SAE: Sequential Anchored Ensembles

Arnaud Delaunoy  
University of Liège  
a.delaunoy@uliege.be

Gilles Louppe  
University of Liège  
g.louppe@uliege.be

Abstract

Computing the Bayesian posterior of a neural network is a challenging task due to the high-dimensionality of the parameter space. Anchored ensembles approximate the posterior by training an ensemble of neural networks on anchored losses designed for the optima to follow the Bayesian posterior. Training an ensemble, however, becomes computationally expensive as its number of members grows since the full training procedure is repeated for each member. In this note, we present Sequential Anchored Ensembles (SAE), a lightweight alternative to anchored ensembles. Instead of training each member of the ensemble from scratch, the members are trained sequentially on losses sampled with high auto-correlation, hence enabling fast convergence of the neural networks and efficient approximation of the Bayesian posterior. SAE outperform anchored ensembles, for a given computational budget, on some benchmarks while showing comparable performance on the others and achieved 2nd and 3rd place in the light and extended tracks of the NeurIPS 2021 Approximate Inference in Bayesian Deep Learning competition.

1 Introduction

Accurate uncertainty quantification has become of high importance for machine learning tasks where incorrect prediction could have severe consequences. Bayesian deep learning is a popular framework for capturing those uncertainties. It consists in estimating the Bayesian posterior

$$p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{p(D)},$$

where \( D \) is a dataset and \( \theta \) the parameters of the neural network. Predictions are then made through marginalization over the parameters

$$p(y \mid x, D) = \int p(y \mid x, \theta)p(\theta \mid D)d\theta,$$

where \( y \) are the outputs and \( x \) are the inputs. In practice, inferring the Bayesian posterior is challenging due to the high-dimensionality of the parameter space; scalable techniques to approximate this posterior are hence needed.

One of such techniques is called anchored ensembles [Pearce et al., 2020]. It builds an ensemble of neural networks with a randomized objective function designed for the optima to approximately follow the Bayesian posterior \( p(\theta \mid D) \). Under the assumption of a normal prior \( p(\theta) = \mathcal{N}(\mu_{\text{prior}}, \Sigma_{\text{prior}}) \) and likelihood \( p(D \mid \theta) \), the optima \( \theta^* \) obtained by minimizing the anchored loss \( -(\log p(D \mid \theta) + \log p_{\text{anc}}(\theta)) \), where \( p_{\text{anc}} = \mathcal{N}(\theta_{\text{anc}}, \Sigma_{\text{prior}}) \), approximately follow the Bayesian posterior \( p(\theta \mid D) \) if \( \theta_{\text{anc}} \sim p(\theta) \).
Algorithm 1 Anchored Ensembling (AE)

\[
\text{for } i \in 1, \ldots, N \text{ do} \\
\theta_{\text{anc},i} \sim p(\theta) \quad \triangleright \text{Sample anchor} \\
\theta_{\text{init},i} \leftarrow \text{INIT()} \quad \triangleright \text{Initialize NN} \\
\theta^*_i \leftarrow \text{TRAIN}(\theta_{\text{anc},i}, \theta_{\text{init},i}) \quad \triangleright \text{Minimize loss} \\
\theta^*_i \leftarrow \arg\min_{\theta} \left\{ \log p(D | \theta) + \log p_{\text{anc}}(\theta) \right\}
\]

2 Sequential anchored ensembles

In this work, we propose Sequential Anchored Ensembles (SAE) which stem from the observation that if \(\theta_{\text{anc},i-1}\) and \(\theta_{\text{anc},i}\) are close so should be \(\theta^*_i\) and \(\theta^*_{i-1}\). Therefore, training all the members of the ensemble from scratch is inefficient. We exploit this to reduce the computational cost of the training procedure by training the elements of the ensemble sequentially starting from the previous solution \(\theta^*_{i-1}\). This is illustrated in Figure 1 and summarized in Algorithm 2.

![Figure 1: While anchored ensembles start the training from scratch at a high loss, sequential anchored ensembles start from the previously obtained solution and hence at a lower loss allowing to build an ensemble of 21 members in the time anchored ensembles built an ensemble of 2 members. Each peak corresponds to the start of a new training procedure, the fast convergence of sequential anchored ensembles allows to train much more ensemble members than anchored ensembles in the same time.](image)

Let us consider the construction of an ensemble composed of \(N\) members \(\theta^*_1, \ldots, \theta^*_N\), with \(\theta^*_i\) constructed based on the anchor \(\theta_{\text{anc},i}\). A first anchor is sampled from the prior and the corresponding neural network is trained in a classical way. The next anchors are sampled such that two consecutive anchors are close to each other and the corresponding neural networks are trained starting from the previous optimum. As consecutive anchors are close, so are the solutions, and the training procedure is then expected to converge much faster. To sample consecutive anchors that are close to each other but eventually span the prior, we rely on an MCMC procedure. As the parameters are assumed independent under the prior, we run a separate chain for each parameter of the neural network so that some transitions are accepted by the Metropolis-Hastings algorithm at each step. The parameters of the anchors being independently normally distributed, the distribution is easy to navigate as opposed to an MCMC procedure directly performed on the posterior. To benefit from parallelization and decrease ensemble correlation, the algorithm can be run multiple times with different initializations.

Algorithm 2 Sequential Anchored Ensembling (SAE)

\[
\theta_{\text{anc},1} \sim p(\theta) \\
\theta_{\text{init},1} \leftarrow \text{init()} \\
d_1 \leftarrow \text{RANDCHOICE}(_{\text{SIZE}}\{\theta_{\text{anc},1}\}, \{-1, 1\}) \\
\theta^*_1 \leftarrow \text{TRAIN}(\theta_{\text{anc},1}, \theta_{\text{init},1}) \\
\text{for } i \in 2, \ldots, N \text{ do} \\
\theta_{\text{anc},i}, d_i \leftarrow \text{MH}\_\text{UPDATE}((\theta_{\text{anc},i-1}, d_{i-1})) \\
\theta_{\text{init},i} \leftarrow \theta^*_{i-1} \\
\theta^*_i \leftarrow \text{TRAIN}(\theta_{\text{anc},i}, \theta_{\text{init},i}) \\
\]

\[
\theta^*_i \leftarrow \arg\min_{\theta} \left\{ \log p(D | \theta) + \log p_{\text{anc}}(\theta) \right\} \\
\]

\[
\text{Sample first anchor} \quad \triangleright \text{Sample anchor} \\
\text{Initialize NN} \quad \triangleright \text{Initialize NN} \\
\text{Direction} \quad \triangleright \text{Direction} \\
\text{Long training} \quad \triangleright \text{Start from previous optimum} \\
\text{i}^{th} \text{ step of the SAE algorithm} \quad \triangleright \text{Alg.} [3] \\
\]
while making small transitions for training to remain fast. We have found that a guided walk
The true posterior used for comparison has been computed by a Hamiltonian Monte-Carlo procedure
The code used to conduct the experiments is available at
The algorithm can be performed with any MCMC procedure, however, to be efficient, the MCMC
procedure should eventually span the whole prior space so that the optima span the whole posterior
while small transitions for training to remain fast. We have found that a guided walk
Metropolis-Hastings procedure with Gaussian transitions [Gustafson, 1998] performs well. This
procedure is illustrated in Algorithm 3. The difference between a classical Metropolis-Hastings
algorithm and its guided walk version is that the latter always performs transitions in the same
direction until a rejection occurs. On rejection, the transition direction is inverted until the next
rejection. As we know the prior to be normally distributed, rejection cannot occur until the point of
maximal density is passed. The anchors will then consistently evolve in the same direction hence
allowing to efficiently span the prior with short steps.

3 Experiments

The code used to conduct the experiments is available at [https://github.com/ADelau/SAE-Sequential-Anchored-Ensembles](https://github.com/ADelau/SAE-Sequential-Anchored-Ensembles) To assess the efficiency of SAE, we compare
how close the approximated predictive density $\int p(y | x, \theta)p(\theta | D)d\theta$ both for anchored ensembles and sequential anchored ensembles. For
a given computational budget, the sequential anchored ensembles will be composed of more members
than the anchored ensembles but each member will be trained for a shorter time and built sequentially.
The true posterior used for comparison has been computed by a Hamiltonian Monte-Carlo procedure
[Izmailov et al., 2021] in the context of the NeurIPS 2021 Approximate Inference in Bayesian Deep
Learning competition [Wilson et al., 2021].

|                  | Cifar10 Resnet-20 | Cifar10(-C) Alexnet | IMDB | DermaMNIST | UCI-Gap |
|------------------|------------------|--------------------|------|------------|--------|
|                  | Ag.  | TV  | Ag.  | TV  | Ag.  | TV  | Ag.  | TV  | W2   |
| epochs 1000      | AE   | 0.849| 0.201| 0.726| 0.262| 0.892| 0.109| 0.877| 0.104| 0.148|
| epochs 10,000    | SAE  | 0.856| 0.176| 0.772| 0.212| 0.887| 0.110| 0.880| 0.098| 0.159|
|                  | AE   | 0.862| 0.199| 0.746| 0.236| 0.926| 0.086| 0.897| 0.089| 0.137|
|                  | SAE  | 0.903| 0.133| 0.787| 0.200| 0.916| 0.099| 0.893| 0.086| 0.143|

Table 1: Comparison of the performance of anchored ensembles and sequential anchored ensembles. While the computational budget is split uniformly between ensemble members for anchored ensembles, sequential anchored ensembles perform many short training procedures allowing to construct more ensemble members for the same computational budget. The reported values are the median performance over at least 20 runs. Cifar10(-C) corresponds to a dataset composed of 20% of uncorrupted samples and 80% of samples with a corruption level of 4.
worse on UCI-Gap and shows similar performance on the other datasets. It shows that the sequential procedure is able to navigate the posterior and sometimes approximates the posterior density more efficiently than traditional anchored ensembles. Interestingly, we observe that when both methods yield a similar agreement, SAE usually yields a lower total variation. SAE hence tends to provide better-calibrated posteriors even when classification performance is similar. The reason why it performs better on some datasets than others is still unclear. However, we hypothesize that some posteriors are easier to navigate leading to high performance using SAE while others are harder and benefit more from different initializations of the weights.

4 Related work

Building an ensemble in a sequential manner has already been proven successful in the context of non-anchored deep ensembles. Snapshot ensembles [Huang et al., 2017] and Fast Geometric Ensembling [Garipov et al., 2018] build a deep ensemble by recording weights during the training procedure. They make use of a cyclical learning rate schedule to increase diversity. Stochastic Weights Averaging [Izmailov et al., 2018] directly averages the weights in place of the predictions leading to lower memory requirements and prediction complexity. Sequential anchored ensembling follows similar ideas but uses anchors to create diversity in the ensemble following the Bayesian posterior. Improvements of those methods include low-precision training [Yang et al., 2019] and Simplicial Pointwise Random Optimization [Benton et al., 2021] which build simplexes of low loss. Such ideas could possibly be adapted for the sequential anchored ensembles setting, improving its performance.

Acknowledgments and Disclosure of Funding

The authors would like to thank Tim Pearce for his insightful comments and feedback. Arnaud Delaunoy would like to thank the National Fund for Scientific Research (F.R.S.-FNRS) for his scholarship. Gilles Louppe is recipient of the ULiège - NRB Chair on Big Data and is thankful for the support of the NRB.

References

Gregory Benton, Wesley Maddox, Sanae Lotfi, and Andrew Gordon Gordon Wilson. Loss surface simplexes for mode connecting volumes and fast ensembling. In Proceedings of the 38th International Conference on Machine Learning, volume 139, pages 769–779. PMLR, 18–24 Jul 2021.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnn. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, pages 8803–8812, 2018.

Paul Gustafson. A guided walk metropolis algorithm. Statistics and computing, 8(4):357–364, 1998.

Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get m for free. International Conference on Learning Representations, 2017.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. arXiv preprint arXiv:1803.05407, 2018.

Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Wilson. What are bayesian neural network posteriors really like? arXiv preprint arXiv:2104.14421, 2021.

Tim Pearce, Felix Leibfried, and Alexandra Brintrup. Uncertainty in neural networks: Approximately bayesian ensembling. In International conference on artificial intelligence and statistics, pages 234–244. PMLR, 2020.

Andrew Gordon Wilson, Pavel Izmailov, Matthew D Hoffman, Yarin Gal, Yingzhen Li, Melanie F Pradier, Sharad Vikram, Andrew Foong, Sanae Lotfi, and Sebastian Farquhar. Evaluating approximate inference in bayesian deep learning. 2021.
A Additional experiments

In this section, we provide results on smaller budgets and all the experimental details, results are shown in Table 2. Details about the architectures can be found in Table 3. The allocation of the computational budget for the UCI-Gap dataset can be found in Table 5 while the others can be found in Table 4. The metrics used to compare the approximate posterior to the true one are, for classification, the agreement

\[
\text{agreement}(p, \hat{p}) = \frac{1}{n} \sum_{i=1}^{N} I[\arg \max_{j} \hat{p}(y_j|x_i) = \arg \max_{j} p(y_j|x_i)]
\]

and the total variation

\[
\text{TV}(p, \hat{p}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \sum_{j} |p(y_j|x_i) - \hat{p}(y_j|x_i)|,
\]

where \(p\) is the true predictive density and \(\hat{p}\) is the approximate one. For regression we use a point-wise Wasserstein-2 distance based on samples from the predictive distribution

\[
W_2(p, \hat{p}) = \inf_{I} \sqrt{\sum_{i \in I} \sum_{j} |p_i - \hat{p}_j|^2}.
\]
| Agreement       | budget (epochs) | 200   | 500   | 1000  | 10,000 |
|-----------------|----------------|-------|-------|-------|--------|
| Cifar10 Resnet-20 | AE             | 0.804±0.010 | 0.833±0.009 | 0.849±0.006 | 0.862±0.005 |
|                 | SAE            | 0.782±0.012 | 0.830±0.010 | 0.856±0.006 | 0.903±0.004 |
| Cifar10-C Alexnet | AE             | 0.660±0.007 | 0.706±0.005 | 0.726±0.004 | 0.746±0.003 |
|                 | SAE            | 0.738±0.006 | 0.762±0.006 | 0.772±0.005 | 0.787±0.002 |
| IMDB            | AE             | 0.829±0.008 | 0.865±0.007 | 0.892±0.004 | 0.926±0.001 |
|                 | SAE            | 0.831±0.005 | 0.868±0.006 | 0.887±0.007 | 0.916±0.002 |
| DermaMNIST      | AE             | 0.802±0.014 | 0.865±0.011 | 0.877±0.007 | 0.897±0.005 |
|                 | SAE            | 0.836±0.009 | 0.869±0.011 | 0.880±0.012 | 0.893±0.008 |

| Total variation | budget (epochs) | 200   | 500   | 1000  | 10,000 |
|-----------------|----------------|-------|-------|-------|--------|
| Cifar10 Resnet-20 | AE             | 0.225±0.017 | 0.211±0.015 | 0.201±0.012 | 0.199±0.003 |
|                 | SAE            | 0.223±0.014 | 0.194±0.012 | 0.176±0.011 | 0.133±0.003 |
| Cifar10-C Alexnet | AE             | 0.335±0.005 | 0.283±0.005 | 0.262±0.006 | 0.236±0.006 |
|                 | SAE            | 0.245±0.004 | 0.221±0.006 | 0.212±0.004 | 0.200±0.003 |
| IMDB            | AE             | 0.180±0.007 | 0.131±0.005 | 0.109±0.003 | 0.086±0.001 |
|                 | SAE            | 0.160±0.003 | 0.124±0.004 | 0.110±0.004 | 0.099±0.002 |
| DermaMNIST      | AE             | 0.153±0.012 | 0.117±0.006 | 0.104±0.007 | 0.089±0.005 |
|                 | SAE            | 0.143±0.023 | 0.111±0.006 | 0.098±0.011 | 0.086±0.004 |

| W_2            | budget (epochs) | 200   | 500   | 1000  | 10,000 |
|----------------|----------------|-------|-------|-------|--------|
| UCI-Gap        | AE             | 0.174±0.357 | 0.152±0.083 | 0.148±0.142 | 0.137±0.039 |
|                 | SAE            | 0.196±0.341 | 0.169±0.149 | 0.159±0.162 | 0.143±0.096 |

Table 2: Comparison of the performance of anchored ensembles and sequential anchored ensembles for various computational budgets. While the computational budget is split uniformly between ensemble members for anchored ensembles, sequential anchored ensembles perform many short training procedures allowing to construct more ensemble members for the same computational budget. The reported values are the median, minimal and maximal performance over 100 runs for the UCI-Gap dataset and over 20 runs for the other datasets.

| Dataset  | Architecture                  |
|----------|-------------------------------|
| Cifar10 Resnet-20 | 20-layers CNN       |
| Cifar10 Alexnet   | 5-layers CNN            |
| IMDB              | 1 Conv layer + LSTM cell|
| DermaMNIST        | 2-layers CNN            |
| UCI-Gap           | 2-layers MLP             |

Table 3: Architectures used with each dataset.
| Budget (epochs) | 200        | 500        | 1000       | 10,000     |
|----------------|------------|------------|------------|------------|
| AE             | 2 members of 100 epochs | 5 × 100   | 10 × 100   | 100 × 100  |
|                | 2 members  | 5 members  | 10 members | 100 members|
| SAE            | 1 chain of 100 first epochs | 2(100 + 75 × 2) | 3(100 + 116 × 2) | 10(100 + 450 × 2) |
|                | 50 sequential trainings of 2 epochs | 152 members | 351 members | 4510 members |

Table 4: Allocation of the computational budget for the Cifar10, Cifar10(-C), IMDB and DermaMNIST datasets.

| Budget (epochs) | 200        | 500        | 1000       | 10,000     |
|----------------|------------|------------|------------|------------|
| AE             | 2 members of 100 epochs | 5 × 100   | 10 × 100   | 100 × 100  |
|                | 2 members  | 5 members  | 10 members | 100 members|
| SAE            | 1 chain of 100 first epochs | 2(100 + 75 × 2) | 3(100 + 116 × 2) | 10(100 + 450 × 2) |
|                | 10 sequential trainings of 10 epochs | 32 members | 72 members | 910 members |

Table 5: Allocation of the computational budget for the UCI-Gap dataset.