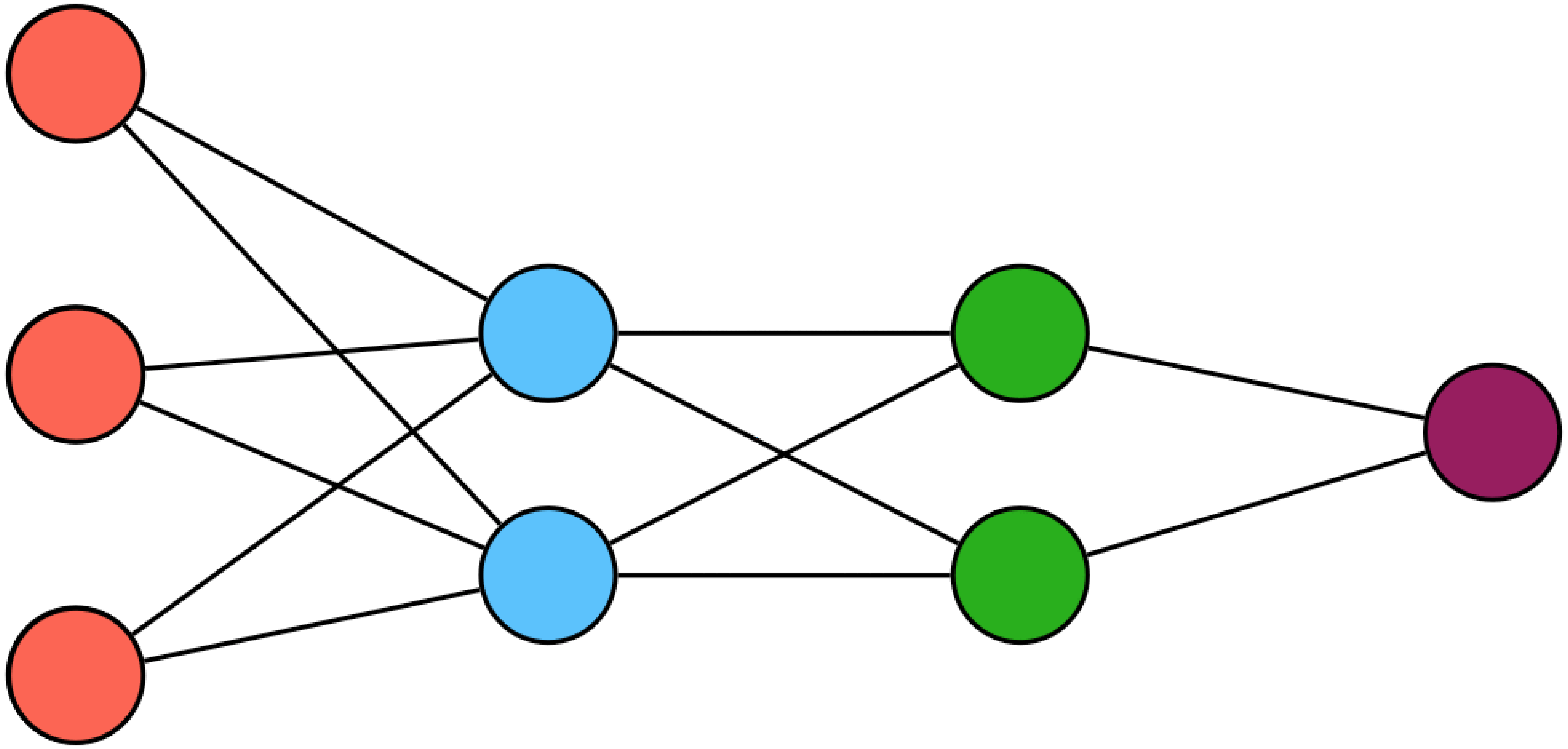
Energy-Aware Neural Architecture Optimization With Splitting Steepest Descent

Dilin Wang[1], Lemeng Wu[1], Meng Li[2], Vikas Chandra[2], Qiang Liu[1]

[1] UT Austin [2] Facebook

EMC2 Workshop @ NeurIPS 2019

Splitting yields adaptive net structure optimization



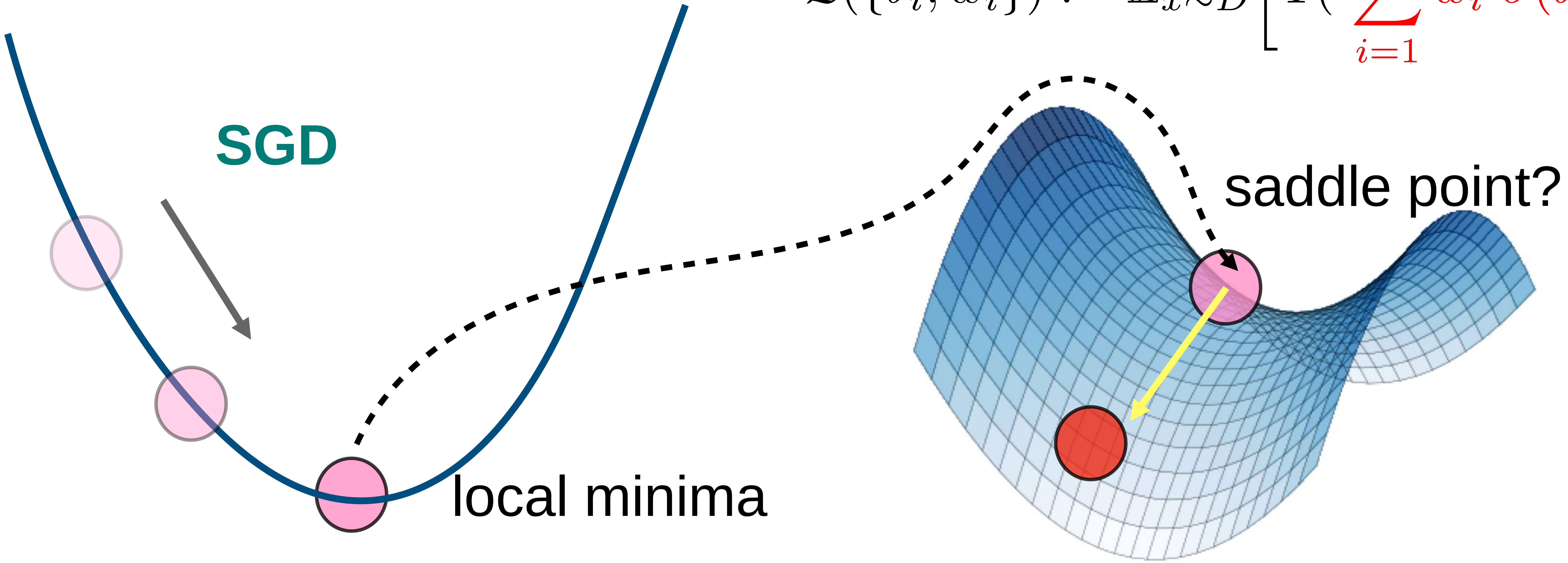
Questions

- Why splitting?
- What neurons should be split first?
- How to split a neuron optimally?

Intuition: escaping local minima

- Splitting θ into m copies $\{w_i, \theta_i\}_{i=1}^{m}$:

$$\mathcal{L}(\{\theta_i, w_i\}) := \mathbb{E}_{x\sim D}\left[\Phi(\ \sum_{i=1}^{m} w_i\ \sigma(\theta_i, x)\ )\right]$$



- A simple network:

$$\mathcal{L}(\theta) := \mathbb{E}_{x\sim D}\left[\Phi(\ \sigma(\theta, x)\ )\right].$$

- *Smooth loss change:*

$$\sum_{i=1}^{m} w_i = 1,\ \ ||\theta_i - \theta||_2 \leq \epsilon$$

Splitting Steepest Descent

- How to choose m and $\{\theta_i, w_i\}$ optimally?

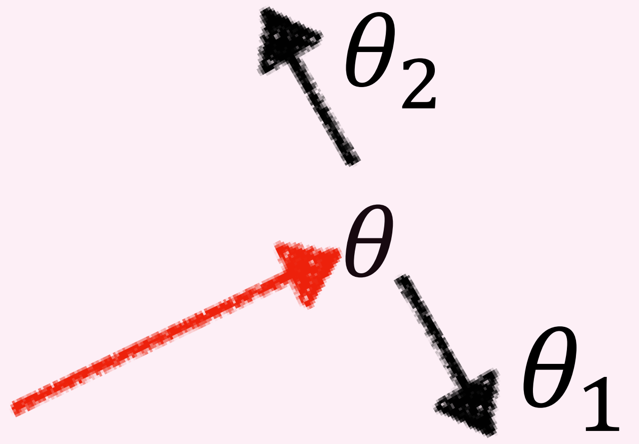
$$\min_{m,\{\theta_i,w_i\}_{i=1}^m} \left\{ \mathcal{L}(\{\theta_i, w_i\}) - \mathcal{L}(\theta) \qquad \text{s.t.} \quad ||\theta_i - \theta||_2 \le \epsilon, \ \sum_{i=1}^m w_i = 1, \quad w_i > 0, \ \forall \ i \right\}.$$

Splitting-index, minimum eigenvalue

$$= \frac{\epsilon^2}{2} \underbrace{\min\left\{ \lambda_{\min}(S(\theta)), \quad 0 \right\}}_{\text{CLOSED-FORM}} + \mathcal{O}(\epsilon^3) \ \text{ with } \ S(\theta) = \mathbb{E}_{x\sim D}\left[\nabla_\sigma \Phi(\sigma(\theta, x)) \ \nabla^2_{\theta\theta}\sigma(\theta, x) \right],$$

Splitting-matrix

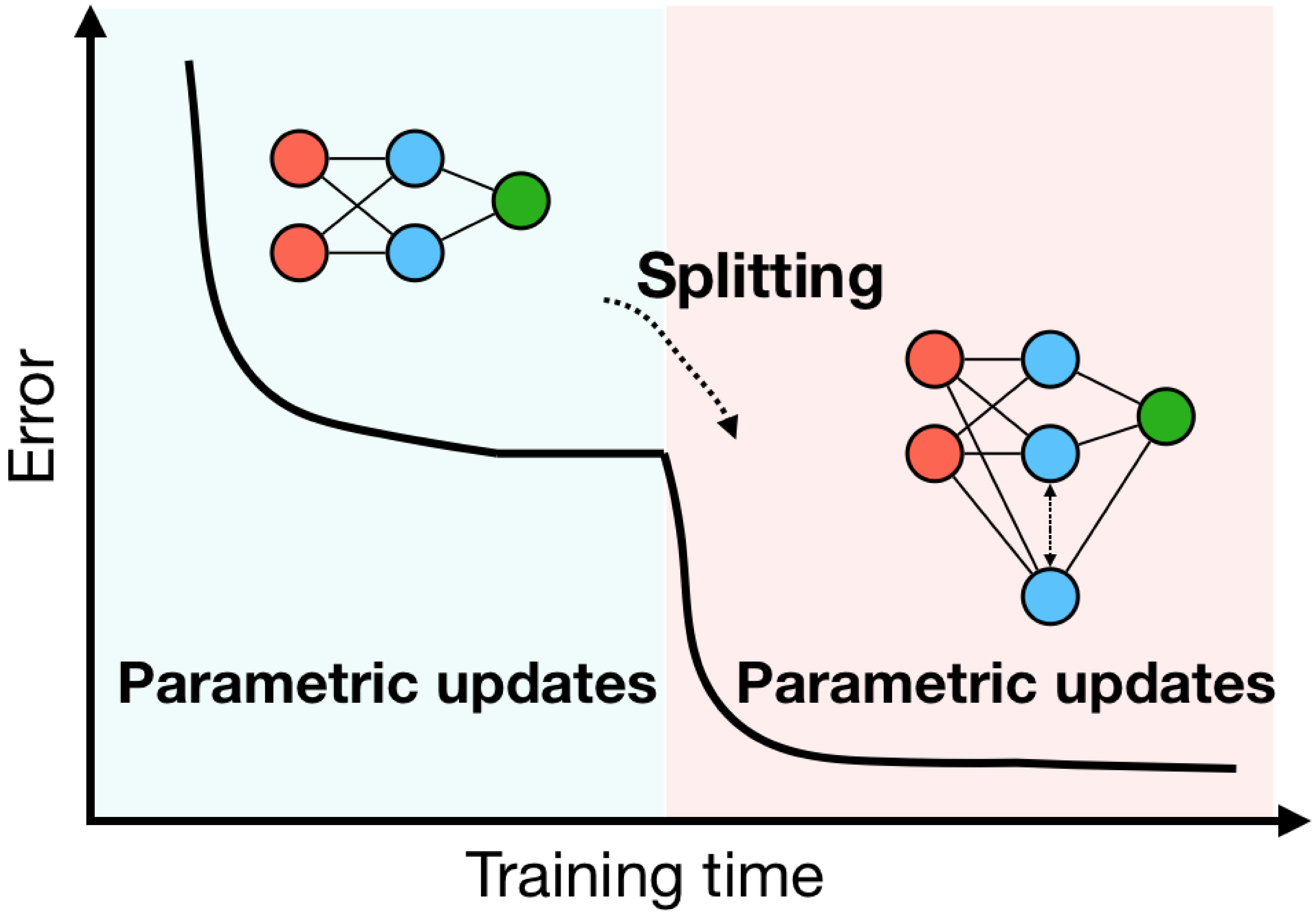
- Optimal splitting strategy

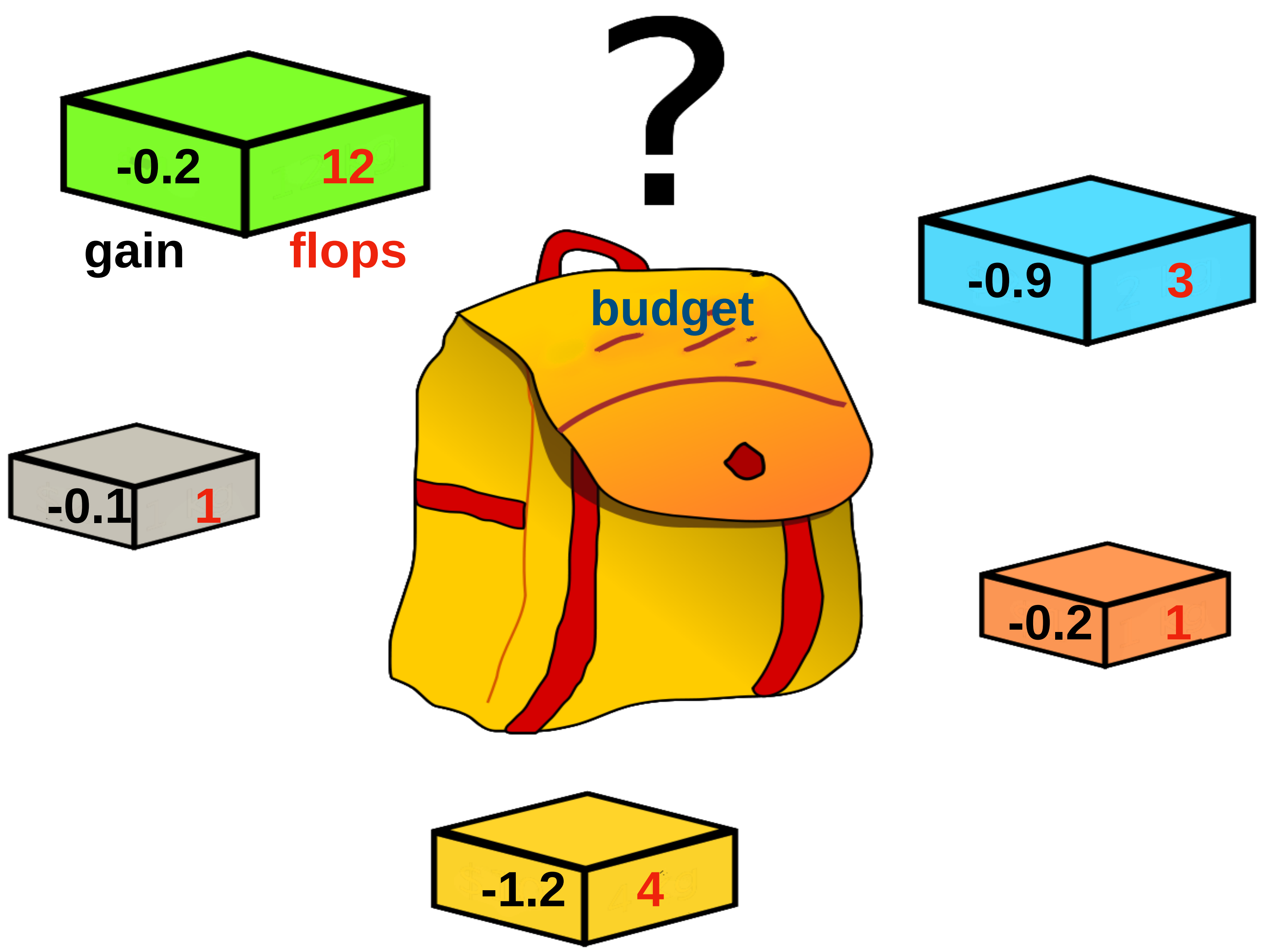


$$\lambda_{\min} S(\theta) \ge 0, \quad \text{no splitting}$$

$$\lambda_{\min} S(\theta) < 0, \quad m = 2, \ \theta_1 = \theta + \epsilon v_{\min}(S(\theta)), \quad \theta_2 = \theta - \epsilon v_{\min}(S(\theta)), \quad w_1 = w_2 = 1/2.$$

Our Algorithm





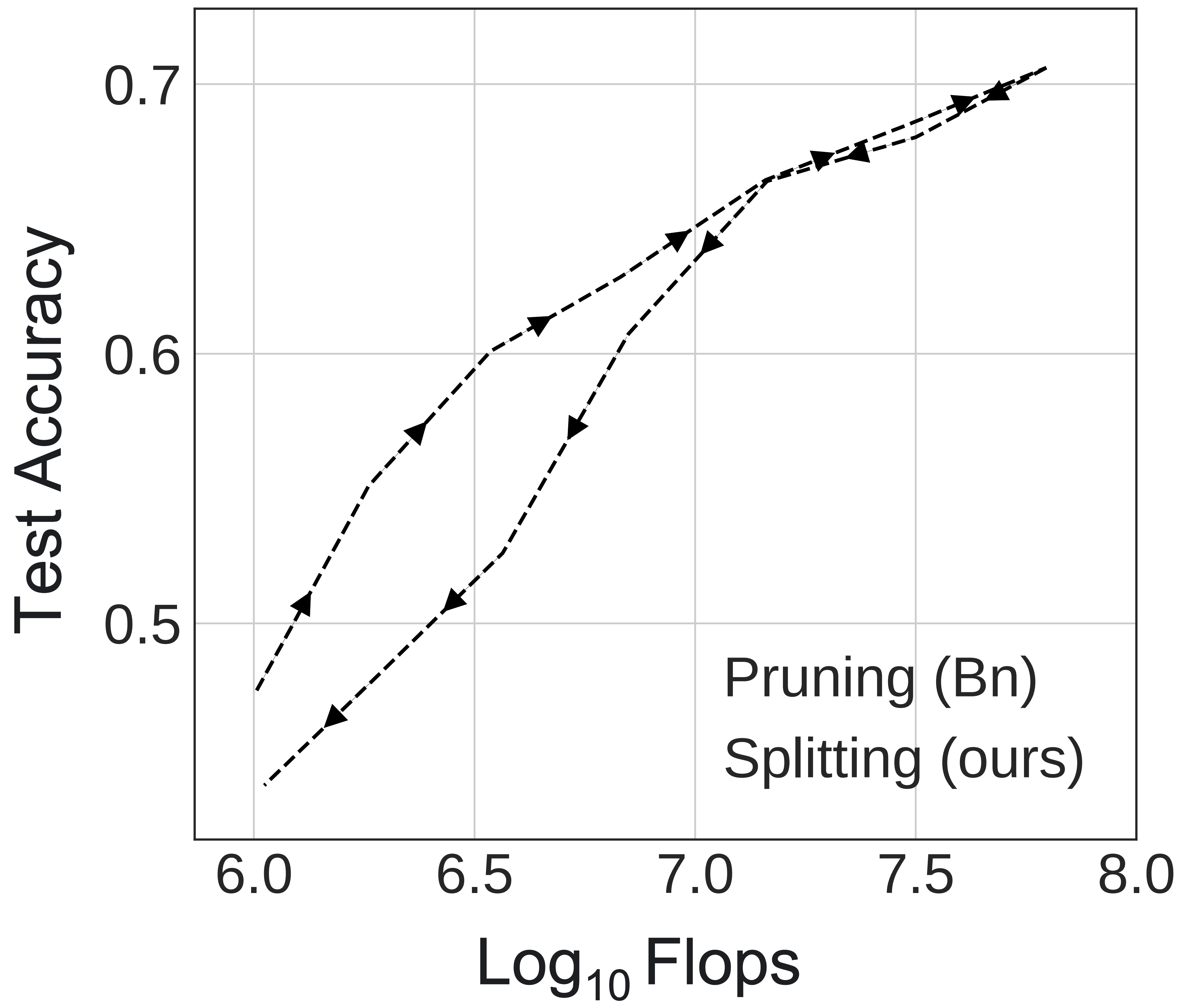
$$\min_{\beta} \quad \sum_{\ell=1}^{n} \beta_\ell \underbrace{\lambda_{\min}(S(\theta^\ell))}_{gain}$$

$$s.t. \quad \beta_\ell \in \{0,1\}$$

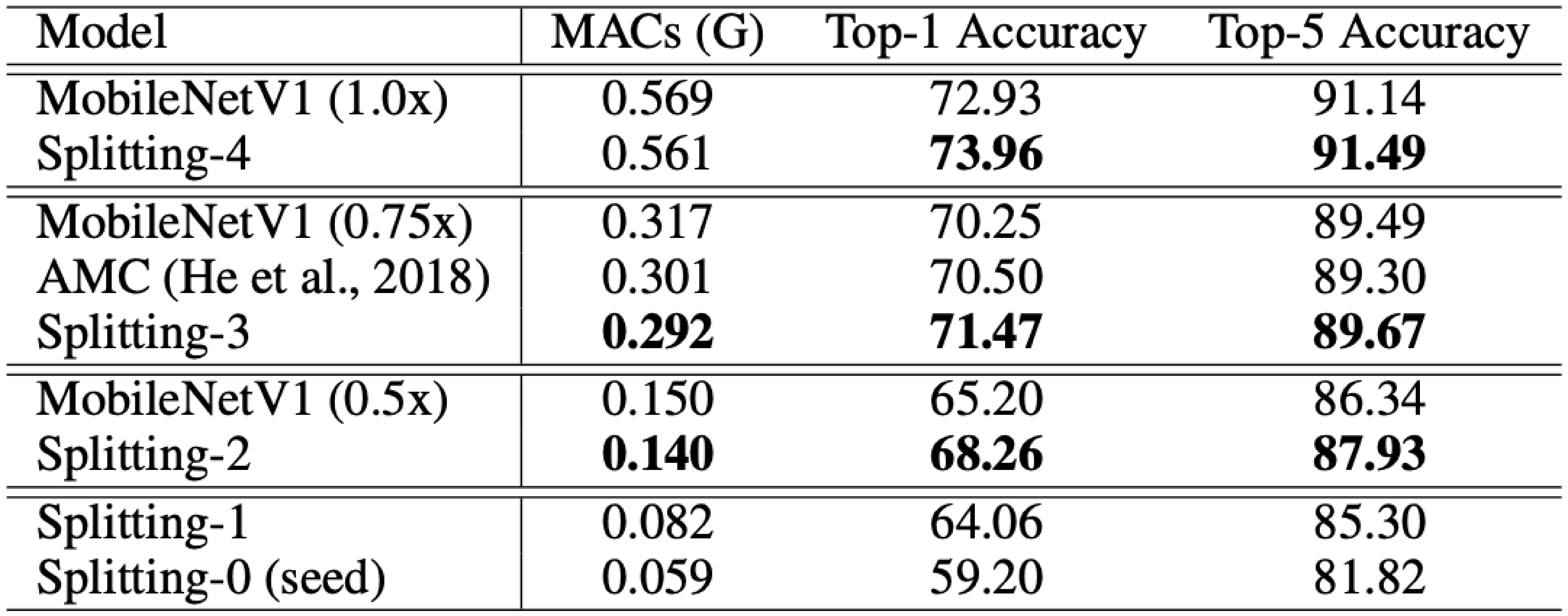
$$\sum_{\ell=1}^{n} \underbrace{e_\ell}_{flops} \beta_\ell \leq \text{budget}$$

Image Classification Results using MobileNetV1

Results on CIFAR100



Results on ImageNet



Model	MACs (G)	Top-1 Accuracy	Top-5 Accuracy
MobileNetV1 (1.0x)	0.569	72.93	91.14
Splitting-4	0.561	**73.96**	**91.49**
MobileNetV1 (0.75x)	0.317	70.25	89.49
AMC (He et al., 2018)	0.301	70.50	89.30
Splitting-3	**0.292**	**71.47**	**89.67**
MobileNetV1 (0.5x)	0.150	65.20	86.34
Splitting-2	**0.140**	**68.26**	**87.93**
Splitting-1	0.082	64.06	85.30
Splitting-0 (seed)	0.059	59.20	81.82

Thank You!