gradients, it has been shown to work well in practice, and Huijben et al. (2020a) keep τ at a fixed value during training.

Huijben et al. (2020a) propose two regimes of DPS. First, Top-1 sampling, an expressive form of DPS where each of the M selected samples are separately conditioned on all N candidate samples, resulting in $M \times N$ trainable logits $\phi_{m,n}$. Second, Top-M sampling (called Top-K in their paper), a constrained form where all M samples together are conditioned on all N candidate samples, i.e. the logits ϕ_n are shared between the M rows of $\boldsymbol{A}$, resulting in only N trainable logits. While Top-1 sampling is more expressive, Huijben et al. (2020a) noticed slightly better results for the Top-M regime, possibly thanks to the smaller number of trainable logits, therefore facilitating optimization. For scaleability reasons, we thus choose to continue with Top-M sampling in this work and refer to this regime as DPS in the rest of this paper. We refer the reader to Huijben et al. (2020a) for more details regarding DPS.

3.3. A-DPS: Active Deep Probabilistic Subsampling

We have seen how DPS enables the learning of a sampling scheme that selects M out of N samples. However, these samples are selected simultaneously. A-DPS selects its samples in an iterative fashion, separating the logits into I acquisition steps, i.e. $\boldsymbol{\phi}^i$ with $i \in \{0, 1, 2, \ldots, I-1\}$ and $I = M$.

Active acquisition is then achieved by introducing dependency between samples, i.e. the sampling distribution at acquisition step i should depend on the information acquired in previous acquisition steps. To that end, we introduce a context vector $\boldsymbol{c_i}$, that encodes information about the current task. We then condition the sampling distribution on this context by learning a transformation $\boldsymbol{\phi} = g_\kappa(\boldsymbol{c})$, where $g_\kappa(.)$ is a function that is differentiable with respect to its input and parameters κ. Thus, instead of optimizing the parameters directly (as DPS does), we optimize $g_\kappa(\boldsymbol{c})$, which we will refer to as the sampling model.

The question then arises how to best generate this context from previous samples. Here, we follow the *analysis-by-synthesis* principle, and let the analysis (the sampling model) depend on the synthesis (the task model). This way, the task model can inform the sampling model what information it needs to achieve its assigned task. The iterative *analysis-by-synthesis* scheme of A-DPS is formalized as follows:

Algorithm 1 A-DPS

Input: acquisition steps I
Data: input signal $\boldsymbol{x}$ and associated task s
$\boldsymbol{c}^0, \boldsymbol{d}, l, i = 0$
while $i < I$ **do**
  $\boldsymbol{\phi}^i = g_{\boldsymbol{\kappa}}(\boldsymbol{c}^i)$
  $\boldsymbol{d}$ += DPS$(\boldsymbol{\phi}^i)$
  $\boldsymbol{y}^i = \boldsymbol{d} \quad \boldsymbol{x}$
  $\hat{s}^i, \boldsymbol{c}^{i+1} = f_{\boldsymbol{\theta}}(\boldsymbol{y}^i)$
  l += $\mathcal{L}(\hat{s}^i, s)$
  i += 1
end

Where DPS() signifies the operation to create a sampling mask from logits as described in section 3.2. By accumulating the loss over all acquisition steps we train in a semi-greedy fashion, which promotes the network to select more interesting samples early on. We visualize the architecture of the A-DPS framework in Fig.1 and discuss its computational complexity in the Appendix.

4. Experiments

To show the applicability of A-DPS on both classification as well as reconstruction tasks we evaluate its performance in two experiments. First, we will compare A-DPS with DPS at different subsampling ratios on an MNIST classification example in Section 4.1. Second, we will compare A-DPS with contemporary CS and deep learning methods on an MRI example in sections 4.2 and 4.3, leveraging the fast MRI knee dataset (Zbontar et al., 2018).

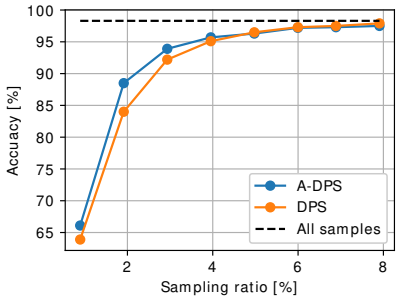

Figure 2. A-DPS outperforms DPS on classification accuracy for MNIST when the sampling ratio is less than 5%.The 'all samples' line indicates the accuracy achieved without subsampling. Both DPS and A-DPS approach this upper bound.

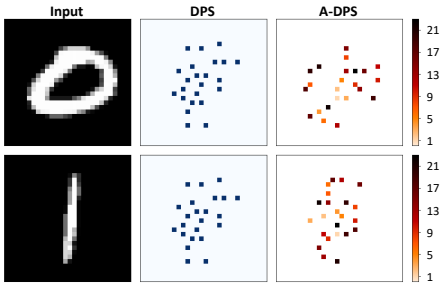

Figure 3. A-DPS uses different sampling patterns for different digits in classification, while DPS uses only one static sampling patterns across the entire dataset. The color scale indicates the order in which samples are taken. The sampling ratio is 2%, resulting in 15 samples.

4.1. MNIST

Experiment setup Classification performance at different sampling rates was tested on the MNIST database (LeCun et al., 1998), consisting of 70,000 grayscale images of 28×28 pixels of handwritten digits between 0 and 9. We split the original 60,000 training images into 50,000 training and 10,000 validation images. We keep the original 10,000 testing examples. We train both DPS top-M and A-DPS to take partial measurements in the pixel-domain at different sampling rates.

Reaching back to Fig.1 and algorithm 1, DPS top-M sampling only consists of the DPS sampling and task model ($f_{\boldsymbol{\theta}}(.)$). All M samples are selected at the same time and used once by $f_{\boldsymbol{\theta}}(.)$ to predict which digit the network is looking at. In the case of A-DPS however, only 1 sample is taken at a time and used as input for $f_{\boldsymbol{\theta}}(.)$. Here, $f_{\boldsymbol{\theta}}(.)$ also creates a context that is used by the sampling network $g_{\boldsymbol{\kappa}}(.)$ to select the next sample. A-DPS iterates through this loop M times in order to select all the samples. We keep $f_{\boldsymbol{\theta}}(.)$ the same for both DPS and A-DPS. Resulting in the fact that the last iteration of A-DPS is similar to that of DPS top-M (i.e. M samples are selected and fed through $f_{\boldsymbol{\theta}}(.)$).

Task model In the classification network $f_{\boldsymbol{\theta}}(.)$ all 784 (28×28) zero-masked samples are used as input for 5 fully connected layers. The fully connected layers have 784, 256, 128, 128, and 10 nodes, respectively. Moreover, all but the last layers are activated by leaky ReLU activation functions with a negative slope of 0.2. The last layer uses a softmax activation function to output class label probabilities. The first three layers also have a dropout of 30%.

The output vector of the fourth layer is used as the context vector for A-DPS. The sampling network $g_{\boldsymbol{\kappa}}(.)$ consists of an LSTM with a hidden size of 128, followed by two linear layers with output sizes of 256 and 784, respectively. Moreover, after the first layer a leaky ReLU activation function is used with a negative slope of 0.2, and a dropout of 30% is applied. The last layer is not followed by any activation function as its output are the unnormalized logits $\boldsymbol{\phi}_i$ used to create the next sampling mask.

Training details Both sampling strategies were trained to minimize the categorical cross-entropy loss. The temperature parameter was fixed to 2. We employ SGD with the Adam solver (Kingma & Ba, 2015) (lr = $2e-4$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1e-7$) to minimize the loss function. Training was performed on batches of 256 examples for 100 epochs.

Results The resulting accuracy on the test set is shown in Fig. 2. A-DPS outperforms DPS especially when the sampling ratio is less than 5%. It is hypothesized that it is especially important to select those candidate samples that carry a lot of information based on the previous samples for very low data rates. In Fig. 2 we also show the upper bound on accuracy our task model can achieve without any subsampling, both DPS and A-DPS approach this limit very quickly. Two examples of the selected sampling masks at $r = 2\%$ are displayed in Fig. 3. Here, it is shown how DPS selects all samples at once, while A-DPS selects them in an iterative fashion, resulting in different sampling patterns for the two examples.

To analyze the sampling patterns across the entire test set we plot all of the patterns together in Fig. 4 for a sampling ratio of 3%. Here we show the relative chance to sample a pixel at each acquisition step. The same candidate sample (index 489 in this case) is always sampled first, as the context is zero there for all examples. After the first step the sampling patterns diverge with a preference for candidate samples near the center of the image.

We also employ t-SNE (Van Der Maaten & Hinton, 2008) to see if we can observe clustering in the sample patterns

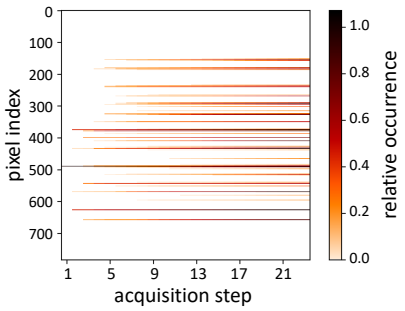

Figure 4. Visualization of all sampling patterns across the test set for a sampling ratio of 3%. The color scale indicates the relative occurrence of the sample at the current acquisition step.

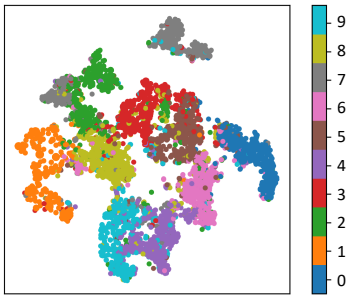

Figure 5. Sampling patterns selected by A-DPS are clustered per digit when applying t-SNE analysis on them. The color scale indicates the ground truth label of the corresponding image. This image was made using a sampling ratio of 3%.

generated by A-DPS on this task. t-SNE maps the multi-dimensional sampling patterns to points in 2D space. In this 2D space t-SNE aims to preserve spatial relationships from the higher dimension, i.e. similar high dimensional vectors get mapped close to together, while dissimilar ones are mapped further apart. The resulting plot is shown in Fig. 5, where each dot is colored with the ground truth label (digit) of the corresponding image. The clustering in this figure indicates how similar selected sampling patterns are. For example, the sampling patterns for digit zero and one tend to be dissimilar from one another. Interestingly, the digit seven seems to have two sampling patterns associated with it, one dissimilar from all others, while the other one is close to that of digit two.

4.2. MRI with (learned) fixed baselines

Experiment setup To show the applicability of A-DPS, we demonstrate its performance on line-based MRI. We make use of the NYU fastMRI database of knee MRI volumes (Zbontar et al., 2018). Only the single-coil measurements were selected, from which the outer slices were removed. The resulting data was split into 8,000 training, 2,000 validation, and 3,000 testing MRI slices. All slices were cropped to the central 208×208 pixels and normalized between 0 and 1. The subsampling operation on one of these MRI slices is then performed in k-space (Fourier-space):

$$\boldsymbol{Y} = |\mathcal{F}^H \boldsymbol{D} \quad \mathcal{F}\boldsymbol{X}|, \tag{6}$$

where $|.|$ is the magnitude operator. Moreover, $\boldsymbol{X} \in R^{N \times N}$ is the fully sampled ground truth image and $\boldsymbol{Y} \in R^{N \times N}$ is the subsampled image, both in the pixel domain. In this case N is equal to 208. Furthermore, $\mathcal{F}$ and $\mathcal{F}^H$ denote the forward and inverse 2D-Fourier transform, respectively. $\boldsymbol{D} \in \{0,1\}^{N \times N}$ denotes the sampling mask in k-space.

Normally $\boldsymbol{Y}$ would be complex, due to the asymmetrical nature of MRI measurements and the incomplete subsampling mask. Here, we choose to take the magnitude of Y to simplify reconstruction. We hypothesize that doing so does not significantly change the problem, as the imaginary part of fully sampled images in the NYU fastMRI dataset is very small compared to the real part.

Task model To reconstruct an estimate of the original image $\hat{\boldsymbol{X}}$ from the partial measurement $\boldsymbol{Y}$ a deep unfolded proximal gradient method is used (Mardani et al., 2018), in which K iterations of a proximal gradient method are unfolded as a feed forward neural network following:

$$\begin{aligned} &\hat{\boldsymbol{X}}^{(k+1)} = \\ &\mathcal{P}^{(k)}_{(\zeta)} \left\{ \hat{\boldsymbol{X}}^{(k)} - \alpha^{(k)}_{(\psi)} \left(|\mathcal{F}^H \boldsymbol{D} \quad \mathcal{F}\hat{\boldsymbol{X}}^{(k)}| - \boldsymbol{Y} \right) \right\}, \end{aligned} \tag{7}$$

where $\mathcal{P}^{(k)}_{(\zeta)}(.)$ is a trainable image-to-image proximal mapping and $\alpha^{(k)}_{(\psi)}$ is the step size, parameterized by ζ and ψ, respectively. We implement this proximal gradient method for $k = 3$ steps, with the trainable step size $\alpha^{(k)}_{(\psi)}$ implemented as a 3×3 convolutional layer. Each proximal mapping is implemented as a series of 4 convolutions with 16, 16, 16, and 1 feature(s) each and a kernel size of 3×3. All convolutions but the last are followed by ReLU activation functions.

We will compare A-DPS to to several relevant sampling baselines, namely, random uniform, low-pass, variable density (VDS), greedy mask selection (Sanchez et al., 2020), LOUPE (Bahadir et al., 2019; Bahadir et al., 2020), and DPS. We compare A-DPS to the active baselines of Zhang et al. (2019) and Pineda et al. (2020) in section 4.3.

Under a random uniform regime all N lines are equally likely to be sampled, while under a low-pass regime the M lines closest to the DC frequency will be selected. VDS on the other hand is a heuristic regime that employs a probability density from which the desired amount of samples are drawn. Following (Lustig et al., 2007), we here use a polynomial probability density function with a decay factor of 6.

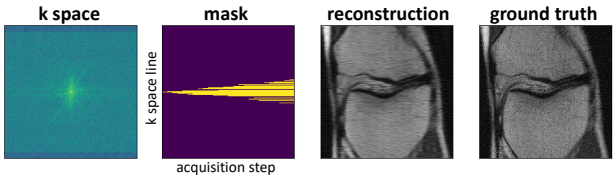

Figure 6. A-DPS MR image reconstruction of a test-set image by adaptively sampling 26 k-space lines. From left to right: 1) k-space, 2) sequence of line indices selected by A-DPS, 3) reconstructed image, 4) fully sampled MR image (ground truth).

For the greedy mask selection we follow the approach by Sanchez et al. (2020) and first optimize the sampling mask using the NESTA solver (Becker et al., 2011). After this, we fix the sampling mask and train our proximal gradient network. Results for both reconstruction algorithms are reported.

To generate a sampling mask $\boldsymbol{D}$ using A-DPS we use a sampling network $g_{\boldsymbol{\kappa}}(.)$. As context the sampling network takes the current reconstruction as input. This image is analyzed using 3 convolutional layers with kernels sizes of 3×3 followed by ReLU activation functions. The output features are of sizes 16, 32, and 64, respectively. The final feature map is aggregated into a feature vector using global average pooling. This feature vector is then fed into an LSTM cell of size 64. The output of the LSTM is transformed by a fully connected layer to the logits of size 208 used to create the sampling mask for the next acquisition step.

Training details To promote the reconstruction of visually plausible images, we leverage both a Mean Squared Error (MSE) and adversarial loss (Ledig et al., 2016). To that end we introduce a discriminator network that is trained to distinguish between real and reconstructed MR images. The discriminator is implemented using three convolutional layers with kernel sizes of 3×3, stride 2, and 64 feature maps, each with Leaky ReLU activations. After the last convolutional layer the feature maps are aggregated into a feature vector using global average pooling, with a dropout rate of 40%, which is mapped to a single output probability using one fully connected layer followed by a sigmoid activation function. Next to the MSE loss and adversarial loss, we add a third loss term that penalizes the MSE loss between the discriminator features of real and generated images. The total loss function is a weighted summation of these three losses, with weights 1, $5e-6$, and $1e-7$, respectively.

All sampling mask selection strategies were then trained using SGD on batches of 2 images for a total of 10 epochs. We again employ the Adam solver (lr = $2e-4$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1e-7$) to minimize the loss function, and set the temperature parameter to 2. We choose $M = 26$, which results in an acceleration factor of 8 ($r = 12.5\%$).

Table 1. Average results over 5 runs on the hold-out test set of size 208×208 for an acceleration factor of 8 compared to the (learned) fixed baselines.

Sampling Model	NMSE	PSNR	SSIM
Random uniform	0.4645	17.7	0.268
Low pass	0.0474	24.4	0.507
Variable density	0.0450	24.6	0.540
Greedy Mask NESTA	0.0425	23.3	0.494
Greedy Mask Prox. Grad.	0.0442	24.9	0.530
LOUPE	0.0476	25.0	0.567
DPS	0.0401	25.3	0.568
A-DPS (proposed)	**0.0389**	**25.5**	**0.582**

Results We score the different strategies based on 3 metrics: the normalized mean square error (NMSE), the peak signal-to-noise ratio (PSNR), and the structural similarity index (SSIM) (Wang et al., 2004). The averaged results over 5 runs on the hold-out test set for an acceleration factor of 8 are shown in Table 1. A-DPS outperforms all other baselines on the three metrics. An example of an A-DPS reconstruction is shown in Fig. 6, while a comprehensive overview of all baselines for this example can be found in the Appendix.

4.3. MRI with active baselines

Training details We also compare A-DPS with the models created by Zhang et al. (2019) and Pineda et al. (2020) using the implementations and checkpoints provided by Pineda et al. (2020)[2]. We here compare using their "Evaluator" and "DS-DDQN" checkpoints. Moreover, different preprocessing is used. No cropping or removal of the outer slices is applied. The input size of the k space is however cropped (where necessary) to a size of 368×640. Moreover, reconstructions are only scored on the central 320×320, as outside of that range there is mostly background. We use the same validation-testing split as the implementation of Pineda et al. (2020), resulting in a total of 34,742 training, 1,785 validation, and 1,851 testing images.

We use the exact same DPS and A-DPS models as in the previous experiment, with only the minor change in output size of the sampling model $g_{\boldsymbol{\kappa}}(.)$ to account for the larger k space. Due to computational constraints training is performed for 5 epochs on batches of size 1. We again employ the same loss function and Adam solver with lr = $2e-4$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1e-7$.

Results We again compare the models for an acceleration factor of 8. Note that the amount of samples taken is

[2] https://github.com/facebookresearch/active-mri-acquisition

Table 2. Results on the hold-out test set of size 368×640 for an acceleration factor of 8 compared to the active baselines.

Sampling Model	NMSE	PSNR	SSIM
(Zhang et al., 2019)	0.0398	28.8	0.610
(Pineda et al., 2020)	0.0371	29.2	0.623
DPS	0.0360	30.1	0.650
A-DPS (proposed)	**0.0342**	**30.2**	**0.654**

different in that case. Whereas in the previous experiment 26 lines are sampled, now that number has jumped up to 46 (to account for the larger images used). We compare A-DPS and DPS with the baselines in the scenario-30L, which means that we always sample the 30 lines closest to DC. Resulting only in 16 candidate samples that the models need to choose. The results of this comparison are shown in Table 2.

As can be seen from Table 2, A-DPS outperforms all other baselines. Important to note here is that DPS and A-DPS use a different reconstruction network when compared to the baselines. A-DPS and DPS make use of the proximal gradient network specified in section 4.2. It is jointly trained with the sampling model, and has 93,919 parameters. The baselines on the other hand make use of an encoder-decoder resnets as proximal operator, also followed by data consistency, and has 294,180,864 parameters. It cannot be jointly trained with the sampling model, but insead needs to be trained seperatly on random masks, a clear drawback of these baselines.

5. Conclusion

We proposed a generalization of DPS, which enables active acquisition, called A-DPS. We demonstrated its applicability on both an MNIST classification task as well as an MRI reconstruction task. Moreover, we found that the adaptive nature of A-DPS improves performance over other sampling pattern selection methods on downstream task performance. We find that A-DPS uses qualitatively differing sampling strategies depending on the context in the MNIST experiment. On a critical note, the black-box nature of A-DPS comes with the traditional machine learning challenges of out-of-distribution generalization and overfitting. This means that in a practical application, the sub-sampling regime could obfuscate the information required to recognize failure cases.

Future work includes exploration on how to improve conditioning of the sampling scheme on earlier acquired information and meta-information (such as resolution, sampling ratio, and weighting). Potential future applications include 3D and dynamic MRI, CT, ultrasound, radar, video, and MIMO systems.

References

Bahadir, C. D., Dalca, A. V., and Sabuncu, M. R. Learning-based Optimization of the Under-sampling Pattern in MRI. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11492 LNCS:780–792, 1 2019.

Bahadir, C. D., Wang, A. Q., Dalca, A. V., and Sabuncu, M. R. Deep-learning-based optimization of the under-sampling pattern in mri. *IEEE Transactions on Computational Imaging*, 6:1139–1152, 2020.

Bakker, T., van Hoof, H., and Welling, M. Experimental design for mri by greedy policy search, 2020.

Ballé, J., Laparra, V., and Simoncelli, E. P. End-to-end Optimized Image Compression. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 11 2017.

Baraniuk, R. and Steeghs, P. Compressive radar imaging. In *IEEE National Radar Conference - Proceedings*, pp. 128–133. Institute of Electrical and Electronics Engineers Inc., 2007. ISBN 1424402840. doi: 10.1109/RADAR.2007.374203.

Becker, S., Bobin, J., and Candès, E. J. NESTA: A fast and accurate first-order method for sparse recovery. *SIAM Journal on Imaging Sciences*, 4(1):1–39, 1 2011. ISSN 19364954. doi: 10.1137/090756855.

Carson, W. R., Chen, M., Rodrigues, M. R. D., Calderbank, R., and Carin, L. Communications-Inspired Projection Design with Application to Compressive Sensing. *SIAM Journal on Imaging Sciences*, 5(4):1185–1212, 10 2012. doi: 10.1137/120878380.

Donoho, D. L. Compressed Sensing. *IEEE Transactions on Information Theory*, 52(4):1289, 2006. doi: 10.1109/TIT.2006.871582.

Ender, J. H. On compressive sensing applied to radar. *Signal Processing*, 90(5):1402–1414, 5 2010. ISSN 01651684. doi: 10.1016/j.sigpro.2009.11.009.

Gumbel, E. J. Statistical theory of extreme values and some practical applications. *NBS Applied Mathematics Series*, 33, 1954.

Han, Y. S., Yoo, J., and Ye, J. C. Deep Residual Learning for Compressed Sensing CT Reconstruction via Persistent Homology Analysis. 11 2016.

Herrmann, F. J., Friedlander, M. P., and Yilmaz, Fighting the curse of dimensionality: Compressive sensing in exploration seismology. *IEEE Signal Processing Magazine*, 29(3):88–100, 2012. ISSN 10535888. doi: 10.1109/MSP.2012.2185859.

Huijben, I. A. M., Veeling, B. S., and Van Sloun, R. J. G. Deep Probabilistic Subsampling for Task-Adaptive Compressed Sensing. In *International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia, 2020a.

Huijben, I. A. M., Veeling, B. S., and Van Sloun, R. J. G. Learning sub-sampling and signal recovery with applications in ultrasound imaging. *IEEE Transactions on Medical Imaging*, 2020b.

Huijben, I. A. M., Veeling, B. S., and Van Sloun, R. J. G. Learning Sampling and Model-Based Signal Recovery for Compressed Sensing MRI. In *International Conference on Acoustics, Speech and Signal Processing , ICASSP 2020*, Barcelona, Spain, 2020c.

Jang, E., Gu, S., and Poole, B. Categorical Reparameterization with Gumbel-Softmax. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 11 2016.

Ji, S., Xue, Y., and Carin, L. Bayesian compressive sensing. *IEEE Transactions on Signal Processing*, 56(6):2346–2356, 2008. doi: 10.1109/TSP.2007.914345.

Jin, K. H., Unser, M., and Yi, K. M. Self-Supervised Deep Active Accelerated MRI. 1 2019.

Kingma, D. P. and Ba, J. L. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, 12 2015.

Lai, Z., Qu, X., Liu, Y., Guo, D., Ye, J., Zhan, Z., and Chen, Z. Image reconstruction of compressed sensing MRI using graph-based redundant wavelet transform. *Medical Image Analysis*, 27:93–104, 1 2016. ISSN 13618423. doi: 10.1016/j.media.2015.05.012.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2323, 1998. ISSN 00189219. doi: 10.1109/5.726791.

Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., and Shi, W. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. *CVPR*, 2017-January:105–114, 9 2016.

Li, G., Zhu, Z., Yang, D., Chang, L., and Bai, H. On projection matrix optimization for compressive sensing systems. *IEEE Transactions on Signal Processing*, 61 (11):2887–2898, 2013.

Lustig, M., Donoho, D., and Pauly, J. M. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine*, 58(6):1182–1195, 12 2007. ISSN 07403194. doi: 10.1002/mrm.21391.

Maddison, C. J., Mnih, A., and Teh, Y. W. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Machine Learning*, 2017.

Mardani, M., Sun, Q., Vasawanala, S., Papyan, V., Monajemi, H., Pauly, J., and Donoho, D. Neural Proximal Gradient Descent for Compressive Imaging. 6 2018.

Pineda, L., Basu, S., Romero, A., Calandra, R., and Drozdzal, M. Active mr k-space sampling with reinforcement learning. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pp. 23–33, Cham, 2020. Springer International Publishing. ISBN 978-3-030-59713-9.

Ravishankar, S. and Bresler, Y. Adaptive sampling design for compressed sensing mri. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 3751–3755, 2011.

Sanchez, T., Gozcu, B., van Heeswijk, R. B., Eftekhari, A., Ilicak, E., Cukur, T., and Cevher, V. Scalable Learning-Based Sampling Optimization for Compressive Dynamic MRI. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8584–8588, Barcelona, Spain, 4 2020. doi: 10.1109/icassp40776.2020.9053345.

Sanders, J. N., Saikin, S. K., Mostame, S., Andrade, X., Widom, J. R., Marcus, A. H., and Aspuru-Guzik, A. Compressed sensing for multidimensional spectroscopy experiments. *Journal of Physical Chemistry Letters*, 3(18):2697–2702, 9 2012. ISSN 19487185. doi: 10.1021/jz300988p.

Theis, L., Shi, W., Cunningham, A., and Huszár, F. Lossy Image Compression with Compressive Autoencoders. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 3 2017.

Van Der Maaten, L. and Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image Quality Assessment: From Error Visibility to Structural Similarity. *Search Results Web results IEEE Transactions on Image Processing*, 13(4), 2004.

Weiss, T., Senouf, O., Vedula, S., Michailovich, O., Zibulevsky, M., and Bronstein, A. PILOT: Physics-Informed Learned Optimized Trajectories for Accelerated MRI. 9 2019.

Yang, Y., Sautière, G., Ryu, J. J., and Cohen, T. S. Feedback recurrent autoencoder. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3347–3351, 2020.

Zbontar, J., Knoll, F., Sriram, A., Muckley, M. J., Bruno, M., Defazio, A., Parente, M., Geras, K. J., Katsnelson, J., Chandarana, H., Zhang, Z., Drozdzal, M., Romero, A., Rabbat, M., Vincent, P., Pinkerton, J., Wang, D., Yakubova, N., Owens, E., Zitnick, C. L., Recht, M. P., Sodickson, D. K., and Lui, Y. W. fastmri: An open dataset and benchmarks for accelerated MRI. *CoRR*, abs/1811.08839, 2018.

Zhang, Z., Romero, A., Muckley, M. J., Vincent, P., Yang, L., and Drozdzal, M. Reducing Uncertainty in Undersampled MRI Reconstruction with Active Acquisition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June: 2049–2053, 2 2019.