SAE: Sequential Anchored Ensembles

Arnaud Delaunoy

Advisor: Gilles Louppe

December 9, 2021

University of Liège, Belgium

Bayesian deep learning

Notations:

- $\boldsymbol{D}$: Dataset
- $\boldsymbol{\theta}$: Neural network parameters
- $\boldsymbol{x}$: Inputs
- $\boldsymbol{y}$: Outputs

We want to compute $p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{D}) = \int p(\boldsymbol{y} \mid \boldsymbol{x}, \theta) p(\boldsymbol{\theta} \mid \boldsymbol{D}) d\,\boldsymbol{\theta}$,

where $p(\boldsymbol{\theta} \mid \boldsymbol{D}) = \frac{p(\boldsymbol{D} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\boldsymbol{D})}$.

Anchored ensembles [Pearce et al., 2020]

Training: Ensemble of N neural networks such that $\boldsymbol{\theta}^*_{1,..,N} \sim p(\boldsymbol{\theta} \mid \boldsymbol{D})$.

Prediction: $p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{D}) \simeq \frac{1}{N} \sum_{i=1}^{N} p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}^*_i)$.

Anchored ensembles

Idea:	Inject noise in the training procedure for the optima to be sampled from the Bayesian posterior

Anchored Ensembling

for i in $1, .., N$ **do**

$\boldsymbol{\theta}_{\text{anc},i} \sim p(\boldsymbol{\theta})$ (Sample anchor)

$\boldsymbol{\theta}_{\text{init},i} \leftarrow \text{init}()$ (Initialize NN)

$\boldsymbol{\theta}_i^* \leftarrow \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{D} \mid \boldsymbol{\theta}) p_{\text{anc},i}(\boldsymbol{\theta})$

end for

where $p_{\text{anc},i} = \mathcal{N}(\boldsymbol{\theta}_{\text{anc},i}, \boldsymbol{\Sigma}_{\text{prior}})$.

Anchored ensembles

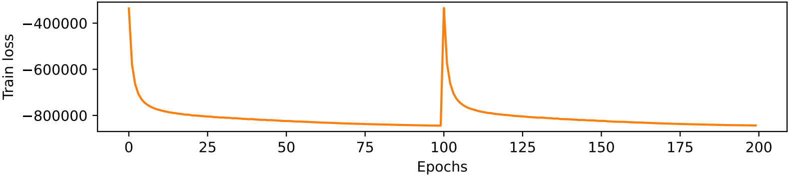
Hypotheses:

- Normal prior: $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_{\text{prior}}, \boldsymbol{\Sigma}_{\text{prior}})$
- Normal likelihood $p(\boldsymbol{D} \,|\, \boldsymbol{\theta})$ (also works for classification in practice)

If $\boldsymbol{\theta}_{\text{anc}} \sim p(\boldsymbol{\theta})$ then $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{D} \,|\, \boldsymbol{\theta}) p_{\text{anc}}(\boldsymbol{\theta}) \sim p(\boldsymbol{\theta} \,|\, \boldsymbol{D})$ (approximately).

Anchored ensembles

Training an ensemble is computationally expensive.



Sequential Anchored ensembles

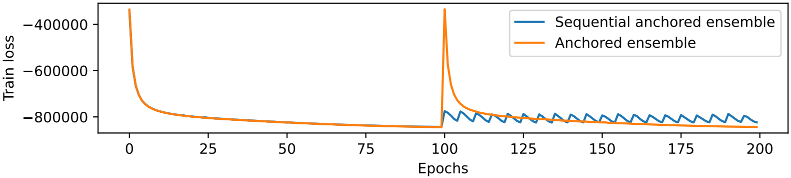
If $\boldsymbol{\theta}_{\text{anc},i}$ is close to $\boldsymbol{\theta}_{\text{anc},i-1}$, then $\boldsymbol{\theta}_i^*$ is close to $\boldsymbol{\theta}_{i-1}^*$

Sequential Anchored Ensembling (SAE)

$\boldsymbol{\theta}_{\text{anc},1} \sim p(\boldsymbol{\theta})$ (Sample first anchor)
$\boldsymbol{\theta}_{\text{init},1} \leftarrow \text{init}()$ (Initialize NN)
$\boldsymbol{\theta}_1^* \leftarrow \text{train}(\boldsymbol{\theta}_{\text{anc},1}; \boldsymbol{\theta}_{\text{init},1})$ (Long)
for i in $2, .., M$ **do**
    $\boldsymbol{\theta}_{\text{anc},i} \leftarrow \text{mcmc\_step}(\boldsymbol{\theta}_{\text{anc},i-1})$
    $\boldsymbol{\theta}_{\text{init},i} \leftarrow \boldsymbol{\theta}_{i-1}^*$
    $\boldsymbol{\theta}_i^* \leftarrow \text{train}(\boldsymbol{\theta}_{\text{anc},i}; \boldsymbol{\theta}_{\text{init},i})$ (Short)
end for

- Allow to build larger ensembles that AE
- SAE ensemble's members are correlated
- Can run SAE multiple times to benefit from different initializations

Sequential Anchored ensembles

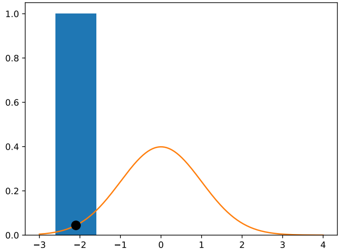


Guided-walk Metropolis-Hastings

For SAE to work well, we need:

- $\boldsymbol{\theta}_{\text{anc},i+1}$ close to $\boldsymbol{\theta}_{\text{anc},i}$ (short training)
- $\boldsymbol{\theta}_{\text{anc},(1,..,M)}$ covers $p(\boldsymbol{\theta})$ well

Guided walk Metropolis-Hastings [Gustafson, 1998]



Guided-walk Metropolis-Hastings

Guided walk Metropolis-Hastings

$y \leftarrow \theta_{\text{anc},i-1} + d_{i-1}|z|, \quad z \sim \mathcal{N}(0, \sigma_{\text{step}})$

$\alpha \leftarrow \min\left(\frac{p(y)}{p(\theta_{\text{anc},i-1})}, 1\right)$

$u \sim \mathcal{U}(0, 1)$

if $u < \alpha$ **then**

  $\theta_{\text{anc},i} \leftarrow y$

  $d_i \leftarrow d_{i-1}$

else

  $\theta_{\text{anc},i} \leftarrow \theta_{\text{anc},i-1}$

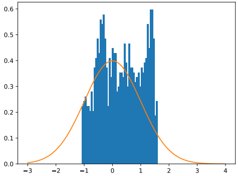
  $d_i \leftarrow -d_{i-1}$

end if

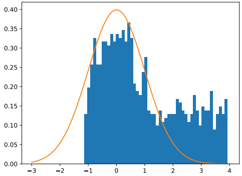
How to choose σ_{step} ?

- Should be as small as possible
- Should span the prior $\rightarrow$ we can verify this!

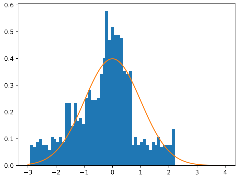
(a) $\sigma_{\text{step}} = 0.01$



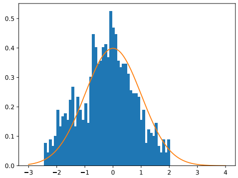
(b) $\sigma_{\text{step}} = 0.02$



(c) $\sigma_{\text{step}} = 0.03$



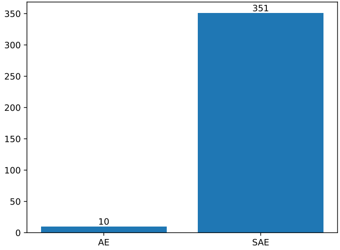
(d) $\sigma_{\text{step}} = 0.05$



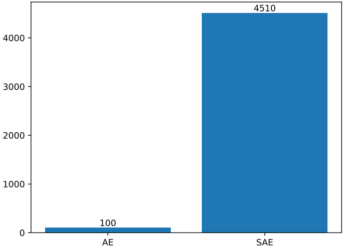
Order of magnitude

Number of members in the ensemble

(a) 1000 epochs



(b) $10,000$ epochs

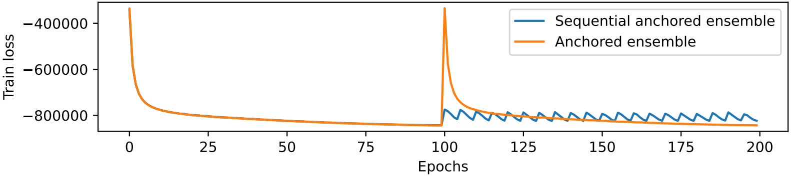


Results

		Cifar10 Resnet		Cifar10-C Alexnet		IMDB		DermaMNIST		UCI-Gap
		Ag.	TV	Ag.	TV	Ag.	TV	Ag.	TV	W_2
1000 epochs	AE	0.849	0.201	0.726	0.262	**0.892**	**0.109**	0.877	0.104	**−0.148**
	SAE	**0.856**	**0.176**	**0.772**	**0.212**	0.887	0.110	**0.880**	**0.098**	−0.178
10,000 epochs	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**	−0.185

Summary

Mail: a.delaunoy@uliege.be **Twitter:** @ArnaudDelaunoy



References

- Tim Pearce, Felix Leibfried, and Alexandra Brintrup. Uncertainty in neural networks: Approximatelybayesian ensembling. InInternational conference on artificial intelligence and statistics, pages234–244. PMLR, 2020.
- Paul Gustafson. A guided walk metropolis algorithm.Statistics and computing, 8(4):357–364, 1998.