Appendix A. Further Details on Theory

A.1. CountSketch Theory

We use a theorem from Cohen et al. (2016) that builds on several results for CountSketch matrices, giving theoretical guarantees for sketching quadratic forms of matrices. The theorem is re-phrased for our purposes, showing the quality of approximation by the sketch in preserving ℓ_2-norms of vectors in the subspace spanned by the columns of W, the matrix that is being sketched. There is a trade-off in the quality of approximation by the sketch and its size, given by the dimension t of the columns of the sketch matrix S. The quality of approximation depends on the ℓ_2-norm of θ and the spectrum of W, namely the operator norm $\|W\|_2$ and the Frobenius norm $\|W\|_F$.

Before we describe the theorem from Cohen et al. (2016), we define a moment property for sketch distributions.

Definition 1 (OSE Moment Property)**.** A distribution $\mathcal{D}$ over $\mathbb{R}^{t\times n}$ satisfies the (ϵ, δ, d)-OSE moment property if there exists some $l \geq 2$ such that for all matrices $U \in \mathbb{R}^{n\times d}$ with orthonormal columns, $\mathbb{E}_{S\sim\mathcal{D}}[\|(SU)^\top(SU) - I\|_2^l] < \epsilon^l \cdot \delta$.

The result of Cohen et al. (2016) shows that when the sketch matrix S is drawn from a distribution $\mathcal{D}$ over $\mathbb{R}^{n\times d}$ satisfying the (ϵ, δ, k)-OSE moment property, the sketch SW approximately preserves the ℓ_2 norm of all vectors in the column-span of W upto additive error that depends on ϵ, k and the spectrum of W.

Theorem 3 (Theorem 6, Cohen et al. (2016))**.** *Let $W \in \mathbb{R}^{n\times d}$ be a matrix and $k \in \mathbb{R}^+$, $\epsilon, \delta \in (0, 1/2]$ be constants. A matrix $S \in \mathbb{R}^{t\times n}$ drawn from a distribution $\mathcal{D}$ over $\mathbb{R}^{t\times n}$ satisfying the $(\epsilon, \delta, 2k)$-OSE moment property has the property that for all $\theta \in \mathbb{R}^d$ simultaneously,*

$$\left|\|SW\theta\|_2^2 - \|W\theta\|_2^2\right| \leq \epsilon\|\theta\|_2^2 \left(\|W\|_2^2 + \frac{\|W\|_F^2}{k}\right) \tag{7}$$

with probability at least $1 - \delta$ where the probability is taken over the distribution $\mathcal{D}$.

A corollary to the theorem is that when the matrix $W \in \mathbb{R}^{n\times d}$ has stable rank r and the distribution $\mathcal{D}$ satisfies the $(\epsilon, \delta, 2r)$-OSE moment property, the quantity $\|SW\theta\|_2^2$ is close to $\|W\theta\|_2^2$ up to additive error $\epsilon\|W\|_2^2\|\theta\|_2^2$ – an error rate that is the standard benchmark for subspace embeddings.

A.2. Proof of Theorem 2

Fix $\tau \in \mathbb{N}$ to be the number of tasks, $\alpha \in [0, 1]$ to be a constant and let $W_1, \ldots, W_\tau \in \mathbb{R}^{n\times m}$ be the matrices given by (1) for each of the τ tasks. Let $S_1, \ldots, S_\tau \in \mathbb{R}^{t\times n}$ be τ matrices drawn i.i.d from a distribution $\mathcal{D}$ satisfying the $(\epsilon, \delta, 2k)$-OSE moment property for constants $k \in \mathbb{N}^+$ and $\epsilon, \delta \in (0, 1/2]$.

We first define the following concatenated matrices: i) $\boldsymbol{S} = [S_1; S_2; \ldots; S_\tau]$ where $[A; B]$ is the column-concatenation for two matrices A, B with the same number of rows, and ii) $\boldsymbol{W} = [\alpha_1 W_1|\alpha_2 W_2|\ldots|\alpha_\tau W_\tau]$ where $\alpha_i = \sqrt{\alpha(1-\alpha)^{\tau-i}}$ and $[A|B]$ is the row-concatenation for two matrices A, B with the same number of columns. We can then write the online sketched importance matrix $\widetilde{W}_\tau$ from (5) as

$$\widetilde{W}_\tau = \sum_{i=1}^{\tau} \sqrt{\alpha(1-\alpha)^{\tau-i}} \cdot S_i W_i = \boldsymbol{S}\boldsymbol{W}. \tag{8}$$

We can also check that the regularizer used when learning on task $\tau + 1$, given by (6), can be written as

$$\widetilde{\mathcal{R}}_\tau(\theta) = \frac{1}{2n}\|\widetilde{W}_\tau(\theta - \theta^*)\| = \frac{1}{2n}\|\boldsymbol{S}\boldsymbol{W}(\theta - \theta^*)\|_2^2. \tag{9}$$

In order to prove Theorem 2, it is sufficient to prove that the concatenated sketch $\boldsymbol{S}$ satisfies the $(\epsilon, \delta, 2k)$-OSE moment property since combining this with Theorem 3 from Cohen et al. (2016) will yield the result.

Our proof of the OSE moment property of the concatenated sketch relies on the OSE moment property of a family of distributions on matrices called the Sparse Johnson-Lindenstrauss Transform (SJLT) which we define next, reproducing the definition of the SJLT from Cohen (2016).

Definition 2. A Sparse Johnson-Lindenstrauss Transform (SJLT) of dimension $t \in \mathbb{N}^+$ for points in $\mathbb{R}^n$ is a random matrix S in $\mathbb{R}^{t\times n}$ with exactly $1 \leq s \leq t$ non-zero entries in each column such that $S_{ij} = \eta_{ij}\sigma_{ij}/\sqrt{s}$; here each σ_{ij} is a Rademacher random variable and $\{\eta_{ij}\}$ is a collection of Bernoulli random variables with the following properties:

1. For all $j \in [n], \sum_{i=1}^{t} \eta_{ij} = s$. That is, each column has *exactly* s non-zero entries.
2. For all $i \in [t], j \in [n], \mathbb{E}\eta_{ij} = s/t$.
3. The η_{ij} are negatively correlated: $\forall X \subseteq [t] \times [n]$, $\mathbb{E}\prod_{(i,j)\in X} \eta_{ij} \leq \prod_{(i,j)\in X} \mathbb{E}\eta_{ij} \leq (s/t)^{|X|}$.

We can check then that the matrix $\boldsymbol{S}$ is a s-sparse $t \times n\tau$-dimensional SJLT when each of the matrices $S_1, \ldots, S_\tau$ are drawn independently from a s-sparse, t-dimensional SJLT for points in $\mathbb{R}^n$. Properties 1 and 2 follow easily since $\boldsymbol{S} = [S_1; \ldots; S_\tau]$ and the fact that matrices $S_1, \ldots, S_\tau$ are drawn independently. Property 3 follows by partitioning every set $X \subset [t] \times [n\tau]$ by the S matrix in which it appears and then applying Property 3 for each part in the partition.

One can check that the CountSketch distribution, as defined in Algorithm 1, is a 1-sparse SJLT. It was shown in Nelson and Nguyên (2013) and Meng and Mahoney (2013) that the 1-sparse SJLT satsifies the $(\epsilon, \delta, 2k)$-OSE moment property for $t = \Omega(k^2/(\epsilon^2\delta))$

We have now shown that the matrix $\boldsymbol{S}$ is a 1-sparse SJLT, satisfying the (ϵ, δ, k) moment property with $t = \Omega(k^2/(\epsilon^2\delta))$ rows. Hence, by applying Theorem 3 to $\boldsymbol{S}$ and using the definition of $\boldsymbol{S}$, Theorem 2 follows.

Appendix B. Further Details on Experiments

B.1. Synthetic Experiments

Setups. For the regularization matrix induced by EWC and MAS, we compare the performance of various approaches to approximating the importance matrix including:

(i) a diagonal approximation;
(ii) a block-diagonal approximation, with a sequence of 50×50 non-zero blocks along the diagonal;
(iii) sketched approximation with $t = 50$;
(iv) a rank-1 SVD;
(v) a low rank (rank $= 50$) SVD;
(vi) the full importance matrix.

We use a small multi-layer perceptron with the architecture $2 \to 128 \to 64 \to 2$ and with ReLU activation function. For all algorithms, we use ADAM as the optimizer with learning rate 10^{-3}. The minibatch size is 100, and we use the importance parameter $\lambda = 10^3$ and the online learning parameter $\alpha = 0.5$ for all experiments. We repeat all toy example experiments 5 times with different fixed seeds, and report the average accuracy on all tasks. These toy example experiments are conducted on one RTX2080Ti GPU.

Online Learning in Synthetic Experiments. For non-sketched approaches, the regularizer (2) in SR methods is approximated by

$$\widetilde{\mathcal{R}}(\theta) := \frac{1}{2}(\theta - \theta_A^*)^\top \widetilde{\Omega}(\theta - \theta_A^*) \tag{10}$$

where $\widetilde{\Omega}$ approximates the importance matrix Ω. The online extension of Sketched SR (see Section 4) applies moving average on the sketch $\widetilde{W}$, and cannot be directly applied on the regularizer in Equation 10. To ensure faithful comparison, moving average is applied on the importance matrix $\widetilde{\Omega}$ in synthetic experiments according to Equation (4).

This corresponds to our observation in the permuted MNIST experiments.

B.2. Sketched versus Diagonal Regularization

B.2.1. Dataset: CIFAR-100 Distribution Shift

For our CIFAR-100 experiment, we follow the 5-task *CIFAR-100 Distribution Shift* dataset introduced in Ramasesh et al. (2021). In our experiment, all 5 tasks are 5-class classification problems, where each class is one of the 20 superclasses of the CIFAR-100 dataset. For instance, we take the five superclasses *aquatic mammals*, *fruits and vegetables*, *household electrical devices*, *trees*, and *vehicles-1*. The corresponding subclasses for Task 1 are (1) *dolphin*, (2) *apple*, (3) *lamp*, (4) *maple tree*, and (5) *bicycle*, while for Task 2, they are (1) *whale*, (2) *orange*, (3) *television*, (4) *willow*, and (5) *motorcycle*. Figure 7 shows sample images and five random augmentations for the classes in both tasks.

B.2.2. Setup

Permuted-MNIST We use a multi-layer perceptron with the architecture $784 \to 1024 \to 512 \to 256 \to 10$ with ReLU activation function and no bias to learn this classification task. We use ADAM as the optimizer with learning rate 10^{-4} and the online learning parameter $\alpha = 0.25$ for all algorithms. The minibatch size is 100. For each algorithm, a grid search on the regularization coefficient $\lambda \in \{10^i \mid i = 2, 3, \ldots, 6\}$ is used to determine the optimal hyperparameter for the reported results. We uses 50 sketches in Sketched SR to approximate the full importance matrix. All permuted-MNIST experiments are repeated 5 times with

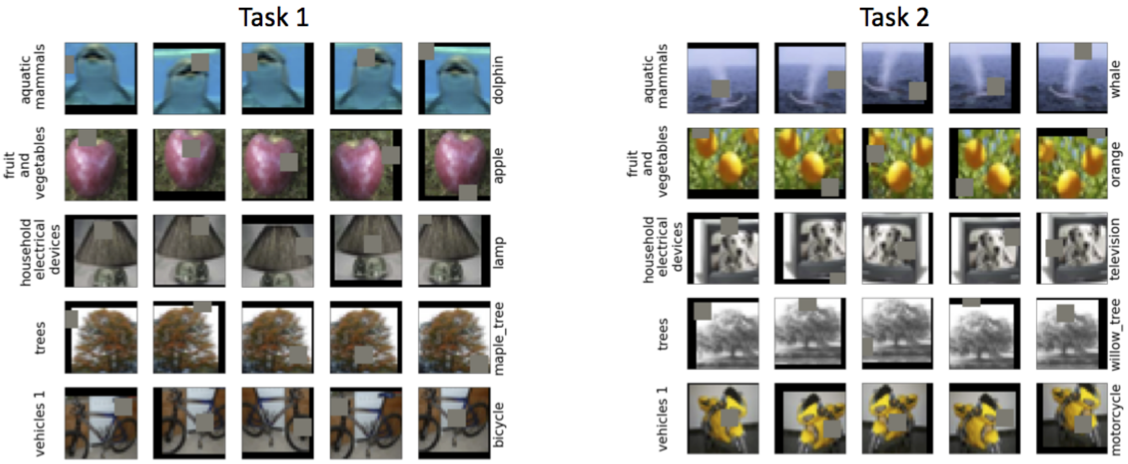


Figure 7: Sample images with 5 random augmentations for Task 1 and Task 2 in our CIFAR-100 experiment. The five superclasses for both tasks are represented by each row (labelled on the left), while the corresponding subclasses for each task are represented by rows within the task (labelled on the right).

different fixed seeds, and we report average accuracy on all tasks. We run permuted-MNIST experiments on a Tesla K80.

CIFAR-100 In all experiments, we used a Wide-ResNet (Zagoruyko and Komodakis, 2016) as our backbone. The network has 16 layers, a widening factor of 4, and a dropout rate of 0.2. We leveraged random flip, translation, and cutout (DeVries and Taylor, 2017) as augmentation. We use ADAM as our optimizer for all experiments, with learning rate 10^{-3} and momentum 0.9. The importance parameter λ for each algorithm is: 10^5 for EWC, 10^2 for Sketched EWC, 10^5 for MAS, 10^3 for Sketched MAS. The minibatch size is 64. The online learning parameter is $\alpha = 0.25$ for all experiments. In Sketched SR algorithms, we uses 50 sketches to approximate the full importance matrix. All reported results are averaged over 10 runs with different random seeds.

B.3. Additional Experiments

100-task Online Learning. Sketched SR methods store more parameters for the regularizer, requiring more memory than diagonal SR methods. However with the online learning regime, when the task number grows, sketched SR method use less parameters in memory than the corresponding offline diagonal SR method. In this experiment, we consider the sketched SR method on 100 consecutive permuted-MNIST tasks. The results are shown in Figure 8. We can see that sketched EWC significantly outperforms online diagonal EWC, especially when number of tasks grows, with storing less parameters than offline EWC in this experiment setting.

Effects of the Sketch Size per Task. We further study the effects of the size of the sketch t (See Equation 3) on the performance of sketched SR on each task. The results are

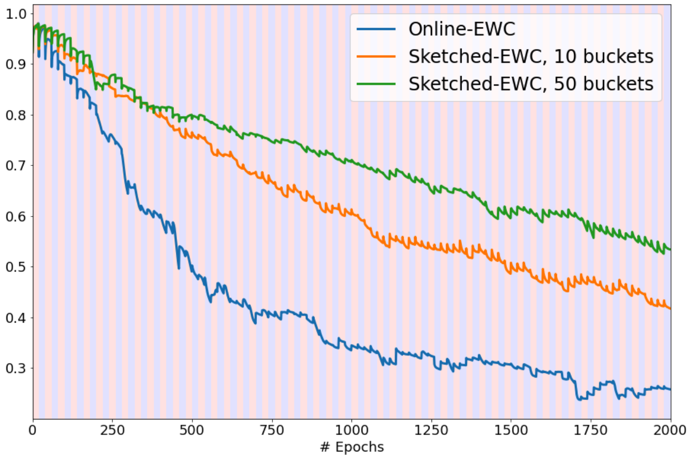
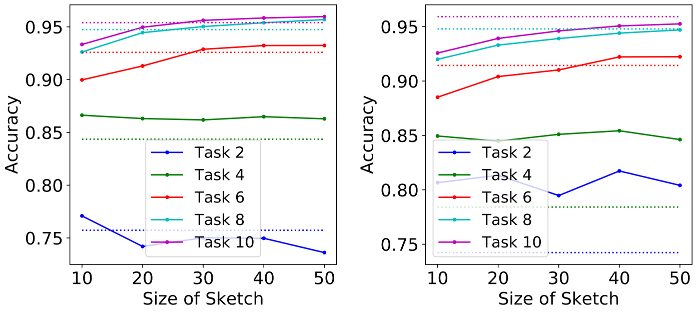


Figure 8: The averaged accuracy (over all tasks) of sketched EWC (resp. MAS, SCP) vs. diagonal EWC (resp. MAS, SCP) on 100-task permuted-MNIST. For the sketched methods, we set $t = 50$. The plots suggest that online sketched methods outperform their diagonal counterparts more significantly when number of tasks grows.



(a) Sketched EWC (b) Sketched MAS

Figure 9: Effect of the sketch size (t) on task accuracy of sketched methods for learning 10 permuted-MNIST tasks. Dotted line represents the accuracy of diagonal methods on the corresponding task with the same color. We can immediately observe that as the number of sketches increases, sketched methods tend to perform better in later tasks (Task ID ≥ 6).

shown in Figure 9. From the plot we see a clear trade-off between the size of the sketch and the accuracy on later tasks, where the accuracy consistently increases as the size of sketches grows. This directly shows that increasing of the size of sketches improves learning capability for new tasks (known in the literature as *intransigence*), with the expense of more computation resources.