Table 1: μ accuracies $\pm\sigma$ of the main experiments.

	dis. MNIST, M=300	perm. MNIST	cifar-10
Random	37.5 ± 1.35	72.7 ± 1.45	28.6 ± 1.23
GSS-IQP	75.9 ± 2.54	77.3 ± 0.54	-
GSS-Clust	75.7 ± 2.17	79.9 ± 0.69	22.5 ± 0.43
FSS-Clust	75.8 ± 1.56	77.8 ± 0.33	26.7 ± 1.47
GSS-Greedy	82.6 ± 2.9	77.3 ± 0.50	33.5 ± 1.62

We thank the reviewers for their valuable comments. We want to emphasize that our work is the first to tackle the problem of online continual learning with *no task boundaries*, i.e. differently from other methods e.g. GEM, iCaRL, we receive no information when task T1 switches to T2, which makes the input stream highly non iid at the boundary. Our method is thus applicable to scenarios of continual learning when task id is unknown. We will expand the related work and improve the presentation of the figures and the tables.

Error bars? larger variance for greedy? Tab. 1 reports the main results with their standard deviation estimated over the different runs. This confirms that the advantage of our method is statistically significant.

R1: **What if no feasible region exists (empty $\tilde{C}$)?** This is a good point. When the data contains outliers it is an open question how to distinguish outliers from under-represented examples. As for this writing, we don't consider the existence of outliers. Nevertheless, our method should still work with outliers: 1) Our surrogate considers pairs of samples, although the outlier is likely to be selected, pairs that don't contain the outliers will not be affected. 2) Rigorous satisfaction of the constraints is not possible in non-convex neural networks, it is converted to soft regularization.

Fig 2 , bigger range of angle/surrogate values? The range of the solid angle is already from 0 to 1 which is the full range of solid angle. In the log scale it may seem truncated because we omit zeros which is negative infinity in log scale.

Which constraints are typically satisfied or violated? The samples are selected to be diverse, and the discarded samples usually have similar counterparts in the buffer. We found the selected samples to cover the different patterns learned classes leading to a balanced buffer even with an imbalanced data stream.

iCaRL CIFAR-10 low. We consider online incremental classification, differently from iCaRL, Ln 294. iCaRL performs offline training with large batches and repeatedly revisiting over the classes of a task.

Computational improvement of GSS-greedy? For our greedy alternative, the major cost corresponds to the estimation of the n gradients of the drawn samples, n was fixed to 10 in all our experiments. For clustering based sample selection, the features/gradients of the buffer samples and the corresponding distances need to be computed at each step.

R2: **Clarification on high level picture** We start from the constrained optimization view of continual learning like in GEM, which needs to be relaxed by constraint selection. Instead of random selection, we perform constraint selection by minimizing the solid angle formed by the constraints. We propose a surrogate for the solid angle objective, and show their relation numerically. Finally we test the effectiveness of the surrogate on continual learning benchmarks.

High dimensional space and when the buffer is large? We only perform the numerical analysis of the monotonous correlation of the solid angle and the surrogate up to 200 dimension, because the complexity of Monte-Carlo integration increases exponentially with the dimension. Empirically we verified with experiments with a buffer size up to 1k and show it is still effective. We also showed another interpretation of our surrogate i.e. to maximize the diversity of the samples in the "gradient" space. We believe diversity maximization is likely to still work in high dimensional space.

Maximising diversity is similar to existing coreset algorithms / coverage maximisation? A key difference is that we consider the diversity in the gradient space instead of data space. The coverage maximization reported are performed in the data space using euclidean metric, which is not working equally well for high dimensional data like image, (Fig. 7 in the coverage maximisation paper). Our contribution is the proposal of using gradient information to measure diversity which is supported by the constrained optimization (Eq. 1). We considered two incremental clustering baselines, in the feature space and in the gradient space, and showed that our approach is superior (Tab. 1,3,4 in the paper).

Methods that combine replay with parameter regularisation criticized? We state that hybridizing prior focused with replay based is necessary to achieve better performance on long sequences, Ln 27-29. Also, there's a difference between the two "regularizations". The one mentioned in the intro is the prior focused regularizing the parameters while the regularization mentioned later is about converting the hard constraint into a regularization term i.e. rehearsal.

More standard benchmarks? We opt for more realistic setting with online training with non stationary distribution and never ending training. This is already the standard setting considered in GEM but with multi-head. Our method aims to solve the hard setting and not the task incremental setting where a task oracle is used at training and test time.

R3: **Consistency analysis for the greedy method? As increasing n to M, is it equivalent to the IQP method?** For greedy, the drawn samples are used to estimate a score of the new sample indicating its similarity to the buffer samples. IQP is exact and solves the surrogate, Eq.7 at each step. Greedy keeps scores of buffer samples once added which relate to their chance of being replaced. We fixed n to 10 on all settings. We abate the effect of n with dis. MNIST, M= 300. We get the following accuracies n=1:67.3, n=5:79.0, n=10:82.6, n=20:79.6, n=30:79.9, n=40:78.4, n=50:78.3.

Batch size effect. We want to be as close as possible to the full online setting. Batch size of 10, used in GEM, seems a good approximation. We run dis.MNIST, M=300, with different batch sizes. The mean accuracies of GSS-greedy are B=5:82.9, B=10:82.6, B=20:84.23, B=50: 80.2, B=100:72.9. Large batch sizes lead less parameter updates.

Random high accuracy on T4. Its buffer has very few samples from previous tasks, 24 for T1-T3, and more from the recent T4 (127) & T5 (849) given M= 1k. As such, it forgets less the more recent task at the cost of older ones.

Backward Transfer. It is shown in GEM with multi head, i.e. when evaluating a task, only its classes are considered. Under shared head this is different. When a task is first evaluated, fewer classes are learned and others will have low accuracy. At the end of sequence, all classes compete which makes backward transfer hard to achieve.