Direction Matters: On the Implicit Bias of Stochastic Gradient Descent with Moderate Learning Rate

Jingfeng Wu et al., ICLR 2021

Presenter: Sai Anuroop Kesanapalli   Buddy: Hikaru Ibayashi

CSCI 699: Computational Perspectives on the Frontiers of Machine Learning
Spring 2023

University of Southern California

Table of Contents

Table of Contents

Preliminary

We introduce some notation here:

- $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ be a pair of d-dimensional feature vector and 1-dimensional label

Preliminary

We introduce some notation here:

- $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ be a pair of d-dimensional feature vector and 1-dimensional label
- Consider a linear regression problem with square loss defined as $l(x, y; w) := (w^T x - y)^2$, where $w \in \mathbb{R}^d$ is the model parameter

Preliminary

We introduce some notation here:

- $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ be a pair of d-dimensional feature vector and 1-dimensional label
- Consider a linear regression problem with square loss defined as $l(x, y; w) := (w^T x - y)^2$, where $w \in \mathbb{R}^d$ is the model parameter
- Let $\mathcal{D}$ be the population distribution over (x, y)

Preliminary

We introduce some notation here:

- $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ be a pair of d-dimensional feature vector and 1-dimensional label
- Consider a linear regression problem with square loss defined as $l(x, y; w) := (w^T x - y)^2$, where $w \in \mathbb{R}^d$ is the model parameter
- Let $\mathcal{D}$ be the population distribution over (x, y)
- Test loss, $L_{\mathcal{D}}(w) := \mathbb{E}_{(x,y)\sim\mathcal{D}}[l(x, y; w)]$

Preliminary

We introduce some notation here:

- $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ be a pair of d-dimensional feature vector and 1-dimensional label
- Consider a linear regression problem with square loss defined as $l(x, y; w) := (w^T x - y)^2$, where $w \in \mathbb{R}^d$ is the model parameter
- Let $\mathcal{D}$ be the population distribution over (x, y)
- Test loss, $L_{\mathcal{D}}(w) := \mathbb{E}_{(x,y)\sim\mathcal{D}}[l(x, y; w)]$
- Training/ empirical loss, $L_{\mathcal{S}}(w) := \frac{1}{n}\sum_{i=1}^{n}(w^T x_i - y_i)^2$, where $\mathcal{S} := \{(x_i, y_i)\}_{i=1}^{n}$ is a training set of n data points drawn i.i.d. from the population distribution $\mathcal{D}$

Preliminary

Gradient Descent (GD):

- $w_{k+1} = w_k - \eta_k \nabla L_{\mathcal{S}}(w_k) = w_k - \frac{2\eta_k}{n} \sum_{i=1}^{n} x_i(x_i^T w_k - y_i)$

Preliminary

Gradient Descent (GD):

- $w_{k+1} = w_k - \eta_k \nabla L_{\mathcal{S}}(w_k) = w_k - \frac{2\eta_k}{n} \sum_{i=1}^{n} x_i (x_i^T w_k - y_i)$

Mini-Batch Stochastic Gradient Descent (SGD):

- $w_{k,j+1} = w_{k,j} - \frac{\eta_k}{b} \sum_{i \in \mathcal{B}_j^k} \nabla l_i(w_{k,j}) = w_{k,j} - \frac{2\eta_k}{b} \sum_{i \in \mathcal{B}_j^k} x_i (x_i^T w_{k,j} - y_i),$

  $j = 1, \ldots, m$ batches
- Note that j indexes batches and k indexes epochs

Preliminary

Condition Number of a matrix A:

- $\kappa(A) = \dfrac{\sigma_{max}(A)}{\sigma_{min}(A)}$, measures the ratio of the maximum relative stretching to the maximum relative shrinking that matrix does to any non-zero vectors

Preliminary

Condition Number of a matrix A:

- $\kappa(A) = \dfrac{\sigma_{max}(A)}{\sigma_{min}(A)}$, measures the ratio of the maximum relative stretching to the maximum relative shrinking that matrix does to any non-zero vectors
- If $D = \text{Diag}(d_i)$ is a diagonal matrix, then $\kappa(D) = \dfrac{\max(d_i)}{\min(d_i)}$

Preliminary

Condition Number of a matrix A:

- $\kappa(A) = \frac{\sigma_{max}(A)}{\sigma_{min}(A)}$, measures the ratio of the maximum relative stretching to the maximum relative shrinking that matrix does to any non-zero vectors
- If $D = \text{Diag}(d_i)$ is a diagonal matrix, then $\kappa(D) = \frac{\max(d_i)}{\min(d_i)}$
- A "problem" with a low condition number is said to be well-conditioned, while a problem with a high condition number is said to be ill-conditioned

Preliminary

Projection operator P:

- A projection on a vector space V is a linear operator $P : V \to V$ such that $P^2 = P$

Preliminary

Projection operator P:

- A projection on a vector space V is a linear operator $P : V \to V$ such that $P^2 = P$
- For example, $P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ projects a point $(x, y, z) \in \mathbb{R}^3$ to its image on the $x - y$ plane, i.e., $(x, y, 0) \in \mathbb{R}^3$

Preliminary

Column space of a matrix A:

- Let K be a field of scalars. Let A be an $m \times n$ matrix, with column vectors $v_1, \ldots, v_n$

Preliminary

Column space of a matrix A:

- Let K be a field of scalars. Let A be an $m \times n$ matrix, with column vectors $v_1, \ldots, v_n$
- A linear combination of these vectors is any vector of the form $c_1 v_1 + \cdots + c_n v_n$, where $c_1, \ldots, c_n$ are scalars

Preliminary

Column space of a matrix A:

- Let K be a field of scalars. Let A be an $m \times n$ matrix, with column vectors $v_1, \ldots, v_n$
- A linear combination of these vectors is any vector of the form $c_1 v_1 + \cdots + c_n v_n$, where $c_1, \ldots, c_n$ are scalars
- The set of all possible linear combinations of $v_1, \ldots, v_n$ is called the column space of A

Preliminary

Column space of a matrix A:

- Let K be a field of scalars. Let A be an $m \times n$ matrix, with column vectors $v_1, \ldots, v_n$
- A linear combination of these vectors is any vector of the form $c_1 v_1 + \cdots + c_n v_n$, where $c_1, \ldots, c_n$ are scalars
- The set of all possible linear combinations of $v_1, \ldots, v_n$ is called the column space of A
- Any linear combination of the column vectors of a matrix A can be written as the product of A with a column vector:

$$A \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} = c_1 \begin{bmatrix} a_{11} \\ \vdots \\ a_{m1} \end{bmatrix} + \cdots + c_n \begin{bmatrix} a_{1n} \\ \vdots \\ a_{mn} \end{bmatrix} = c_1 v_1 + \cdots + c_n v_n$$

Preliminary

Column space of a matrix A:

- Let K be a field of scalars. Let A be an $m \times n$ matrix, with column vectors $v_1, \dots, v_n$
- A linear combination of these vectors is any vector of the form $c_1 v_1 + \cdots + c_n v_n$, where $c_1, \dots, c_n$ are scalars
- The set of all possible linear combinations of $v_1, \dots, v_n$ is called the column space of A
- Any linear combination of the column vectors of a matrix A can be written as the product of A with a column vector:

$$A \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} = c_1 \begin{bmatrix} a_{11} \\ \vdots \\ a_{m1} \end{bmatrix} + \cdots + c_n \begin{bmatrix} a_{1n} \\ \vdots \\ a_{mn} \end{bmatrix} = c_1 v_1 + \cdots + c_n v_n$$

- Therefore, the column space of A consists of all possible products Ax, for $x \in K^n$

Preliminary

Lipschitz continuous gradient:

- Let f be a twice differentiable convex function

Preliminary

Lipschitz continuous gradient:

- Let f be a twice differentiable convex function
- Then f has a Lipschitz continuous gradient if there exists an L such that $\nabla^2 f \preccurlyeq LI$

Preliminary

Lipschitz continuous gradient:

- Let f be a twice differentiable convex function
- Then f has a Lipschitz continuous gradient if there exists an L such that $\nabla^2 f \preccurlyeq LI$
- In other words, the largest eigenvalue of the Hessian of f is uniformly upper bounded by L everywhere

Table of Contents

The Minimum-Norm Bias of SGD/ GD

We consider the case of SGD/ GD optimizing linear regression problem:

- Rewrite training loss as $L_{\mathcal{S}}(w) = \frac{1}{n}||X^T w - Y||_2^2$, where $X \in \mathbb{R}^{d \times n}$ and $Y \in \mathbb{R}^n$

The Minimum-Norm Bias of SGD/ GD

We consider the case of SGD/ GD optimizing linear regression problem:

- Rewrite training loss as $L_{\mathcal{S}}(w) = \frac{1}{n}||X^T w - Y||_2^2$, where $X \in \mathbb{R}^{d\times n}$ and $Y \in \mathbb{R}^n$
- Then its global minima are given by $\mathcal{W}_* := \{w \in \mathbb{R}^d : Pw = w_*,\ w_* := X(X^T X)^{-1} Y\}$, where P is the projection operator onto the column space of X

The Minimum-Norm Bias of SGD/ GD

We consider the case of SGD/ GD optimizing linear regression problem:

- Rewrite training loss as $L_{\mathcal{S}}(w) = \frac{1}{n}||X^T w - Y||_2^2$, where $X \in \mathbb{R}^{d \times n}$ and $Y \in \mathbb{R}^n$
- Then its global minima are given by $\mathcal{W}_* := \{w \in \mathbb{R}^d : Pw = w_*,\ w_* := X(X^T X)^{-1} Y\}$, where P is the projection operator onto the column space of X

The Minimum-Norm Bias of SGD/ GD

We consider the case of SGD/ GD optimizing linear regression problem:

- Rewrite training loss as $L_{\mathcal{S}}(w) = \frac{1}{n}||X^T w - Y||_2^2$, where $X \in \mathbb{R}^{d \times n}$ and $Y \in \mathbb{R}^n$
- Then its global minima are given by $\mathcal{W}_* := \{w \in \mathbb{R}^d : Pw = w_*,\ w_* := X(X^T X)^{-1} Y\}$, where P is the projection operator onto the column space of X
- We focus on overparameterized cases where $\mathcal{W}_*$ is not a singleton

The Minimum-Norm Bias of SGD/ GD

- Notice that every gradient $\nabla l_i(w) = 2x_i(x_i^T w - y_i)$ is spanned in the column space of the data manifold

The Minimum-Norm Bias of SGD/ GD

- Notice that every gradient $\nabla l_i(w) = 2x_i(x_i^T w - y_i)$ is spanned in the column space of the data manifold
- Thus, GD and SGD can never move along the direction that is orthogonal to the data manifold

The Minimum-Norm Bias of SGD/ GD

- Notice that every gradient $\nabla l_i(w) = 2x_i(x_i^T w - y_i)$ is spanned in the column space of the data manifold
- Thus, GD and SGD can never move along the direction that is orthogonal to the data manifold
- This means they implicitly admit the following hypothesis class:

$$\mathcal{H}_{\mathcal{S}} = \{w \in \mathbb{R}^d : P_\perp w = P_\perp w_0\},$$

where w_0 is the initializtion and $P_\perp = I - P$ is the projection operator onto the orthogonal complement to the column space of X

The Minimum-Norm Bias of SGD/ GD

For any global optimum $w \in W_*$, i.e., $Pw = w_*$, consider the following:

The Minimum-Norm Bias of SGD/ GD

For any global optimum $w \in W_*$, i.e., $Pw = w_*$, consider the following:

$$w - w_0 = (P + P_\perp)(w - w_0) \iff ||w - w_0||_2^2 = ||(P + P_\perp)(w - w_0)||_2^2$$

The Minimum-Norm Bias of SGD/ GD

For any global optimum $w \in W_*$, i.e., $Pw = w_*$, consider the following:

$$w - w_0 = (P + P_\perp)(w - w_0) \iff ||w - w_0||_2^2 = ||(P + P_\perp)(w - w_0)||_2^2$$

$$\begin{aligned}
&= ||Pw - Pw_0 + P_\perp(w - w_0)\}||_2^2 \\
&= ||w_* - Pw_0 + P_\perp(w - w_0)||_2^2 \\
&= ||w_* - Pw_0||_2^2 + ||P_\perp(w - w_0)||_2^2 + 2(w_* - Pw_0)^T(P_\perp(w - w_0))
\end{aligned}$$

The Minimum-Norm Bias of SGD/ GD

For any global optimum $w \in W_*$, i.e., $Pw = w_*$, consider the following:

$$w - w_0 = (P + P_\perp)(w - w_0) \iff ||w - w_0||_2^2 = ||(P + P_\perp)(w - w_0)||_2^2$$

$$\begin{aligned}
&= ||Pw - Pw_0 + P_\perp(w - w_0)\}||_2^2 \\
&= ||w_* - Pw_0 + P_\perp(w - w_0)||_2^2 \\
&= ||w_* - Pw_0||_2^2 + ||P_\perp(w - w_0)||_2^2 + 2(w_* - Pw_0)^T(P_\perp(w - w_0)) \\
&= ||w_* - Pw_0||_2^2 + ||P_\perp(w - w_0)||_2^2 + 2(w - w_0)^T P^T P_\perp(w - w_0)
\end{aligned}$$

The Minimum-Norm Bias of SGD/ GD

For any global optimum $w \in W_*$, i.e., $Pw = w_*$, consider the following:

$$w - w_0 = (P + P_\perp)(w - w_0) \iff ||w - w_0||_2^2 = ||(P + P_\perp)(w - w_0)||_2^2$$

$$\begin{aligned}
&= ||Pw - Pw_0 + P_\perp(w - w_0)\}||_2^2 \\
&= ||w_* - Pw_0 + P_\perp(w - w_0)||_2^2 \\
&= ||w_* - Pw_0||_2^2 + ||P_\perp(w - w_0)||_2^2 + 2(w_* - Pw_0)^T(P_\perp(w - w_0)) \\
&= ||w_* - Pw_0||_2^2 + ||P_\perp(w - w_0)||_2^2 + 2(w - w_0)^T {\color{red}P^T P_\perp}(w - w_0)
\end{aligned}$$

The Minimum-Norm Bias of SGD/ GD

For any global optimum $w \in W_*$, i.e., $Pw = w_*$, consider the following:

$$w - w_0 = (P + P_\perp)(w - w_0) \iff ||w - w_0||_2^2 = ||(P + P_\perp)(w - w_0)||_2^2$$

$$\begin{aligned}
&= ||Pw - Pw_0 + P_\perp(w - w_0)\}||_2^2 \\
&= ||w_* - Pw_0 + P_\perp(w - w_0)||_2^2 \\
&= ||w_* - Pw_0||_2^2 + ||P_\perp(w - w_0)||_2^2 + 2(w_* - Pw_0)^T(P_\perp(w - w_0)) \\
&= ||w_* - Pw_0||_2^2 + ||P_\perp(w - w_0)||_2^2 + 2(w - w_0)^T P^T P_\perp(w - w_0) \\
&= ||w_* - Pw_0||_2^2 + ||P_\perp(w - w_0)||_2^2
\end{aligned}$$

The Minimum-Norm Bias of SGD/ GD

For any global optimum $w \in W_*$, i.e., $Pw = w_*$, consider the following:

$$w - w_0 = (P + P_\perp)(w - w_0) \iff ||w - w_0||_2^2 = ||(P + P_\perp)(w - w_0)||_2^2$$

$$\begin{aligned}
&= ||Pw - Pw_0 + P_\perp(w - w_0)\}||_2^2 \\
&= ||w_* - Pw_0 + P_\perp(w - w_0)||_2^2 \\
&= ||w_* - Pw_0||_2^2 + ||P_\perp(w - w_0)||_2^2 + 2(w_* - Pw_0)^T(P_\perp(w - w_0)) \\
&= ||w_* - Pw_0||_2^2 + ||P_\perp(w - w_0)||_2^2 + 2(w - w_0)^T P^T P_\perp(w - w_0) \\
&= ||w_* - Pw_0||_2^2 + ||P_\perp(w - w_0)||_2^2
\end{aligned}$$

- Note here that $||w - Pw_0||_2^2 + ||P_\perp(w - w_0)||_2^2$ is minimized when $P_\perp w = P_\perp w_0$, i.e., $w \in \mathcal{H}_\mathcal{S}$.

The Minimum-Norm Bias of SGD/ GD

- Thus w is the solution found by SGD/ GD when the learning rate is set properly so that the algorithms can find a global optimum

The Minimum-Norm Bias of SGD/ GD

- Thus w is the solution found by SGD/ GD when the learning rate is set properly so that the algorithms can find a global optimum
- Since initialization is usually set to zero, SGD/ GD is biased to find the global optimum that is closest to the initialization, which is referred as the "minimum-norm" bias in literature

Table of Contents

Directional Bias of SGD/ GD: A Toy Example

We now conduct a 2-dimensional case study to motivate the directional bias of SGD in the moderate learning rate regime

- Consider a training set consisting of just two orthogonal points,
$$\mathcal{S} = \left\{ \left(x_1 = \begin{bmatrix} \sqrt{\kappa} \\ 0 \end{bmatrix}, y_1 = 0 \right), \left(x_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, y_2 = 0 \right) \right\}, \ \kappa > 2$$

Directional Bias of SGD/ GD: A Toy Example

We now conduct a 2-dimensional case study to motivate the directional bias of SGD in the moderate learning rate regime

- Consider a training set consisting of just two orthogonal points, $\mathcal{S} = \left\{ \left(x_1 = \begin{bmatrix} \sqrt{\kappa} \\ 0 \end{bmatrix}, y_1 = 0 \right), \left(x_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, y_2 = 0 \right) \right\}$, $\kappa > 2$
- Let $w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$

Directional Bias of SGD/ GD: A Toy Example

We now conduct a 2-dimensional case study to motivate the directional bias of SGD in the moderate learning rate regime

- Consider a training set consisting of just two orthogonal points, $$\mathcal{S} = \left\{ \left(x_1 = \begin{bmatrix} \sqrt{\kappa} \\ 0 \end{bmatrix}, y_1 = 0 \right), \left(x_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, y_2 = 0 \right) \right\}, \ \kappa > 2$$
- Let $w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$
- $L_{\mathcal{S}}(w) = \frac{1}{n} \sum_{i=1}^{n} (w^T x_i - y_i)^2 = \frac{1}{2}(l_1(w) + l_2(w)) = \frac{1}{2}(w_1^2 \kappa + w_2^2)$

Directional Bias of SGD/ GD: A Toy Example

We now conduct a 2-dimensional case study to motivate the directional bias of SGD in the moderate learning rate regime

- Consider a training set consisting of just two orthogonal points, $\mathcal{S} = \left\{ \left(x_1 = \begin{bmatrix} \sqrt{\kappa} \\ 0 \end{bmatrix}, y_1 = 0 \right), \left(x_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, y_2 = 0 \right) \right\}$, $\kappa > 2$
- Let $w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$
- $L_{\mathcal{S}}(w) = \frac{1}{n} \sum_{i=1}^{n} (w^T x_i - y_i)^2 = \frac{1}{2}(l_1(w) + l_2(w)) = \frac{1}{2}(w_1^2 \kappa + w_2^2)$
- $\nabla L_{\mathcal{S}}(w) = 0 \iff \begin{bmatrix} w_1 \kappa \\ w_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

Directional Bias of SGD/ GD: A Toy Example

We now conduct a 2-dimensional case study to motivate the directional bias of SGD in the moderate learning rate regime

- Consider a training set consisting of just two orthogonal points, $\mathcal{S} = \left\{ \left(x_1 = \begin{bmatrix} \sqrt{\kappa} \\ 0 \end{bmatrix}, y_1 = 0 \right), \left(x_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, y_2 = 0 \right) \right\}$, $\kappa > 2$
- Let $w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$
- $L_{\mathcal{S}}(w) = \frac{1}{n} \sum_{i=1}^{n} (w^T x_i - y_i)^2 = \frac{1}{2}(l_1(w) + l_2(w)) = \frac{1}{2}(w_1^2 \kappa + w_2^2)$
- $\nabla L_{\mathcal{S}}(w) = 0 \iff \begin{bmatrix} w_1 \kappa \\ w_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$
- So $w_* = 0$ is the unique minimum of $L_{\mathcal{S}}(w)$

Directional Bias of SGD/ GD: A Toy Example

Further,

- $\nabla^2 L_{\mathcal{S}}(w) = \begin{bmatrix} \kappa & 0 \\ 0 & 1 \end{bmatrix} \implies L_{\mathcal{S}}(w)$ is κ-smooth

Directional Bias of SGD/ GD: A Toy Example

Further,

- $\nabla^2 L_{\mathcal{S}}(w) = \begin{bmatrix} \kappa & 0 \\ 0 & 1 \end{bmatrix} \implies L_{\mathcal{S}}(w)$ is κ-smooth

Directional Bias of SGD/ GD: A Toy Example

Further,

- $\nabla^2 L_{\mathcal{S}}(w) = \begin{bmatrix} \kappa & 0 \\ 0 & 1 \end{bmatrix} \implies L_{\mathcal{S}}(w)$ is κ-smooth
- $\nabla^2 l_1(w) = \begin{bmatrix} 2\kappa & 0 \\ 0 & 0 \end{bmatrix} \implies l_1(w)$ is 2κ-smooth

Directional Bias of SGD/ GD: A Toy Example

Further,

- $\nabla^2 L_{\mathcal{S}}(w) = \begin{bmatrix} \kappa & 0 \\ 0 & 1 \end{bmatrix} \implies L_{\mathcal{S}}(w)$ is κ-smooth
- $\nabla^2 l_1(w) = \begin{bmatrix} 2\kappa & 0 \\ 0 & 0 \end{bmatrix} \implies l_1(w)$ is 2κ-smooth
- $\nabla^2 l_2(w) = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix} \implies l_2(w)$ is 2-smooth

Directional Bias of SGD/ GD: A Toy Example

Further,

- $\nabla^2 L_{\mathcal{S}}(w) = \begin{bmatrix} \kappa & 0 \\ 0 & 1 \end{bmatrix} \implies L_{\mathcal{S}}(w)$ is κ-smooth
- $\nabla^2 l_1(w) = \begin{bmatrix} 2\kappa & 0 \\ 0 & 0 \end{bmatrix} \implies l_1(w)$ is 2κ-smooth
- $\nabla^2 l_2(w) = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix} \implies l_2(w)$ is 2-smooth
- Thus $l_2(w)$ is 2-smooth, but $l_1(w)$, the individual loss for data x_1, is only 2κ-smooth, which is more ill-conditioned compared to $L_{\mathcal{S}}(w)$ and $l_2(w)$

Directional Bias of SGD/ GD: A Toy Example

Let us now analytically solve for the solutions of GD and SGD. Starting with GD, recall that the update step is $w_k = w_{k-1} - \eta \nabla L_{\mathcal{S}}(w_{k-1})$

$$= w_{k-1} - \eta \begin{bmatrix} w_{k-1,1}\kappa \\ w_{k-1,2} \end{bmatrix}$$

Directional Bias of SGD/ GD: A Toy Example

Let us now analytically solve for the solutions of GD and SGD. Starting with GD, recall that the update step is $w_k = w_{k-1} - \eta \nabla L_{\mathcal{S}}(w_{k-1})$

$$= w_{k-1} - \eta \begin{bmatrix} w_{k-1,1}\kappa \\ w_{k-1,2} \end{bmatrix}$$

$$= w_{k-1} - \eta \begin{bmatrix} \kappa & 0 \\ 0 & 1 \end{bmatrix} w_{k-1}$$

Directional Bias of SGD/ GD: A Toy Example

Let us now analytically solve for the solutions of GD and SGD. Starting with GD, recall that the update step is $w_k = w_{k-1} - \eta \nabla L_{\mathcal{S}}(w_{k-1})$

$$= w_{k-1} - \eta \begin{bmatrix} w_{k-1,1}\kappa \\ w_{k-1,2} \end{bmatrix}$$

$$= w_{k-1} - \eta \begin{bmatrix} \kappa & 0 \\ 0 & 1 \end{bmatrix} w_{k-1}$$

$$= \begin{bmatrix} 1 - \eta\kappa & 0 \\ 0 & 1 - \eta \end{bmatrix} w_{k-1}$$

Directional Bias of SGD/ GD: A Toy Example

Let us now analytically solve for the solutions of GD and SGD. Starting with GD, recall that the update step is $w_k = w_{k-1} - \eta \nabla L_{\mathcal{S}}(w_{k-1})$

$$= w_{k-1} - \eta \begin{bmatrix} w_{k-1,1}\kappa \\ w_{k-1,2} \end{bmatrix}$$

$$= w_{k-1} - \eta \begin{bmatrix} \kappa & 0 \\ 0 & 1 \end{bmatrix} w_{k-1}$$

$$= \begin{bmatrix} 1 - \eta\kappa & 0 \\ 0 & 1 - \eta \end{bmatrix} w_{k-1}$$

$$= \begin{bmatrix} (1 - \eta\kappa)^k & 0 \\ 0 & (1 - \eta)^k \end{bmatrix} w_0$$

Directional Bias of SGD/ GD: A Toy Example

Let us now analytically solve for the solutions of GD and SGD. Starting with GD, recall that the update step is $w_k = w_{k-1} - \eta \nabla L_{\mathcal{S}}(w_{k-1})$

$$
\begin{aligned}
&= w_{k-1} - \eta \begin{bmatrix} w_{k-1,1}\kappa \\ w_{k-1,2} \end{bmatrix} \\
&= w_{k-1} - \eta \begin{bmatrix} \kappa & 0 \\ 0 & 1 \end{bmatrix} w_{k-1} \\
&= \begin{bmatrix} 1-\eta\kappa & 0 \\ 0 & 1-\eta \end{bmatrix} w_{k-1} \\
&= \begin{bmatrix} (1-\eta\kappa)^k & 0 \\ 0 & (1-\eta)^k \end{bmatrix} w_0
\end{aligned}
$$

So we have that $w_k^{gd} = \begin{bmatrix} (1-\eta\kappa)^k & 0 \\ 0 & (1-\eta)^k \end{bmatrix} w_0$

Directional Bias of SGD/ GD: A Toy Example

Let us now analytically solve for the solutions of GD and SGD. Starting with GD, recall that the update step is $w_k = w_{k-1} - \eta \nabla L_{\mathcal{S}}(w_{k-1})$

$$= w_{k-1} - \eta \begin{bmatrix} w_{k-1,1}\kappa \\ w_{k-1,2} \end{bmatrix}$$

$$= w_{k-1} - \eta \begin{bmatrix} \kappa & 0 \\ 0 & 1 \end{bmatrix} w_{k-1}$$

$$= \begin{bmatrix} 1-\eta\kappa & 0 \\ 0 & 1-\eta \end{bmatrix} w_{k-1}$$

$$= \begin{bmatrix} (1-\eta\kappa)^k & 0 \\ 0 & (1-\eta)^k \end{bmatrix} w_0$$

So we have that $w_k^{gd} = \begin{bmatrix} (1-\eta\kappa)^k & 0 \\ 0 & (1-\eta)^k \end{bmatrix} w_0$

For $\eta \in \left(\frac{1}{\kappa}, \frac{2}{1+\kappa}\right)$, $|1-\eta\kappa| < |1-\eta| < 1$

Directional Bias of SGD/ GD: A Toy Example

- With moderate learning rate GD is convergent for both directions e_1 and e_2

Directional Bias of SGD/ GD: A Toy Example

- With moderate learning rate GD is convergent for both directions e_1 and e_2
- GD fits e_1 faster since the contraction parameter is smaller, i.e., $|1 - \eta\kappa| < |1 - \eta| < 1$

Directional Bias of SGD/ GD: A Toy Example

- With moderate learning rate GD is convergent for both directions e_1 and e_2
- GD fits e_1 faster since the contraction parameter is smaller, i.e., $|1-\eta\kappa| < |1-\eta| < 1$
- Thus observing the entire optimization path, GD approaches the minimum $w_* = 0$ along e_2, which corresponds to the smaller eigenvalue direction of $\nabla^2 L_{\mathcal{S}}(w)$

Directional Bias of SGD/ GD: A Toy Example

- With moderate learning rate GD is convergent for both directions e_1 and e_2
- GD fits e_1 faster since the contraction parameter is smaller, i.e., $|1 - \eta\kappa| < |1 - \eta| < 1$
- Thus observing the entire optimization path, GD approaches the minimum $w_* = 0$ along e_2, which corresponds to the smaller eigenvalue direction of $\nabla^2 L_{\mathcal{S}}(w)$
- We note this directional bias for GD also holds in the small learning rate regime

Directional Bias of SGD/ GD: A Toy Example

For SGD, recall that the update step is

$$w_{k,j+1} = w_{k-1,j} - \frac{\eta}{b} \sum_{i \in \mathcal{B}_j^k} \nabla l_i(w_{k-1,j})$$

$$= w_{k-1} - 2\eta \begin{bmatrix} w_{k-1,1}\kappa \\ w_{k-1,2} \end{bmatrix}$$

Directional Bias of SGD/ GD: A Toy Example

For SGD, recall that the update step is

$$w_{k,j+1} = w_{k-1,j} - \frac{\eta}{b} \sum_{i \in \mathcal{B}_j^k} \nabla l_i(w_{k-1,j})$$

$$= w_{k-1} - 2\eta \begin{bmatrix} w_{k-1,1}\kappa \\ w_{k-1,2} \end{bmatrix}$$

$$= w_{k-1} - 2\eta \begin{bmatrix} \kappa & 0 \\ 0 & 1 \end{bmatrix} w_{k-1}$$

Directional Bias of SGD/ GD: A Toy Example

For SGD, recall that the update step is

$$
\begin{aligned}
w_{k,j+1} &= w_{k-1,j} - \frac{\eta}{b} \sum_{i \in \mathcal{B}_j^k} \nabla l_i(w_{k-1,j}) \\
&= w_{k-1} - 2\eta \begin{bmatrix} w_{k-1,1}\kappa \\ w_{k-1,2} \end{bmatrix} \\
&= w_{k-1} - 2\eta \begin{bmatrix} \kappa & 0 \\ 0 & 1 \end{bmatrix} w_{k-1} \\
&= \begin{bmatrix} 1 - 2\eta\kappa & 0 \\ 0 & 1 - 2\eta \end{bmatrix} w_{k-1}
\end{aligned}
$$

Directional Bias of SGD/ GD: A Toy Example

For SGD, recall that the update step is

$$\begin{aligned}
w_{k,j+1} &= w_{k-1,j} - \frac{\eta}{b} \sum_{i \in \mathcal{B}_j^k} \nabla l_i(w_{k-1,j}) \\
&= w_{k-1} - 2\eta \begin{bmatrix} w_{k-1,1}\kappa \\ w_{k-1,2} \end{bmatrix} \\
&= w_{k-1} - 2\eta \begin{bmatrix} \kappa & 0 \\ 0 & 1 \end{bmatrix} w_{k-1} \\
&= \begin{bmatrix} 1-2\eta\kappa & 0 \\ 0 & 1-2\eta \end{bmatrix} w_{k-1} \\
&= \begin{bmatrix} (1-2\eta\kappa)^k & 0 \\ 0 & (1-2\eta)^k \end{bmatrix} w_0
\end{aligned}$$

Directional Bias of SGD/ GD: A Toy Example

For SGD, recall that the update step is

$$
\begin{aligned}
w_{k,j+1} &= w_{k-1,j} - \frac{\eta}{b} \sum_{i \in \mathcal{B}_j^k} \nabla l_i(w_{k-1,j}) \\
&= w_{k-1} - 2\eta \begin{bmatrix} w_{k-1,1}\kappa \\ w_{k-1,2} \end{bmatrix} \\
&= w_{k-1} - 2\eta \begin{bmatrix} \kappa & 0 \\ 0 & 1 \end{bmatrix} w_{k-1} \\
&= \begin{bmatrix} 1-2\eta\kappa & 0 \\ 0 & 1-2\eta \end{bmatrix} w_{k-1} \\
&= \begin{bmatrix} (1-2\eta\kappa)^k & 0 \\ 0 & (1-2\eta)^k \end{bmatrix} w_0
\end{aligned}
$$

So we have that $w_k^{sgd} = \begin{bmatrix} (1-2\eta\kappa)^k & 0 \\ 0 & (1-2\eta)^k \end{bmatrix} w_0$

Directional Bias of SGD/ GD: A Toy Example

For SGD, recall that the update step is

$$w_{k,j+1} = w_{k-1,j} - \frac{\eta}{b}\sum_{i\in\mathcal{B}_j^k} \nabla l_i(w_{k-1,j})$$

$$= w_{k-1} - 2\eta \begin{bmatrix} w_{k-1,1}\kappa \\ w_{k-1,2} \end{bmatrix}$$

$$= w_{k-1} - 2\eta \begin{bmatrix} \kappa & 0 \\ 0 & 1 \end{bmatrix} w_{k-1}$$

$$= \begin{bmatrix} 1-2\eta\kappa & 0 \\ 0 & 1-2\eta \end{bmatrix} w_{k-1}$$

$$= \begin{bmatrix} (1-2\eta\kappa)^k & 0 \\ 0 & (1-2\eta)^k \end{bmatrix} w_0$$

So we have that $w_k^{sgd} = \begin{bmatrix} (1-2\eta\kappa)^k & 0 \\ 0 & (1-2\eta)^k \end{bmatrix} w_0$

For $\eta \in \left(\frac{1}{\kappa}, \frac{2}{1+\kappa}\right)$, $|1-2\eta| < 1 < |1-2\eta\kappa|$

Directional Bias of SGD/ GD: A Toy Example

- With moderate learning rate SGD converges along e_2 but oscillates along e_1 since $|1 - 2\eta| < 1 < |1 - 2\eta\kappa|$

Directional Bias of SGD/ GD: A Toy Example

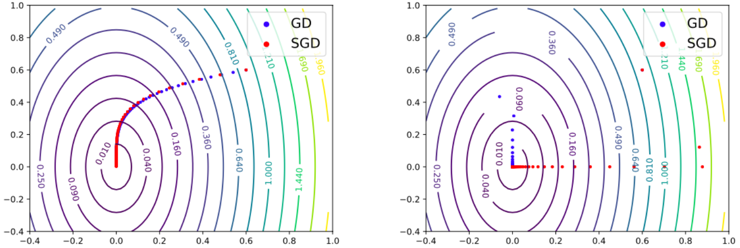
- With moderate learning rate SGD converges along e_2 but oscillates along e_1 since $|1 - 2\eta| < 1 < |1 - 2\eta\kappa|$
- SGD cannot fit e_1 before the learning rate decays, however when this happens, e_2 is already well fitted

Directional Bias of SGD/ GD: A Toy Example

- With moderate learning rate SGD converges along e_2 but oscillates along e_1 since $|1 - 2\eta| < 1 < |1 - 2\eta\kappa|$
- SGD cannot fit e_1 before the learning rate decays, however when this happens, e_2 is already well fitted
- Overall, SGD fits $e2$ first then fits e_1, i.e., SGD converges to the minimum $w_* = 0$ along e_1, which corresponds to the larger eigenvalue direction of $\nabla^2 L_{\mathcal{S}}(w)$

Directional Bias of SGD/ GD: A Toy Example

- With moderate learning rate SGD converges along e_2 but oscillates along e_1 since $|1 - 2\eta| < 1 < |1 - 2\eta\kappa|$
- SGD cannot fit e_1 before the learning rate decays, however when this happens, e_2 is already well fitted
- Overall, SGD fits $e2$ first then fits e_1, i.e., SGD converges to the minimum $w_* = 0$ along e_1, which corresponds to the larger eigenvalue direction of $\nabla^2 L_{\mathcal{S}}(w)$
- In the small learning rate regime, we note that SGD behaves similar to GD and thus goes after the smaller eigenvalue direction in such case



(a) Small learning rate regime

(b) Moderate learning rate regime

Figure 1: Illustration for the 2-D example studied in Section 3. Here $\kappa = 4$ and $w_0 = (0.6, 0.6)$. **(a)**: Small learning rate regime. The small learning rate is $0.1/\kappa$. In this regime SGD and GD behave similarly and they both converge along e_2. **(b)**: Moderate learning rate regime. The initial moderate learning rate is $\eta = 1.1/\kappa$ and the decayed learning rate is $\eta' = 0.1/\kappa$. In this regime GD converges along e_2 but SGD converges along e_1, the larger eigenvalue direction of the data matrix. Please refer to Section 3 for further discussions.

Table of Contents

Directional Bias of SGD/ GD: Main Results

Theorem 1: The directional bias of SGD with moderate LR, informal

Suppose $d \geq$poly(n). Denote $v = \dfrac{n}{\sqrt{d}}$ (which is small). Then with high probability it holds that $\lambda_1 > \lambda_2 + \Theta(\nu), \lambda_{n-1} > \lambda_n + \Theta(\nu)$. Suppose the initialization is set such that $x_i^T(w_0 - w_*) \neq 0$ for every $i \in [n]$. Consider SGD with the following moderate learning rate scheme

$$\eta_k = \begin{cases} \eta \in \left(\dfrac{b}{\lambda_1 - \Theta(v)}, \dfrac{b}{\lambda_2 + \Theta(v)}\right), \ k = 1, \ldots, k_1; \\ \eta' \in \left(0, \dfrac{b}{2\lambda_1}\right), \ k = k_1 + 1, \ldots, k_2, \end{cases}$$

then for ϵ such that poly$(\epsilon) > \nu$, there exist $k_1 = \mathcal{O}\left(\log \dfrac{1}{\epsilon} + k_2\right)$ and $k_2 > 0$ such that with high probability the output of SGD $w^{sgd} := w_{k_2}$ satisfies

$$(1 - \epsilon) \cdot \gamma_1 \leq \frac{(P(w^{sgd} - w_*))^T \cdot XX^T \cdot P(w^{sgd} - w_*)}{||P(w^{sgd} - w_*)||_2^2} \leq \gamma_1$$

, where γ_1 is the largest eigenvalue of the data matrix XX^T.

Directional Bias of SGD/ GD: Main Results

Theorem 2: The directional bias of GD with moderate LR, informal

Under the same conditions as Theorem 1, consider GD with the following moderate or small learning rate scheme

$$\eta_k \in \left(0, \frac{n}{2\lambda_1 + \Theta(v)}\right), \ k = 1, \ldots, k_2$$

, then for any $\epsilon > 0$, if $k_2 > \mathcal{O}\left(\log \frac{1}{\epsilon}\right)$, then with high probability the output of GD $w^{gd} := w_{k_2}$ satisfies

$$\gamma_n \leq \frac{(P(w^{gd} - w_*))^T \cdot XX^T \cdot P(w^{gd} - w_*)}{||P(w^{gd} - w_*)||_2^2} \leq (1 + \epsilon) \cdot \gamma_n$$

, where γ_n is the smallest eigenvalue of the data matrix XX^T restricted in the column space of X.

Directional Bias of SGD/ GD: Main Results

Theorem 2: The directional bias of GD with moderate LR, informal

Under the same conditions as Theorem 1, consider GD with the following moderate or small learning rate scheme

$$\eta_k \in \left(0, \frac{n}{2\lambda_1 + \Theta(v)}\right), \ k = 1, \ldots, k_2$$

, then for any $\epsilon > 0$, if $k_2 > \mathcal{O}\left(\log \frac{1}{\epsilon}\right)$, then with high probability the output of GD $w^{gd} := w_{k_2}$ satisfies

$$\gamma_n \leq \frac{(P(w^{gd} - w_*))^T \cdot XX^T \cdot P(w^{gd} - w_*)}{||P(w^{gd} - w_*)||_2^2} \leq (1 + \epsilon) \cdot \gamma_n$$

, where γ_n is the smallest eigenvalue of the data matrix XX^T restricted in the column space of X.

Thus Theorem 1 and 2 suggest that, when projected onto the data manifold, SGD and GD converge to the optimum along the largest and smallest eigenvalue direction respectively.

Directional Bias of SGD/ GD: Main Results

Theorem 3: The directional bias of SGD with small LR, informal)

Theorem 2 applies to (SGD) with the following small learning rate scheme

$$\eta_k = \eta' \in \left(0, \frac{b}{2\lambda_1 + \Theta(v)}\right), \ k = 1, \ldots, k_2$$

Directional Bias of SGD/ GD: Main Results

Theorem 3: The directional bias of SGD with small LR, informal)

Theorem 2 applies to (SGD) with the following small learning rate scheme

$$\eta_k = \eta' \in \left(0, \frac{b}{2\lambda_1 + \Theta(v)}\right), \; k = 1, \ldots, k_2$$

Theorem 4: Effects of the directional bias, informal (Gist)

- In the moderate learning rate regime, there is a separation between the test error of SGD and that of GD. In detail, early stopped SGD finds a nearly optimal solution thanks to its particular directional bias. In contrast, early stopped GD can only find a suboptimal one.
- In the small learning rate regime, however, SGD no longer admits the dedicated directional bias for moderate learning rate. Instead it behaves similarly as GD, and hence outputs suboptimal solutions when early stopping is adopted.

Directional Bias of SGD/ GD: Main Results

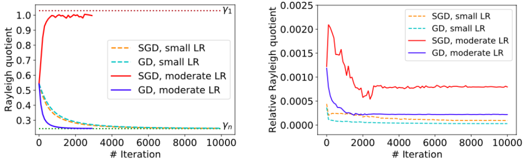
- Under the practically used moderate learning rate, there is a separation between the generalization abilities of SGD and GD

Directional Bias of SGD/ GD: Main Results

- Under the practically used moderate learning rate, there is a separation between the generalization abilities of SGD and GD
- This work gives a theoretical justification of the phenomenon that SGD outperforms GD when the learning rate is moderate

Directional Bias of SGD/ GD: Main Results

- Under the practically used moderate learning rate, there is a separation between the generalization abilities of SGD and GD
- This work gives a theoretical justification of the phenomenon that SGD outperforms GD when the learning rate is moderate



(a) Linear regression on synthetic data (b) Neural network on a subset of FashionMNIST

Figure 2: Comparison of the (relative) Rayleigh quotients. **(a)**: A linear regression example. We randomly draw 100 samples from a 10,000-dimensional space as described in Section 4, where $\zeta \sim \mathcal{U}([0.5, 1])$. The small learning rate scheme is specified by $(\eta', k_2) = (0.2, 10^4)$, and the moderate learning rate scheme is specified by $(\eta, \eta', k_1, k_2) = (1.05, 0.1, 2 \times 10^3, 3 \times 10^3)$. Numerical results show the Rayleigh quotient converges to its maximum for SGD with moderate learning rate, and converges to its minimum for GD and SGD with small learning rate, which verifies Theorems 1, 2 and 3. **(b)**: A neural network example. The plots are averaged over 10 runs. We randomly draw 2,000 samples from FashionMNIST as the training set. The model is a 5-layer convolutional neural network. The small learning rate scheme is specified by $(\eta', k_2) = (10^{-3}, 10^4)$, and the moderate learning rate scheme is specified by $(\eta, \eta', k_1, k_2) = (10^{-2}, 10^{-3}, 2.5 \times 10^3, 10^4)$. Since neural network is non-convex, we compare the *relative Rayleigh quotient* of the concerned algorithms, i.e., the Rayleigh quotient of the convergence directions divided by the maximum absolute eigenvalue of the Hessian (see Appendix D.3).

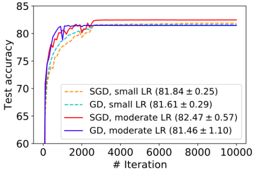


Figure 3: The test accuracy of a neural network on a subset of FashionMNIST. The plots are averaged over 10 runs. The experimental setting is identical to that in Figure 2(b). The plots show that SGD with moderate learning rate achieves the highest test accuracy, and GD and SGD with small learning rate perform similarly, but are worse than the former.

Table of Contents

References

Jingfeng Wu et al. "Direction Matters: On the Implicit Bias of Stochastic Gradient Descent with Moderate Learning Rate". In: *International Conference on Learning Representations*.

Thank you!