Block Mean Approximation for Efficient Second Order Optimization

Yao Lu, Mehrtash Harandi, Richard Hartley, Razvan Pascanu

ANU & Data61 & Google DeepMind

November 25, 2023

Optimization

For $f(\boldsymbol{\theta}) : \mathbb{R}^n \to \mathbb{R}$

$$\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$$

Optimization

For $f(\boldsymbol{\theta}) : \mathbb{R}^n \to \mathbb{R}$

$$\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$$

First order method

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$$

Second order method

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \mathbf{G}^{-1} \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$$

Optimization

For $f(\boldsymbol{\theta}) : \mathbb{R}^n \to \mathbb{R}$

$$\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$$

First order method

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$$

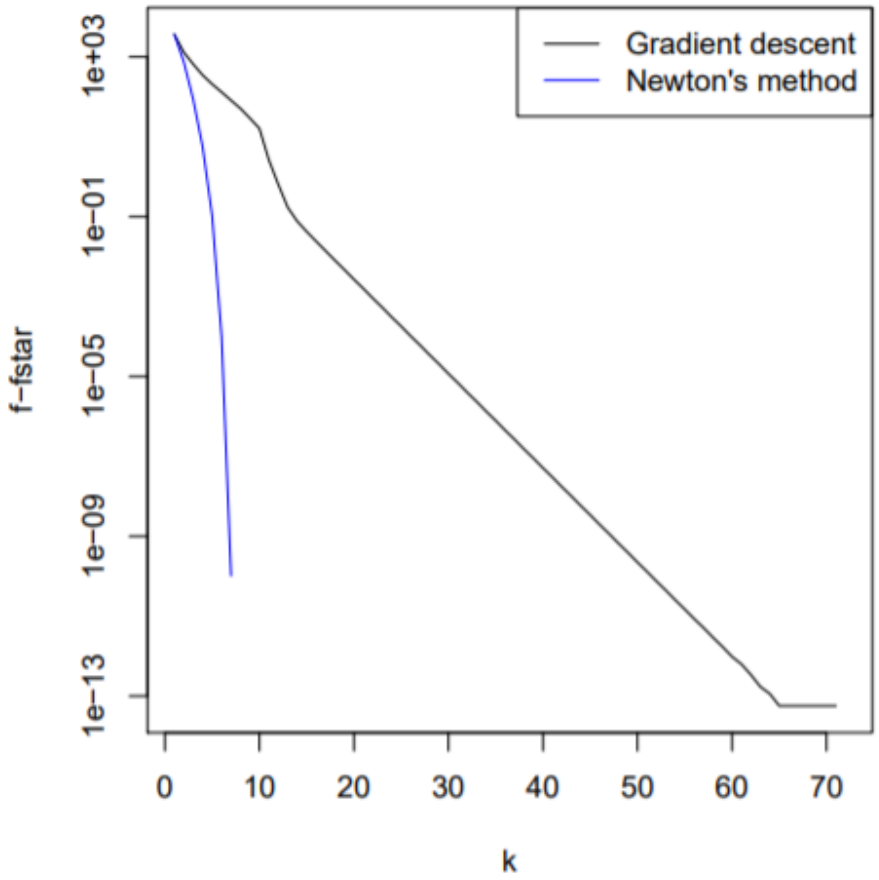
Second order method

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \mathbf{G}^{-1} \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$$

- Newton method
- Natural gradient
- Adaptive gradient

Example: logistic regression

Logistic regression example, with $n = 500$, $p = 100$: we compare gradient descent and Newton's method, both with backtracking



Second Order Optimization

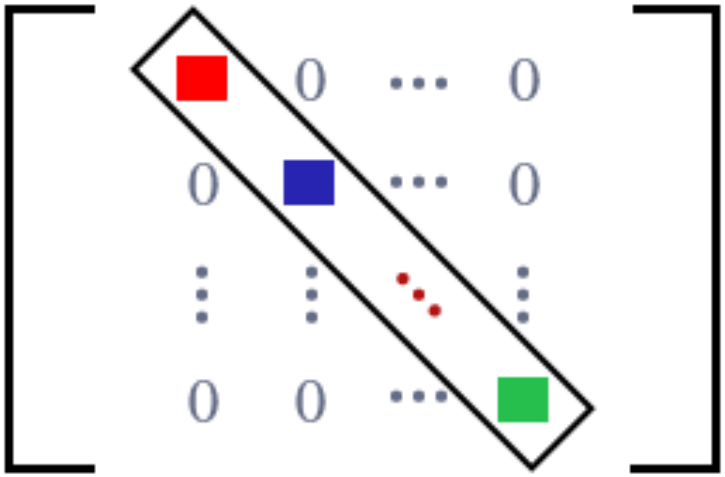
$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \mathbf{G}^{-1} \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$$

Compute $\mathbf{G}^{-1}$ is expensive.

Approximation is needed.

Approximation of $\mathbf{G}$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \mathbf{G}^{-1} \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$$

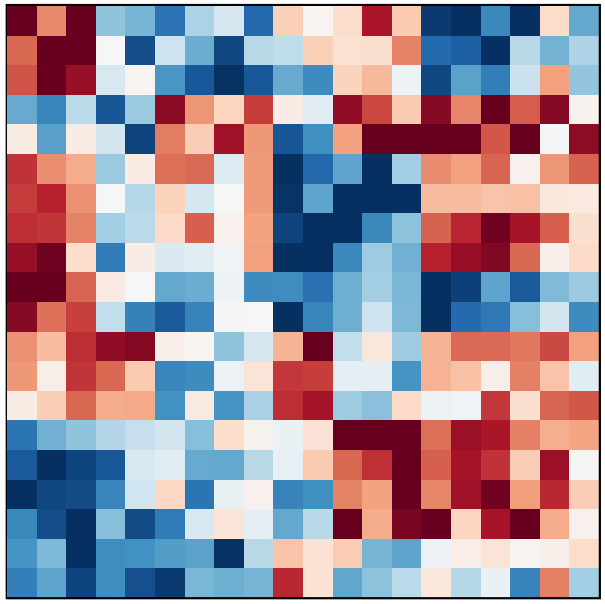


Approximation of $\mathbf{G}$

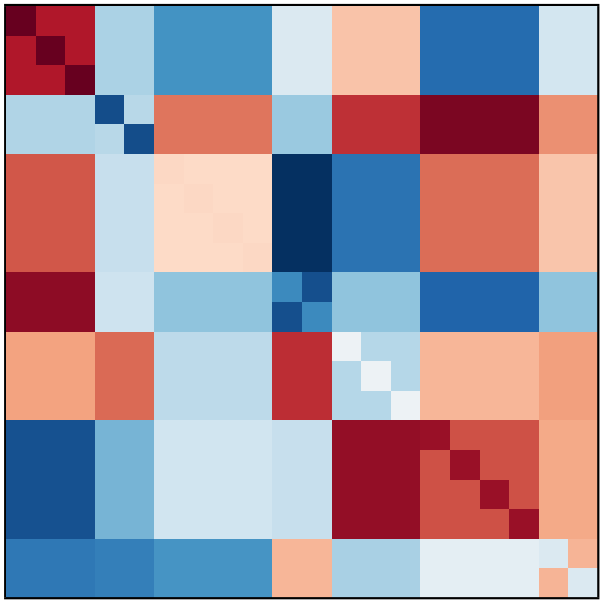
$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \mathbf{G}^{-1} \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$$

- Diagonal
- Block diagonal
- Low rank
- Kronecker product

Block Mean Approximation

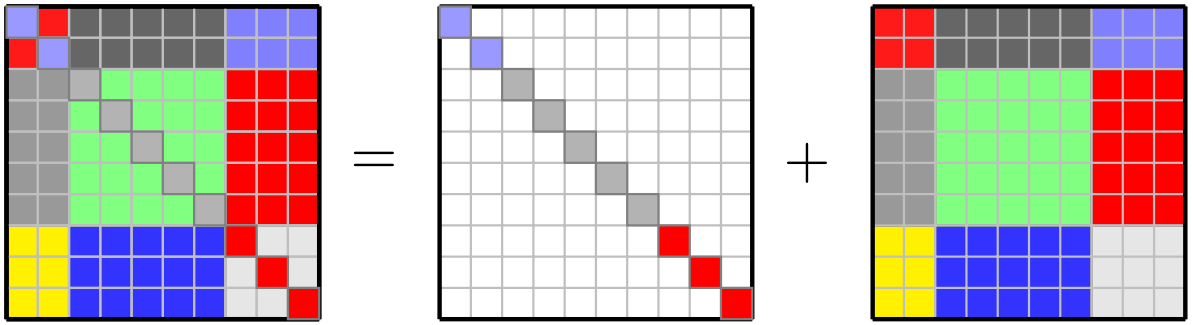


(a) Original

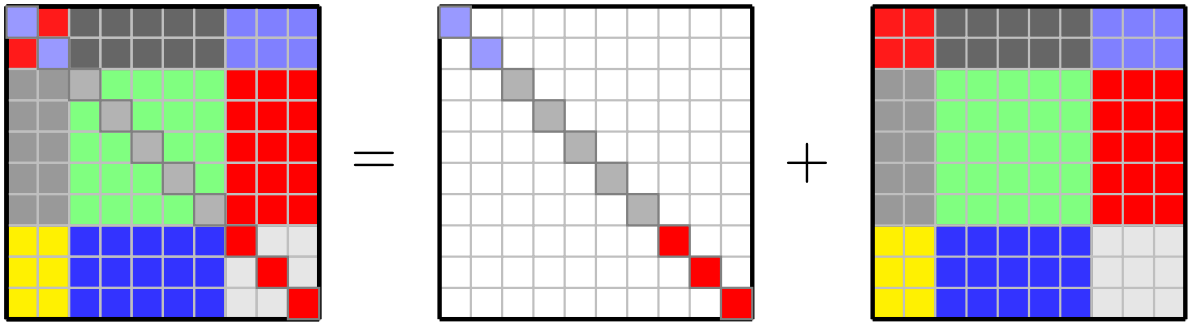


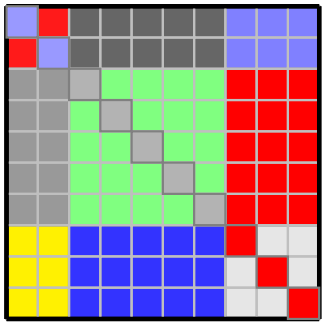
(b) Approximate

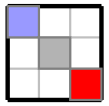
Block Mean Approximation



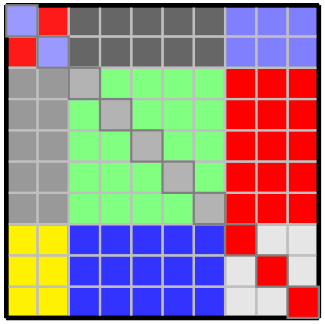
Block Mean Approximation



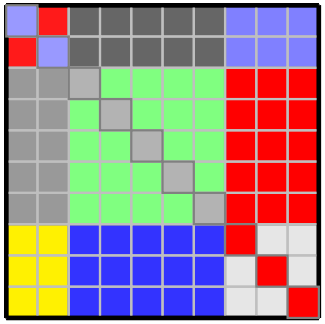


can be represented by  and  with the size of each block.

Block Mean Approximation

The inverse of  has the same block mean structure.

Block Mean Approximation

The inverse of  has the same block mean structure.

It can be computed by inverting a matrix of size  and 

Theorem

For the non-singular matrix $\bar{\mathbf{\Lambda}} + \bar{\mathbf{B}}$*, where* $\bar{\mathbf{\Lambda}}$ *and* $\bar{\mathbf{B}}$ *are the diagonal and full expansion of* $\mathbf{\Lambda}$ *and* $\mathbf{B}$ *with respect to the partition vector* $\mathbf{s}$*,*

$$(\bar{\mathbf{\Lambda}} + \bar{\mathbf{B}})^{-1} = \bar{\mathbf{\Lambda}}^{-1} + \bar{\mathbf{D}} \tag{1}$$

where $\bar{\mathbf{D}}$ *is the full expansion matrix with partition vector* $\mathbf{s}$ *of*

$$\mathbf{D} = (\mathbf{\Lambda S} + \mathbf{SBS})^{-1} - (\mathbf{\Lambda S})^{-1} \tag{2}$$

where $\mathbf{S} = \mathrm{diag}(\mathbf{s})$*.*

Theorem

For the non-singular matrix $\bar{\mathbf{\Lambda}} + \bar{\mathbf{B}}$*, where* $\bar{\mathbf{\Lambda}}$ *and* $\bar{\mathbf{B}}$ *are the diagonal and full expansion of* $\mathbf{\Lambda}$ *and* $\mathbf{B}$ *with respect to the partition vector* $\mathbf{s}$*,*

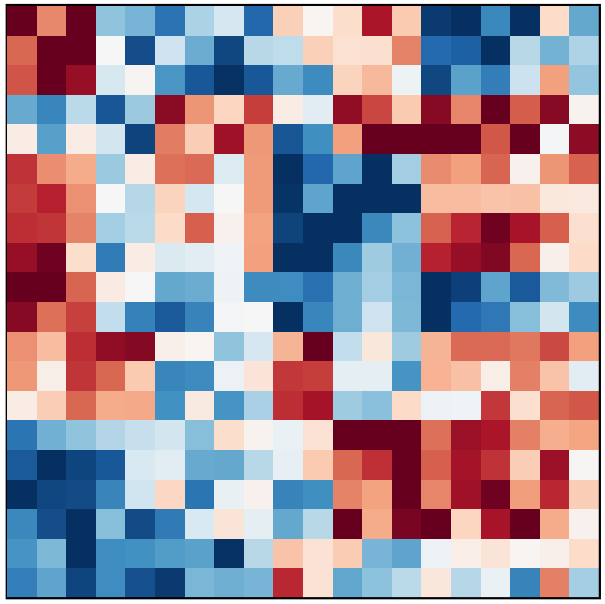
$$(\bar{\mathbf{\Lambda}} + \bar{\mathbf{B}})^{-\frac{1}{2}} = \bar{\mathbf{\Lambda}}^{-\frac{1}{2}} + \bar{\mathbf{D}} \tag{3}$$

where $\bar{\mathbf{D}}$ *is the full expansion matrix with partition vector* $\mathbf{s}$ *of*

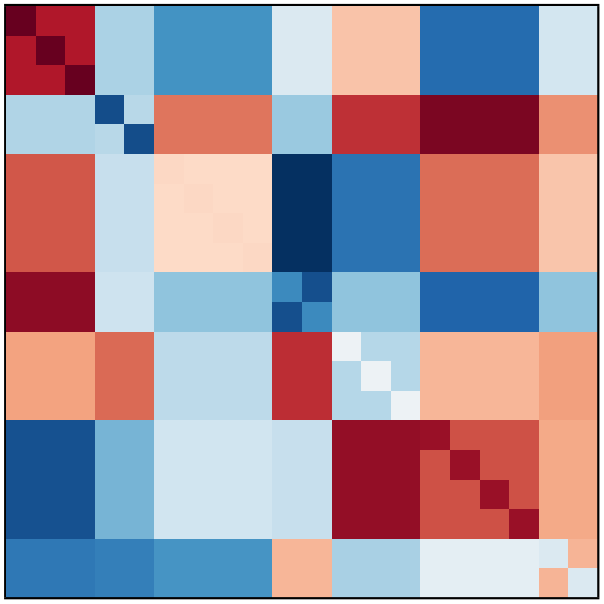
$$\mathbf{D} = \mathbf{S}^{-\frac{1}{2}} \left[(\mathbf{\Lambda} + \mathbf{S}^{\frac{1}{2}} \mathbf{B} \mathbf{S}^{\frac{1}{2}})^{-\frac{1}{2}} - \mathbf{\Lambda}^{-\frac{1}{2}} \right] \mathbf{S}^{-\frac{1}{2}} \tag{4}$$

where $\mathbf{S} = \mathrm{diag}(\mathbf{s})$*.*

Block Mean Approximation



(c) Original



(d) Approximate

Optimal Block Mean Approximation

Proposition

The optimal block mean approximation of $\mathbf{M}$ *with the partition vector* $\mathbf{s}$ *according to the Frobenius norm*

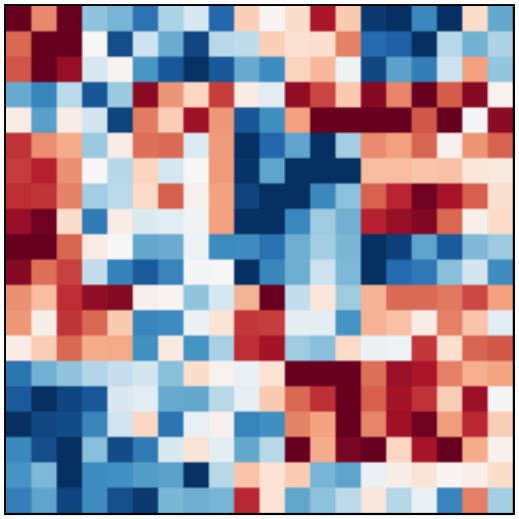
$$\min_{\bar{\mathbf{\Lambda}},\bar{\mathbf{B}}} \|\bar{\mathbf{\Lambda}} + \bar{\mathbf{B}} - \mathbf{M}\|_F^2 \tag{5}$$

is given by

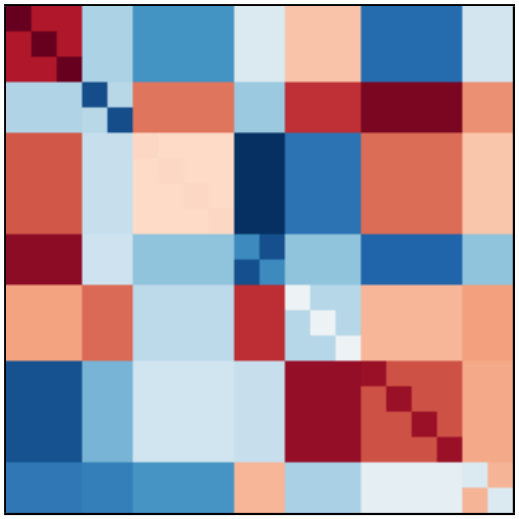
$$b_{ij} = \begin{cases} 0, & i = j, s_i = 1, \\ \frac{\sum_{mn} \mathbf{M}^{ii}_{mn} - \sum_m \mathbf{M}^{ii}_{mm}}{s_i(s_i - 1)}, & i = j, s_i \neq 1, \\ \frac{\sum_{mn} \mathbf{M}^{ij}_{mn}}{s_i s_j}, & i \neq j, \end{cases} \tag{6}$$

$$\lambda_i = \frac{1}{s_i} \sum_m \mathbf{M}^{ii}_{mm} - b_{ii} \ . \tag{7}$$

Block Mean Approximation



(e) Original



(f) Approximate

For neural nets, each block can represent the weights in a layer.

AdaGrad

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \mathbf{H}_t^{-1/2} \mathbf{g}_t$$

$$\widehat{\mathbf{H}}_t = \mathbf{Z}_t \mathbf{F}_t \mathbf{Z}_t \approx \mathbf{H}_t$$

where $\mathbf{Z}_t$ is diagonal and $\mathbf{F}_t$ is a block mean approximation matrix.

Experiments

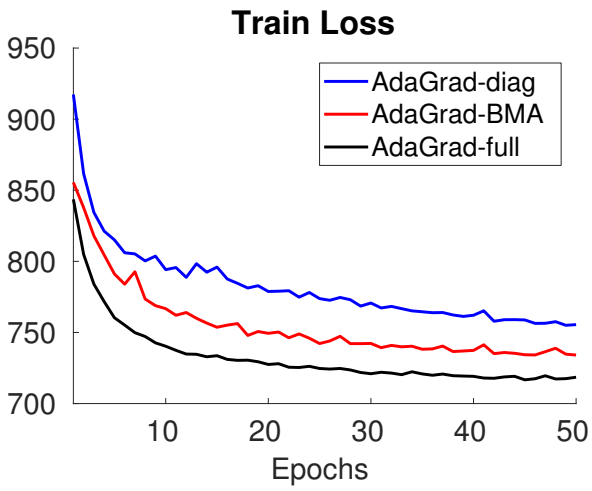
Table 1: Small model

Conv 3x3, 3
Max Pooling 2x2
Conv 3x3, 3
Max Pooling 2x2
Conv 3x3, 3
Max Pooling 2x2
Conv 3x3, 3
Max Pooling 2x2
Fully Connected, 10
Softmax, 10

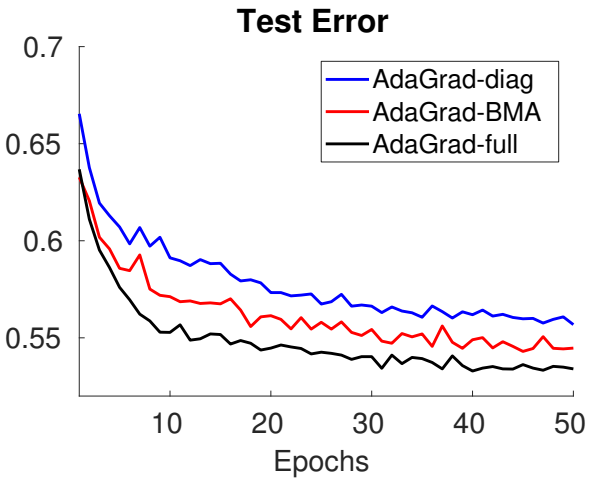
Table 2: Large model

Conv 3x3, 32
Conv 3x3, 32
Conv 3x3, 32
Conv 3x3, 32
Max Pooling 2x2
Conv 3x3, 32
Conv 3x3, 32
Conv 3x3, 32
Conv 3x3, 32
Max Pooling 2x2
Conv 3x3, 32
Conv 3x3, 32
Conv 3x3, 32
Conv 3x3, 32
Max Pooling 2x2
Conv 3x3, 32
Conv 3x3, 32
Conv 3x3, 32
Conv 3x3, 32
Max Pooling 2x2
Fully Connected, 10
Softmax, 10

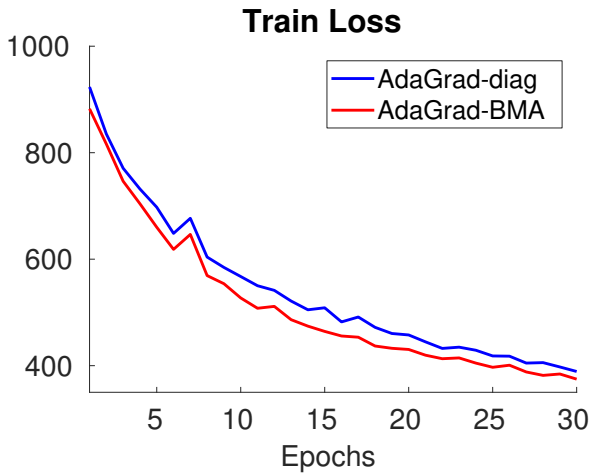
Experiments



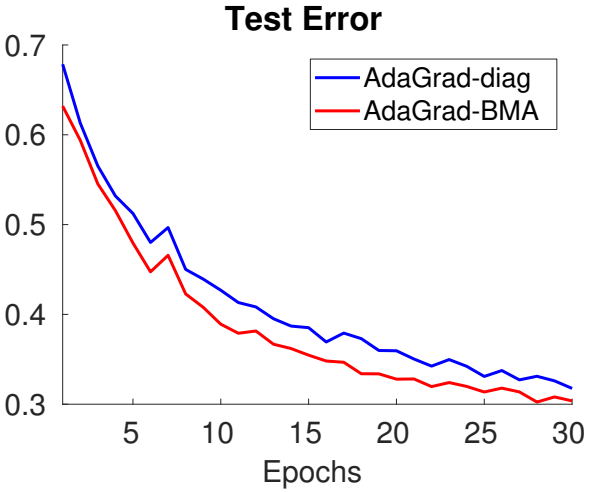
(g) CIFAR-10, small model



(h) CIFAR-10, small model



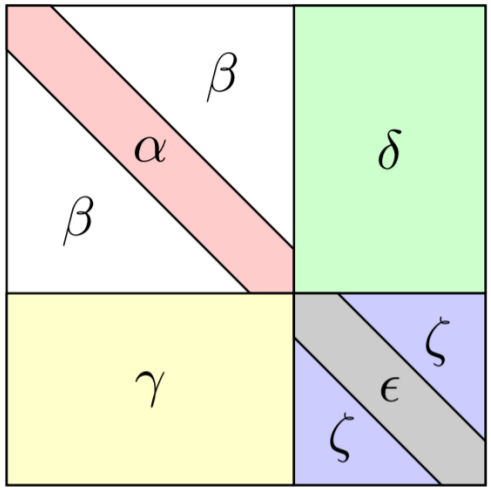
(i) CIFAR-10, large model



(j) CIFAR-10, large model

Open Questions

- The right block structure?



$\mathbf{P}_1\mathbf{W}\mathbf{P}_2$ in *Analytic Study of Families of Spurious Minima in Two-Layer ReLU Neural Networks*, NeurIPS 2021

- Other applications (e.g. Gaussian processes)?