Global Convergence and Stability of SGD

Vivak Patel, Shushu Zhang, and Bowen Tian

NeurIPS 2022

What is SGD? Why do we care?

What is SGD? Why do we care?

Stochastic Gradient Descent (SGD) is a **foundational algorithm** used to train machine learning models that has shown incredible generalization ability over most of its competitors.

What is SGD? Why do we care?

Stochastic Gradient Descent (SGD) is a **foundational algorithm** used to train machine learning models that has shown incredible generalization ability over most of its competitors.

Thus, understanding **when SGD succeeds or fails** at training is essential to robust and reliable learning.

What is SGD? Why do we care?

Stochastic Gradient Descent (SGD) is a **foundational algorithm** used to train machine learning models that has shown incredible generalization ability over most of its competitors.

Thus, understanding **when SGD succeeds or fails** at training is essential to robust and reliable learning.

In particular, understanding SGD's **global convergence behavior** is the starting point for any further analysis of SGD.

PROBLEM

Unfortunately, existing global convergence analyses of SGD **do not** apply to realistic machine learning models.

Failure on Archetype Examples

In our work, we show that existing theory fails for:

Failure on Archetype Examples

In our work, we show that existing theory fails for:

→ A **simple feed forward network** for binary classification with three layers trained with a standard approach.

Failure on Archetype Examples

In our work, we show that existing theory fails for:

→ A **simple feed forward network** for binary classification with three layers trained with a standard approach.

→ A **simple recurrent neural network** for binary classification with a temporal length of four trained with a standard approach.

Failure on Archetype Examples

In our work, we show that existing theory fails for:

- → A **simple feed forward network** for binary classification with three layers trained with a standard approach.
- → A **simple recurrent neural network** for binary classification with a temporal length of four trained with a standard approach.
- → A trivial **Poisson regression problem** for modeling count data given some feature information.

How do we fail?

How exactly does the theory fail on these examples?

How do we fail?

How exactly does the theory fail on these examples?

Assume the **gradient** of the objective (e.g., empirical risk minimization problem) is globally Lipschitz continuous or (L_0, L_1)-smooth.

How do we fail?

How exactly does the theory fail on these examples?

Assume the **gradient** of the objective (e.g., empirical risk minimization problem) is globally Lipschitz continuous or (L_0, L_1)-smooth.

Assume the **variance of the stochastic gradients** is globally bounded, satisfies expected smoothness, or even exists.

What are better assumptions?

What are better assumptions?

Assume the **gradient** is **locally Hölder continuous** for some power $\alpha \in (0, 1]$.

What are better assumptions?

Assume the **gradient** is **locally Hölder continuous** for some power $\alpha \in (0, 1]$.

Assume the $\alpha + 1$-**moment of the stochastic gradients** is upper bounded by an **arbitrary** upper semi-continuous function.

What are better assumptions?

Assume the **gradient** is **locally Hölder continuous** for some power $\alpha \in (0,1]$.

Assume the $\alpha + 1$**-moment of the stochastic gradients** is upper bounded by an **arbitrary** upper semi-continuous function.

Our aforementioned examples **satisfy** these assumptions.

Why is this setting so complicated?

Why is this setting so complicated?

A priori, SGD's iterates can behave **arbitrarily**: converge to a stationary point, converge to a non-stationary point, enter a cycle, have a limit cycle, distinct limit supremum and infimum, vary i.o. between infinity and a finite point, and they can diverge.

Why is this setting so complicated?

A priori, SGD's iterates can behave **arbitrarily**: converge to a stationary point, converge to a non-stationary point, enter a cycle, have a limit cycle, distinct limit supremum and infimum, vary i.o. between infinity and a finite point, and they can diverge.

Current analysis techniques **do not** generalize trivially or readily to this setting!

What do we do?

For SGD with diminishing, matrix-valued step sizes:

What do we do?

For SGD with diminishing, matrix-valued step sizes:

We develop **two novel analysis techniques** that generalize to this more **realistic and complicated setting**.

What do we do?

For SGD with diminishing, matrix-valued step sizes:

We develop **two novel analysis techniques** that generalize to this more **realistic and complicated setting**.

Under this setting, we show that SGD's iterates **either** converge to a stationary point or diverge with probability one. **See Theorem 2.**

What do we do?

For SGD with diminishing, matrix-valued step sizes:

We develop **two novel analysis techniques** that generalize to this more **realistic and complicated setting**.

Under this setting, we show that SGD's iterates **either** converge to a stationary point or diverge with probability one. **See Theorem 2.**

Under an additional, interesting assumption, we show that the objective function **cannot** diverge even if the iterates diverge. **See Theorem 3.**

SUMMARY

We provide the **first** global convergence analysis of SGD under **realistic assumptions** for differentiable machine learning problems.