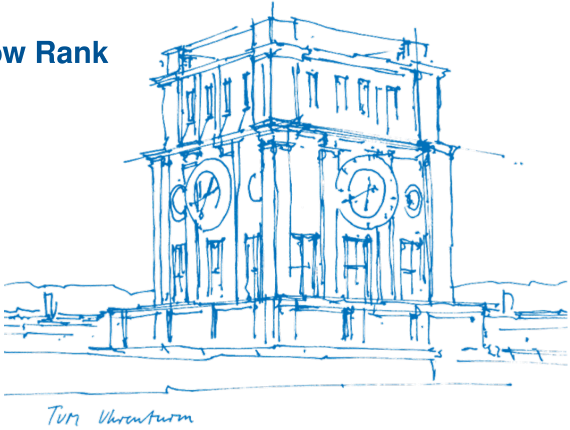
Gradient Descent for Deep Matrix Factorization

Dynamics and Implicit Bias towards Low Rank

Donata Buožytė

July 15th, 2022



Gradient Descent for Deep Matrix Factorization: Dynamics and Implicit Bias towards Low Rank

Hung-Hsu Chou[1], Carsten Gieshoff[1], Johannes Maly[2], and Holger Rauhut[1]

[1]Chair for Mathematics of Information Processing, RWTH Aachen University, Germany
[2]Department of Scientific Computing, KU Eichstaett/Ingolstadt, Germany

August 30, 2021

Outline

Motivation

theoretical studies	real situations
neural network trained with the (stochastic) gradient descent results in zero training error $\rightarrow$ model is overfitting	neural networks trained with the SDG lead to models that generalize pretty well

Motivation

theoretical studies	real situations
neural network trained with the (stochastic) gradient descent results in zero training error $\rightarrow$ model is overfitting	neural networks trained with the SDG lead to models that generalize pretty well

observations:

- optimization algorithms introduce implicit bias towards certain solutions
- SGD converges to linear functions described by a low rank matrix

Motivation

theoretical studies	real situations
neural network trained with the (stochastic) gradient descent results in zero training error $\rightarrow$ model is overfitting	neural networks trained with the SDG lead to models that generalize pretty well

observations:

- optimization algorithms introduce implicit bias towards certain solutions
- SGD converges to linear functions described by a low rank matrix

$\Rightarrow$ key task: understand nature of implicit bias

Feed Forward Neural Networks

General Case

- $h : \mathbb{R}^{n_0} \to \mathbb{R}^{n_N}$, $h(x) = g_N \circ \ldots \circ g_1(x)$, $N > 1$ with:
  - layers: $g_k : \mathbb{R}^{n_{k-1}} \to \mathbb{R}^{n_k}$, $h_k(x) = \sigma(W_k x + b_k)$
  - weight matrices: $W_k \in \mathbb{R}^{n_k \times n_{k-1}}$
  - bias terms: $b_k \in \mathbb{R}^{n_k}$
  - activation function: $\sigma : \mathbb{R} \to \mathbb{R}$, in general non-linear

Feed Forward Neural Networks

General Case

- $h : \mathbb{R}^{n_0} \to \mathbb{R}^{n_N}$, $h(x) = g_N \circ \ldots \circ g_1(x)$, $N > 1$ with:
  - layers: $g_k : \mathbb{R}^{n_{k-1}} \to \mathbb{R}^{n_k}$, $h_k(x) = \sigma(W_k x + b_k)$
  - weight matrices: $W_k \in \mathbb{R}^{n_k \times n_{k-1}}$
  - bias terms: $b_k \in \mathbb{R}^{n_k}$
  - activation function: $\sigma : \mathbb{R} \to \mathbb{R}$, in general non-linear
- loss function: $\mathcal{L} : \mathbb{R}^{n_N} \times \mathbb{R}^{n_N} \to \mathbb{R}_+$
- approach in supervised learning:

$$\min_{W_1,\ldots,W_N} \frac{1}{M} \sum_{i=1}^{M} \mathcal{L}(h(x_i), y_i)$$

Feed Forward Neural Networks

Linear Case

- linear neural networks: $\sigma(x) = x$, $b_k = 0$, hence
$$h_{\text{linear}}(x) = W_N \dots W_1 x$$

Feed Forward Neural Networks
Linear Case

- linear neural networks: $\sigma(x) = x$, $b_k = 0$, hence
$$h_{\text{linear}}(x) = W_N \dots W_1 x$$
- approach with quadratic loss:

$$\min_{W_1,\dots,W_N} \frac{1}{M} \sum_{i=1}^{M} \|W_N \dots W_1 x_i - y_i\|_2^2 \tag{1}$$

Feed Forward Neural Networks
Linear Case

- linear neural networks: $\sigma(x) = x$, $b_k = 0$, hence
$$h_{\text{linear}}(x) = W_N \dots W_1 x$$
- approach with quadratic loss:
$$\min_{W_1,\dots,W_N} \frac{1}{M} \sum_{i=1}^{M} \|W_N \dots W_1 x_i - y_i\|_2^2 \tag{1}$$
- it can be shown: if $\operatorname{span}(\{x_i\}_{i=1}^{M}) = \mathbb{R}^{n_0}$, then the set of minimizers of (1) is equivalent to the set of minimizers of
$$\min_{W_1,\dots,W_N} \|W_N \dots W_1 - \hat{W}\|_F^2 \tag{2}$$

($\hat{W}$ = ground truth matrix)

Frobenius Norm

For $A \in \mathbb{R}^{m \times n}$:

$$
\begin{aligned}
\|A\|_F^2 &= \sum_{i=1}^{m} \sum_{j=1}^{n} (a_{ij})^2 \\
&= \operatorname{trace}(A^T A) \\
&= \operatorname{trace}((U \Sigma V^T)^T (U \Sigma V^T)) \\
&= \operatorname{trace}(\Sigma^T \Sigma) \\
&= \sum_{i=1}^{r} \sigma_i^2
\end{aligned}
$$

Gradient Descent

For the loss function:

$$\mathcal{L}(W) = \frac{1}{2}\|W - \hat{W}\|_F^2, \quad \nabla_W \mathcal{L}(W) = (W - \hat{W}) \tag{3}$$

$$\nabla_{W_j} \mathcal{L}(W) = (W_N \dots W_{j+1})^T \nabla_W \mathcal{L}(W)(W_{j-1} \dots W_1)^T \tag{4}$$

Gradient Descent

For the loss function:

$$\mathcal{L}(W) = \frac{1}{2}\|W - \hat{W}\|_F^2, \quad \nabla_W \mathcal{L}(W) = (W - \hat{W}) \tag{3}$$

$$\nabla_{W_j}\mathcal{L}(W) = (W_N \dots W_{j+1})^T \nabla_W \mathcal{L}(W)(W_{j-1} \dots W_1)^T \tag{4}$$

the gradient descent is defined as:

$$W_j^{(0)} = \alpha W_0 \tag{5}$$

$$W_j^{(k+1)} = W_j^{(k)} - \eta \nabla_{W_j}\mathcal{L}(W^{(k)}) \tag{6}$$

Outline

Dynamics of Gradient Descent with Identical Initialization

Consider the (discrete) dynamics

$$W_j^{(0)} = \alpha W_0$$
$$W_j^{(k+1)} = W_j^{(k)} - \eta \nabla_{W_j} \mathcal{L}(W^{(k)})$$

with an identical initialization $W_0 = I$.

Dynamics of Gradient Descent with Identical Initialization

Now let $(W_j^{(k)})_{j=1}^N$ be solutions of the discrete dynamics with identical initialization and $\hat{W} = V\Lambda V^T$ the eigendecomposition of the symmetric ground truth matrix $\hat{W}$.

Dynamics of Gradient Descent with Identical Initialization

Now let $(W_j^{(k)})_{j=1}^N$ be solutions of the discrete dynamics with identical initialization and $\hat{W} = V\Lambda V^T$ the eigendecomposition of the symmetric ground truth matrix $\hat{W}$.Then:

1. $D^{(k)} = D_j^{(k)} := V^T W_j^{(k)} V \quad \forall j$ are real, diagonal, identical
2. $D^{(k)}$ follows the dynamics $D^{(k+1)} = D^{(k)} - \eta (D^{(k)})^{N-1}((D^{(k)})^N - \Lambda)$

Dynamics of Gradient Descent with Identical Initialization

Hence by definition:

- $W_j^{(k)} = VD^{(k)}V^T \quad \forall j, k$
- $W_N^{(k)} \dots W_1^{(k)} = (VD^{(k)}V^T)^N = V(D^{(k)})^N V^T \quad \forall k$

Dynamics of Gradient Descent with Identical Initialization

Hence by definition:

- $W_j^{(k)} = VD^{(k)}V^T \quad \forall j, k$
- $W_N^{(k)} \dots W_1^{(k)} = (VD^{(k)}V^T)^N = V(D^{(k)})^N V^T \quad \forall k$

Additionally: as $D^{(k)}$ is diagonal, the dynamics can be reformulated as

$$d_{ii}^{(k+1)} = d_{ii}^{(k)} - \eta(d_{ii}^{(k)})^{N-1}((d_{ii}^{(k)})^N - \lambda_i) \tag{7}$$

where λ_i is the corresponding eigenvalue of $\hat{W}$.

Dynamics of Gradient Descent with Identical Initialization

Theorem 2.1 $N \geq 2$, $\lambda \in \mathbb{R}$, $\alpha > 0$, $\{d^{(k)}\}_k$ *the solution of* (7).
Let $M = \max(\alpha, |\lambda|^{\frac{1}{N}})$ *and* η *be s.t.*

$$0 < \eta < \begin{cases} \frac{1}{2NM^{2N-2}} & \textit{if } \lambda \geq 0 \\ \frac{1}{(3N-2)M^{2N-2}} & \textit{if } \lambda < 0. \end{cases}$$

Additionally let $\epsilon \in (0, |\alpha - \lambda_+^{\frac{1}{N}}|)$ *the desired error and* $T = \min\{k : |d^{(k)} - \lambda_+^{\frac{1}{N}}| \leq \epsilon\}$ *the minimal number of steps needed to achieve the desired error.*

Then: $T \leq T_N^{\mathrm{Id}}(\lambda, \epsilon, \alpha, \eta)$.
(Note: T_N^{Id} *is an estimation of the required number of steps with 5 different cases depending on the input values.)*

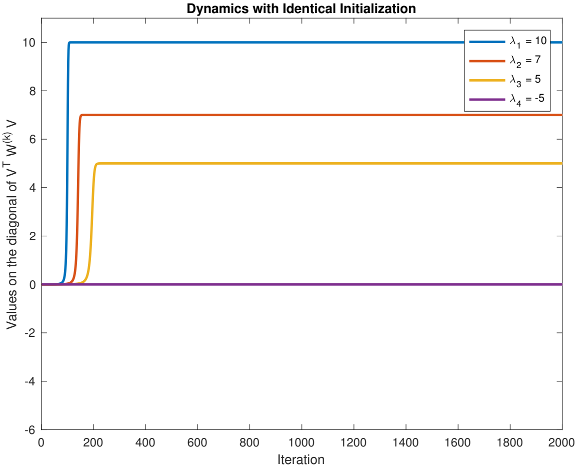
Recovery of positive eigenvalues

Theorem 2.2 *$N \geq 2$, $\hat{W} = V\Lambda V^T \in \mathbb{R}^{n\times n}$ an eigendecomposition the symmetric matrix $\hat{W}$.*
Let $W^{(k)} = W_N^{(k)} \dots W_1^{(k)}$ with $W_j^{(k)}$ defined by (5)-(6) *with loss function* (3)*, $W_0 = I$ and $\alpha > 0$.*
Let $M = \max(\alpha, \|\hat{W}\|_2^{\frac{1}{N}})$ and η be s.t.

$$0 < \eta < \frac{1}{(3N-2)M^{2N-2}}. \tag{8}$$

Then: $\lim_{k\to\infty} W^{(k)} = V\Lambda_+ V^T$ and the error is the diagonal matrix $E^{(k)} = V^T W^{(k)} V - \Lambda_+$.

Recovery of positive eigenvalues: Example



$$\hat{W} = \begin{pmatrix} 7/3 & 10/3 & -4 & 2/3 \\ 10/3 & 10/3 & 5 & 5/3 \\ -4 & 5 & 4 & 1 \\ 2/3 & 5/3 & 1 & -5 \end{pmatrix}$$

$$\Lambda = \begin{pmatrix} 10 & 0 & 0 & 0 \\ 0 & 7 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & -5 \end{pmatrix}$$

$$N = 3, \ \alpha = 0.1$$

Recovery of positive eigenvalues

Conclusion of Theorem 2.2:
the dynamics

$$W_j^{(0)} = \alpha I$$
$$W_j^{(k+1)} = W_j^{(k)} - \eta \nabla_{W_j} \mathcal{L}(W^{(k)})$$

can recover the non-negative eigenvalues of $\hat{W}$ under certain conditions.

Outline

Dynamics of Gradient Descent with Perturbed Initialization

Idea:
instead of initializing all $W_j^{(0)}$ with only αI, perturb the initialization slightly by for example using

$$W_j^{(0)} = \begin{cases} (\alpha - \beta) I & \text{if } j = 1 \\ \alpha I & \text{otherwise} \end{cases} \tag{9}$$

with $0 < \beta < \alpha$.

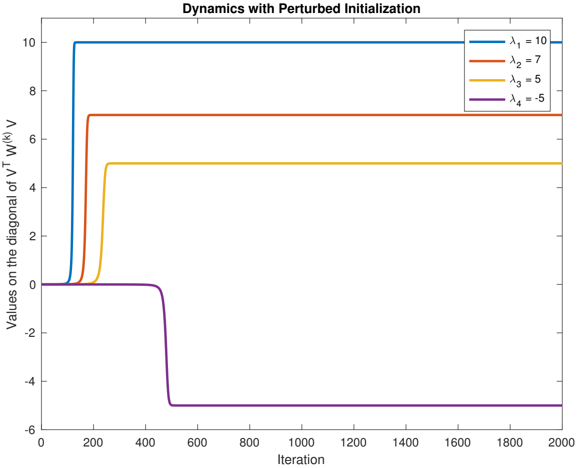
Recovery of arbitrary eigenvalues

Theorem 3.1 *$N \geq 2$, $\hat{W} = V\Lambda V^T \in \mathbb{R}^{n\times n}$ an eigendecomposition the symmetric matrix $\hat{W}$.*
Let $W^{(k)} = W_N^{(k)} \dots W_1^{(k)}$ with $W_j^{(k)}$ defined by (5) *and the perturbed initialization* (9) *with loss function* (3) *and $W_0 = I$.*
Let $M = \max(\alpha, \|\hat{W}\|_2^{\frac{1}{N}})$, $0 < \frac{\beta}{c-1} < \alpha$, $c \in (1,2)$ with c being the maximal real solution of $1 = (c-1)c^{N-1}$ and η be s.t.

$$0 < \eta < \frac{1}{9N(cM)^{2N-2}}. \tag{10}$$

Then: $\lim_{k\to\infty} W^{(k)} = V\Lambda V^T = \hat{W}$ and the error is the diagonal matrix $E^{(k)} = V^T W^{(k)} V - \Lambda$.

Recovery of arbitrary eigenvalues: Example



$$\hat{W} = \begin{pmatrix} 7/3 & 10/3 & -4 & 2/3 \\ 10/3 & 10/3 & 5 & 5/3 \\ -4 & 5 & 4 & 1 \\ 2/3 & 5/3 & 1 & -5 \end{pmatrix}$$

$$\Lambda = \begin{pmatrix} 10 & 0 & 0 & 0 \\ 0 & 7 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & -5 \end{pmatrix}$$

$$N = 3,\ \alpha = 0.1,\ \beta = 0.05$$

Recovery of arbitrary eigenvalues

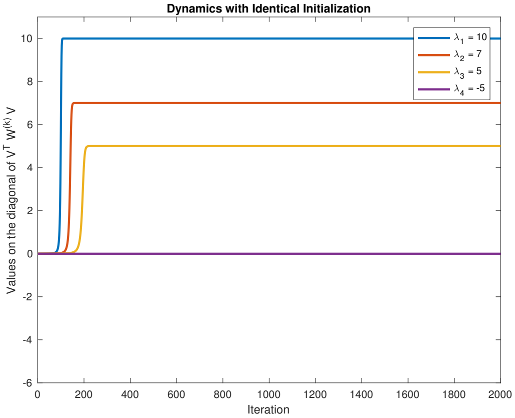
First conclusion of Theorem 3.1:
the dynamics

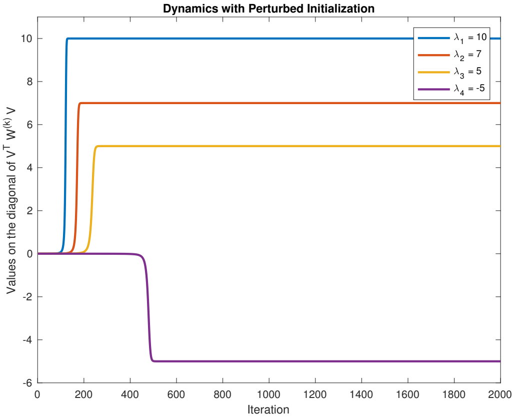
$$W_j^{(0)} = \begin{cases} (\alpha - \beta)I & \text{if } j = 1 \\ \alpha I & \text{otherwise} \end{cases}$$

$$W_j^{(k+1)} = W_j^{(k)} - \eta \nabla_{W_j} \mathcal{L}(W^{(k)})$$

can recover the all eigenvalues of $\hat{W}$ under certain conditions.

Recovery of arbitrary eigenvalues: Example





Recovery of arbitrary eigenvalues

Second conclusion of Theorem 3.1:
there are two different regimes of the dynamic:

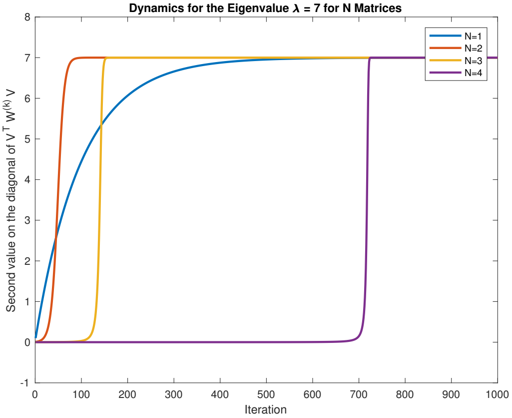
1. if $\lambda_i > 0$: dynamics behave as in 2.2 ("only positive eigenvalues")

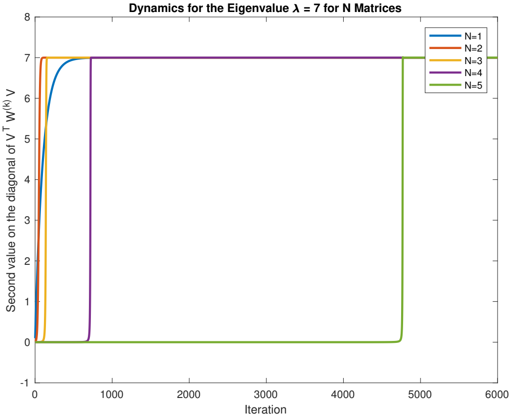
Recovery of arbitrary eigenvalues

Second conclusion of Theorem 3.1:
there are two different regimes of the dynamic:

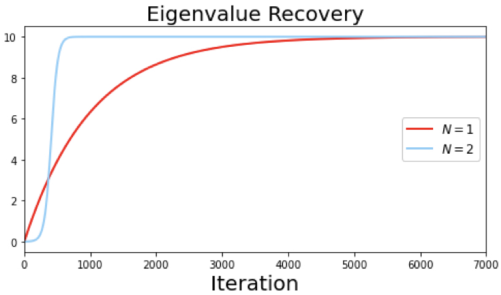
1. if $\lambda_i > 0$: dynamics behave as in 2.2 ("only positive eigenvalues")
2. if $\lambda_i < 0$: at first $(\lambda_i)_+$ is approximated up to β and only after reaching that level forced to take negative values

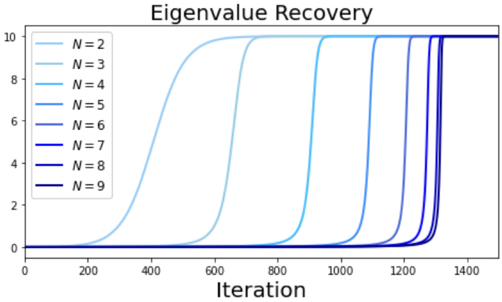
Eigenvalue Recovery for Different Depths: Example





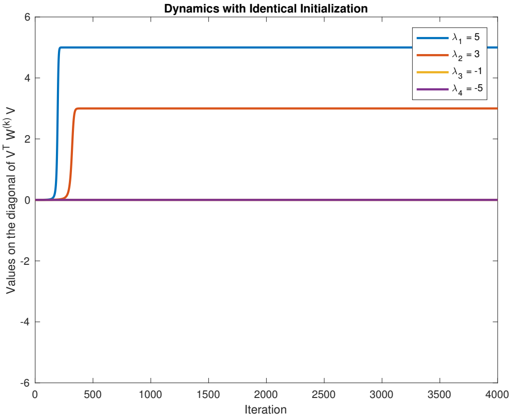
Eigenvalue Recovery for Different Depths

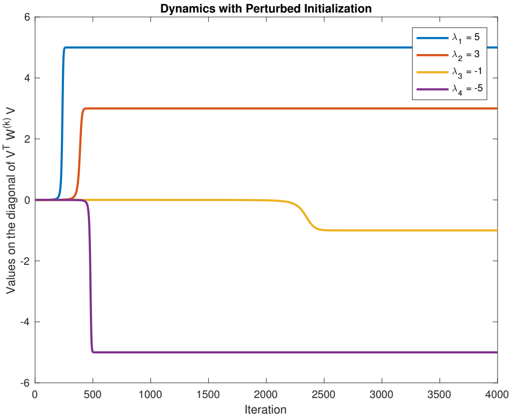




Outline

Implicit Bias Towards Low-Rank: Example





Implicit Bias Towards Low-Rank

The Theorems 2.2 and 3.1 suggest:

- as dominant eigenvalues will be approximated faster: stopping the gradient descent at a suitable finite k will result in a low rank matrix $W^{(k)}$

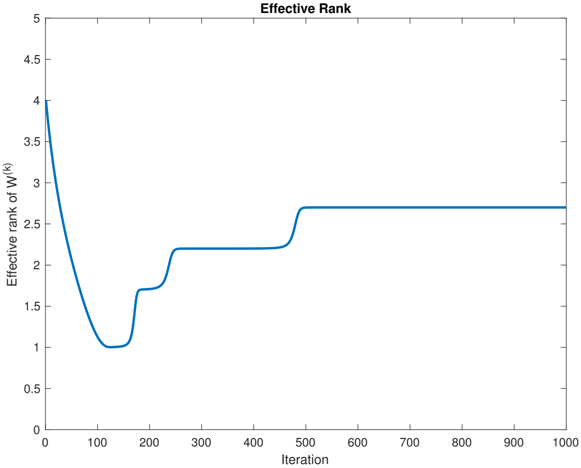
Implicit Bias Towards Low-Rank

The Theorems 2.2 and 3.1 suggest:

- as dominant eigenvalues will be approximated faster: stopping the gradient descent at a suitable finite k will result in a low rank matrix $W^{(k)}$
- it can be shown: effective rank of $W^{(k)}$ drops to one, then monotonically increases to plateau on effective ranks of various low-rank approximations of $\hat{W}$

Effective rank: $1 \leq r(W) = \frac{\|W\|_*}{\|W\|_2} = \frac{\sum_j \sigma(A)_j}{\|W\|_2} \leq \operatorname{rank}(W)$

Effective Rank: Example

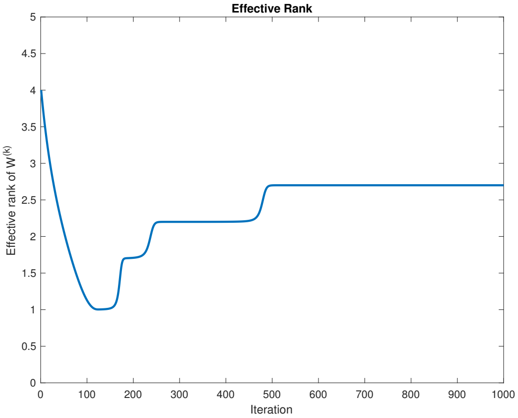


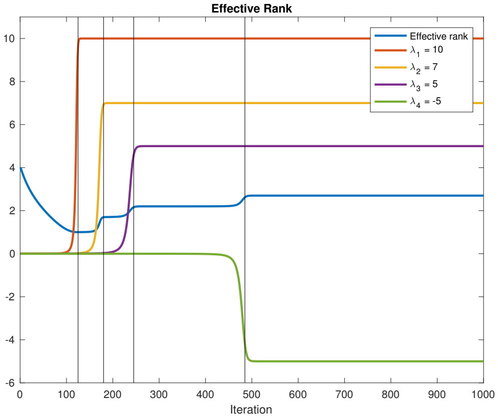
$$\hat{W} = \begin{pmatrix} 7/3 & 10/3 & -4 & 2/3 \\ 10/3 & 10/3 & 5 & 5/3 \\ -4 & 5 & 4 & 1 \\ 2/3 & 5/3 & 1 & -5 \end{pmatrix}$$

$$\Lambda = \begin{pmatrix} 10 & 0 & 0 & 0 \\ 0 & 7 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & -5 \end{pmatrix}$$

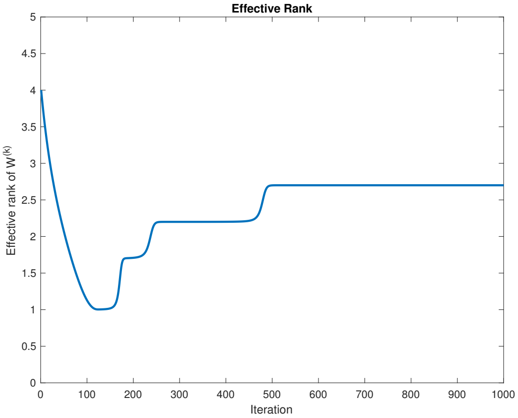
$$N = 3, \ \alpha = 0.1, \ \beta = 0.05$$

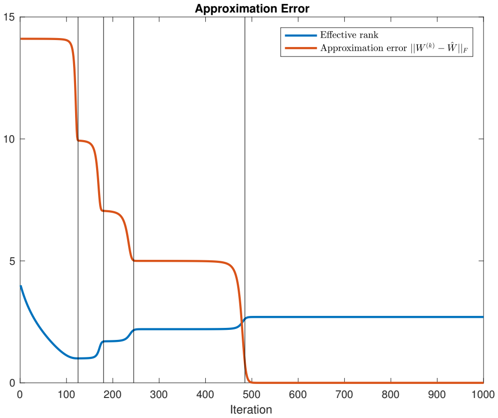
Effective Rank: Example





Effective Rank: Example





Outline

Central Observations

Three central observations:

1. recovery of positive eigenvalues
2. recovery of arbitrary eigenvalues
3. implicit bias towards low-rank

Thank you for your attention!