BMRS: Bayesian Model Reduction for Structured Pruning

Dustin Wright, Christian Igel, Raghavendra Selvan

dw@di.ku.dk; **I'm on the job market!** ·······▸

BACKGROUND: We want to balance **accuracy** and **compression** in a **principled** way. We do so with variational inference and Bayesian Model Reduction (BMR).

BMRS in a nutshell

- Apply multiplicative noise to network structures (Neklyudov et al. 2017)

$$\boldsymbol{h}_i = \theta_i \cdot (\boldsymbol{w}_i \boldsymbol{h}_{i-1})$$
$$q_\phi(\theta_i) = \mathrm{LogN}_{[a,b]}(\theta_i | \mu_i, \sigma_i^2)$$
$$p(\theta_i) = \mathrm{LogU}_{[a,b]}(\theta_i)$$

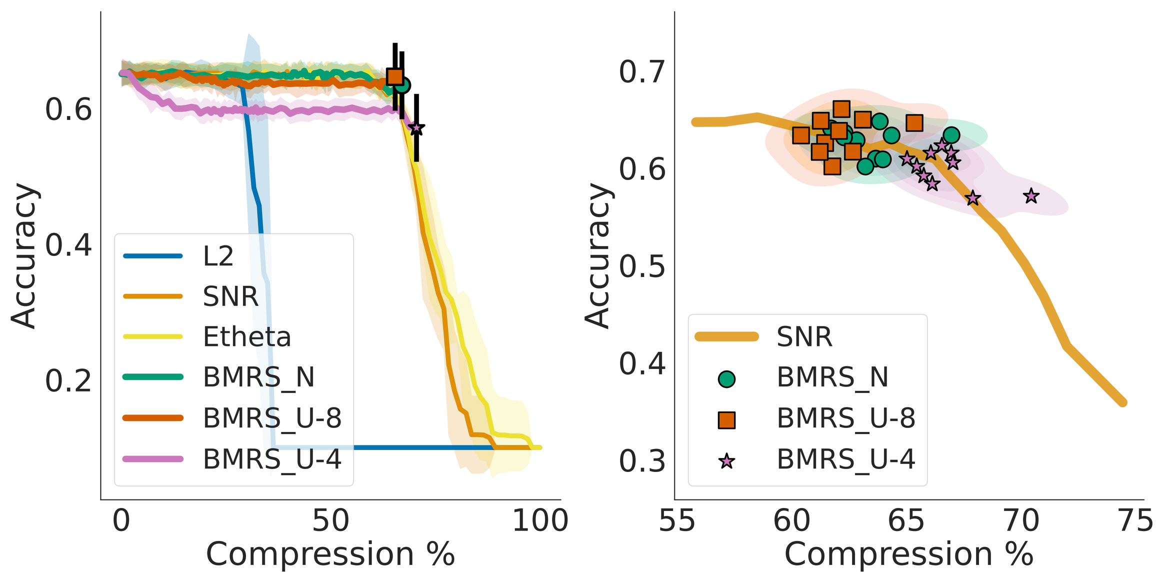
- Use Bayesian Model Reduction (BMR) to find what to prune

$$\Delta F \triangleq \log\frac{\tilde{p}(D)}{p(D)} = \log\mathbb{E}_{\tilde{p}}\left[\frac{q_\phi(\theta)}{p(\theta)}\right]$$

RESULTS

- BMRS finds near-optimal point w/ post-training pruning
- Good compression + maintains accuracy w/ continuous pruning

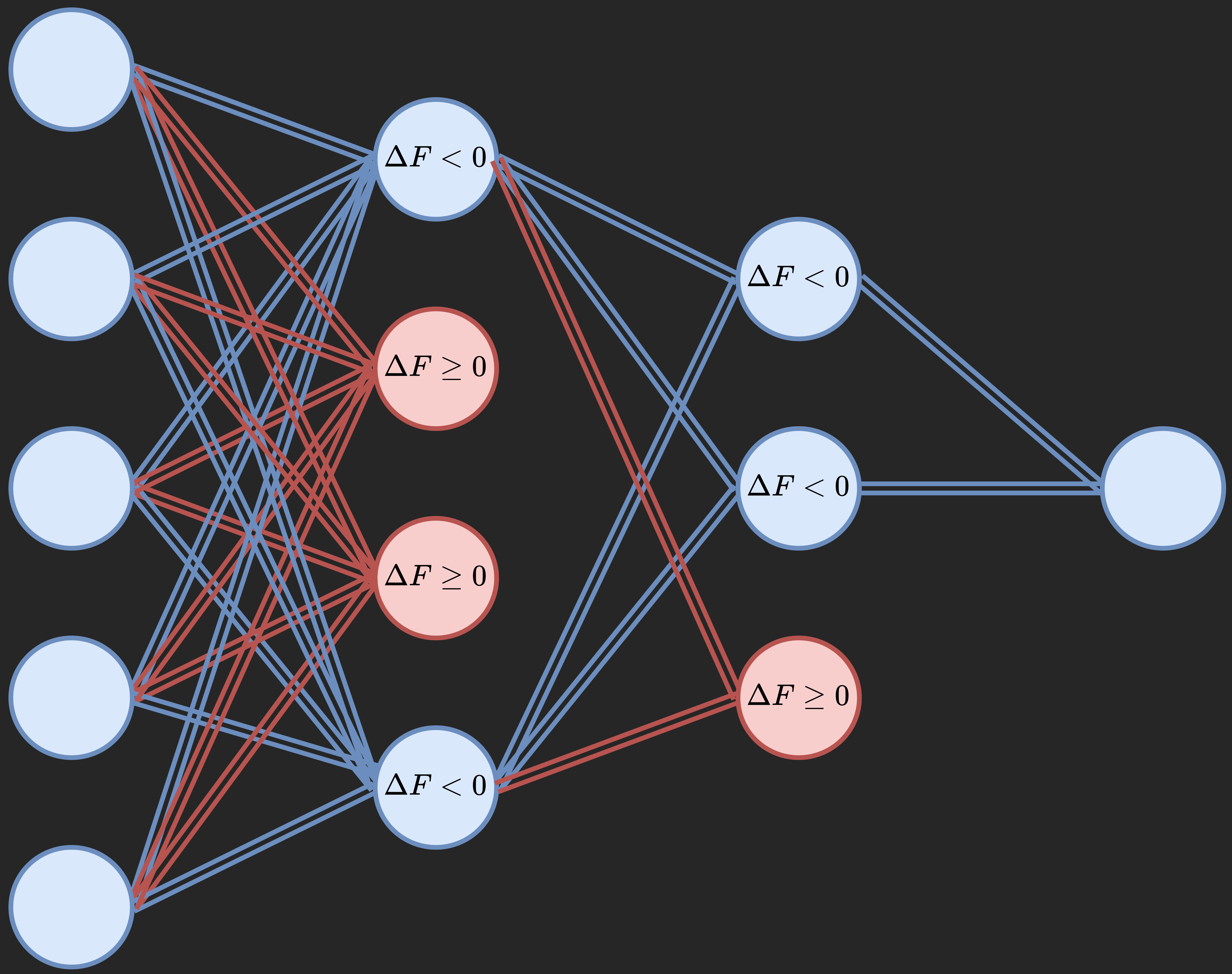
CNN performance on CIFAR10



Continuous pruning results

Pruning Method	MNIST Comp. (%)	MNIST Acc.	Fash-MNIST Comp. (%)	Fash-MNIST Acc.	CIFAR10 Comp. (%)	CIFAR10 Acc.
			MLP			
None	0.00 ± 0.00	97.43 ± 0.14	0.00 ± 0.00	88.17 ± 0.20	0.00 ± 0.00	44.94 ± 0.40
L2	43.11 ± 2.06	10.39 ± 0.32	87.86 ± 2.27	18.23 ± 10.22	42.89 ± 2.64	10.00 ± 0.00
$E[\theta]$	52.08 ± 1.71	96.88 ± 0.15	91.76 ± 0.81	85.59 ± 0.26	77.99 ± 1.54	43.39 ± 0.46
SNR	58.57 ± 2.01	96.92 ± 0.08	99.83 ± 0.00	10.00 ± 0.00	75.93 ± 1.26	43.97 ± 0.46
$\mathrm{BMRS}_{\mathcal{N}}$	48.86 ± 1.32	96.95 ± 0.19	93.20 ± 0.66	84.99 ± 0.35	76.36 ± 1.08	43.59 ± 0.29
$\mathrm{BMRS}_{\mathcal{U}}$-8	48.73 ± 1.90	96.93 ± 0.16	93.02 ± 0.81	85.01 ± 0.32	77.17 ± 0.98	43.45 ± 0.42
$\mathrm{BMRS}_{\mathcal{U}}$-4	54.47 ± 1.74	$\mathbf{96.99 \pm 0.13}$	91.57 ± 0.71	$\mathbf{85.79 \pm 0.34}$	76.63 ± 0.94	$\mathbf{44.06 \pm 0.40}$
			Lenet5			
None	0.00 ± 0.00	99.07 ± 0.09	0.00 ± 0.00	89.16 ± 0.27	0.00 ± 0.00	67.62 ± 0.77
L2	83.42 ± 1.92	11.35 ± 0.00	83.62 ± 1.69	10.00 ± 0.00	52.29 ± 2.18	10.00 ± 0.00
$E[\theta]$	88.29 ± 1.00	51.30 ± 41.12	89.71 ± 0.56	50.93 ± 33.45	66.19 ± 1.36	65.83 ± 0.90
SNR	92.66 ± 5.77	62.70 ± 41.93	98.47 ± 3.45	17.01 ± 21.03	70.29 ± 2.02	$\mathbf{67.68 \pm 0.52}$
$\mathrm{BMRS}_{\mathcal{N}}$	86.90 ± 1.15	95.59 ± 0.94	88.02 ± 1.00	77.90 ± 2.44	62.87 ± 1.64	66.14 ± 0.70
$\mathrm{BMRS}_{\mathcal{U}}$-8	86.11 ± 1.37	95.27 ± 1.02	87.61 ± 0.72	77.23 ± 3.49	62.54 ± 1.49	66.28 ± 1.07
$\mathrm{BMRS}_{\mathcal{U}}$-4	87.58 ± 1.01	$\mathbf{96.66 \pm 0.59}$	88.72 ± 0.73	$\mathbf{81.10 \pm 1.50}$	68.07 ± 1.95	67.66 ± 0.59

Bayesian structured pruning without threshold tuning



Structures are pruned using an efficient calculation based on the statistics of the variational distribution

Paper and code

Also read our *critical* perspective of AI efficiency

Two variants

- $\mathrm{BMRS}_{\mathcal{N}}$ ΔF:

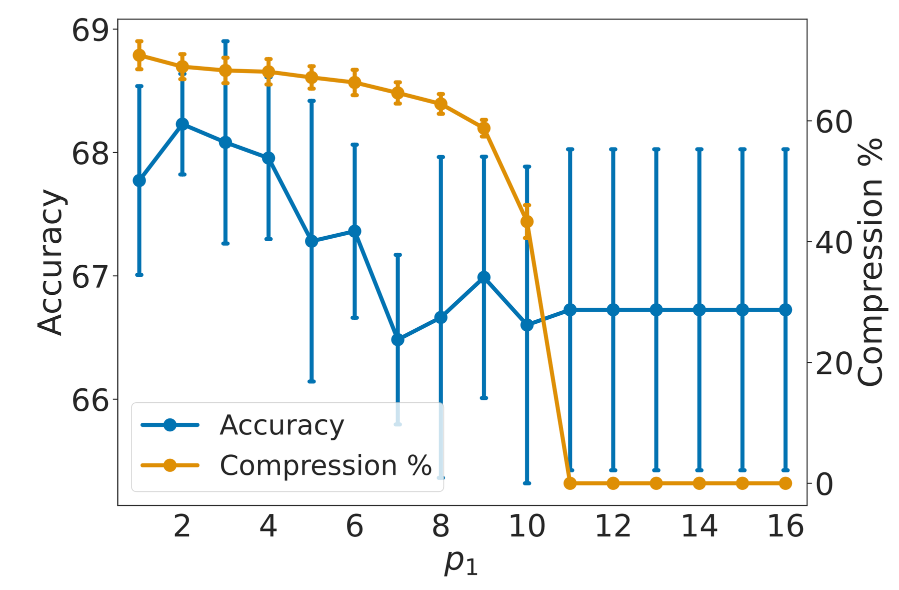
$$\frac{1}{2}\log\frac{\tilde{\sigma}_q^2}{2\pi\tilde{\sigma}_p^2\sigma_q^2} + \log\frac{Z_{\tilde{q}}(\log b - \log a)}{Z_{\tilde{p}}Z_q} - \frac{1}{2}\left(\frac{\mu_q^2}{\sigma_q^2} + \frac{\tilde{\mu}_p^2}{\tilde{\sigma}_p^2} - \frac{\tilde{\mu}_q^2}{\tilde{\sigma}_q^2}\right)$$

- $\mathrm{BMRS}_{\mathcal{U}}$ ΔF:

$$\frac{\log b' - \log a'}{\log b - \log a} \leq q_\phi(a' \leq \theta \leq b')$$

- $\mathrm{BMRS}_{\mathcal{N}}$ is **thresholdless**
- $\mathrm{BMRS}_{\mathcal{U}}$ can be **tuned**…

CNN performance on CIFAR10



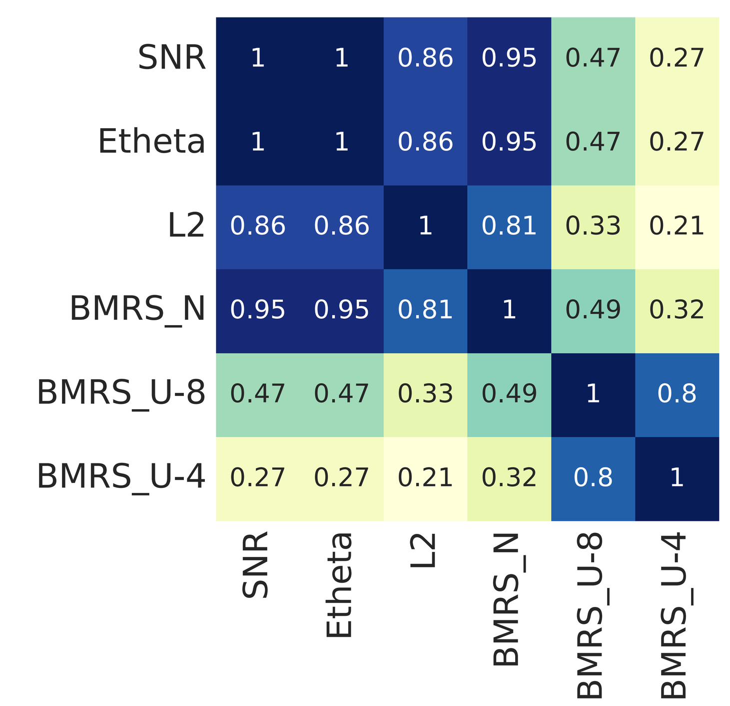
- …and has a connection to **floating point precision**

Mantissas: $m \sim (m\log B)^{-1}$, $\frac{1}{B} \leq m \leq 1$

$$\tilde{p}(\theta): \begin{cases} \left(\theta\log\frac{2^{p_2}}{2^{p_1}}\right)^{-1}, & \frac{1}{2^{p_2}} \leq \theta \leq \frac{1}{2^{p_1}} \\ 0, & \text{otherwise} \end{cases}$$

- Different pruning functions are learned

Spearman rank correlation of pruning methods



University of Copenhagen