Data-Dependent Stability of Stochastic Gradient Descent

A review

M. Zadem K. Abouda TH. Yen Vu

MAP670L - Generalisation properties of algorithms in ML

Academic Year 2019/2020

Introduction

Algorithmic stability

Stability	
Uniform	On-average
- Restrictive - Not data-dependent	- Less restrictive - Data-dependent

- Pessimistic generalization bound.
- Insufficient to give deeper theoretical insights.

Outline

Notations

- $\mathcal{Z}$: Example space
- Training and testing examples are drawn iid from a probability distribution $\mathcal{D}$ over $\mathcal{Z}$.
- $S = \{z_i\}_{i=1}^m \sim \mathcal{D}^m$: A training set drawn according to $\mathcal{D}$.
- $A : \mathcal{Z}^m \mapsto \mathcal{H}$: A learning algorithm, $\mathcal{H}$ a parameter space
- $f(\boldsymbol{w}, z)$: Loss incurred by predicting with parameter $\boldsymbol{w} \in \mathcal{H}$

Outline

Uniform stability of a randomized algorithm

Definition

A randomized algorithm A is ϵ-uniformly stable if :
for all datasets $S, S^{(i)} \in \mathcal{Z}^m$ such that S and $S^{(i)}$ differ in the $i-th$ example we have

$$\sup_{z \in \mathcal{Z}, i \in [m]} \left\{ \mathbb{E}\left[f\left(A_S, z\right) - f\left(A_{S^{(i)}}, z\right) \right] \right\} \leqslant \epsilon$$

Outline

Uniform stability implies generalization in expectation

Theorem

Let A be ϵ-uniformly stable. Then,

$$\left| \underset{S,A}{\mathbb{E}} \left[\widehat{R}_S\left(A_S\right) - R\left(A_S\right) \right] \right| \leqslant \epsilon$$

Outline

On-average stability of a randomized algorithm

Here we introduce a data-dependent notion of stability. Therefore, we denote stability as a function of $\epsilon(\theta)$ where θ is a parameter of the algorithm that captures caracteristics of the given dataset.

Definition

A randomized algorithm A is ϵ-uniformly stable if :
for all datasets $S, S^{(i)} \in \mathcal{Z}^m$ such that S and $S^{(i)}$ differ in the $i-th$ example we have

$$\sup_{i \in [m]} \left\{ \mathbb{E}_{A} \mathbb{E}_{S,z} \left[f\left(A_S, z\right) - f\left(A_{S^{(i)}}, z\right) \right] \right\} \leqslant \epsilon(\theta)$$

Outline

On-average stability implies generalization in expectation

Theorem

Let A be $\epsilon(\theta)$-uniformly stable. Then,

$$\left| \mathbb{E}_{S,A} \left[\widehat{R}_S(A_S) - R(A_S) \right] \right| \leqslant \epsilon(\theta)$$

Outline

SGD algorithm

- The studied SGD algorithm is the following: given a training set $S = \{z_i\}_{i=1}^{m} \sim^{iid} \mathcal{D}^m$, step sizes $\{\alpha_t\}_{t=1}^{T}$, random indices $I = \{j_t\}_{t=1}^{T}$, and an initialization point $\boldsymbol{w}_1$, perform updates

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \alpha_t \nabla f\left(\boldsymbol{w}_t, z_{jt}\right)$$

SGD algorithm

- The studied SGD algorithm is the following: given a training set $S = \{z_i\}_{i=1}^m \sim^{iid} \mathcal{D}^m$, step sizes $\{\alpha_t\}_{t=1}^T$, random indices $I = \{j_t\}_{t=1}^T$, and an initialization point $\boldsymbol{w}_1$, perform updates

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \alpha_t \nabla f\left(\boldsymbol{w}_t, z_{jt}\right)$$

- The variance of the stochastic gradients verifies

$$\mathop{\mathbb{E}}_{S,z_z}\left[\left\|\nabla f\left(\boldsymbol{w}_{S,t}, z\right) - \nabla R\left(\boldsymbol{w}_{S,t}\right)\right\|^2\right] \leqslant \sigma^2 \quad \forall t \in [T]$$

Loss function

The loss function in the following theorems is assumed to be

- Non-negative
- Lipschitz: A loss function f is L-Lipschitz if $\|\nabla f(\boldsymbol{w}, z)\| \leqslant L, \forall \boldsymbol{w} \in \mathcal{H}$ and $\forall z \in \mathcal{Z}$
- β-Smooth: A loss function is β-smooth if $\forall \boldsymbol{w}, \boldsymbol{v} \in \mathcal{H}$ and $\forall z \in \mathcal{Z}, \|\nabla f(\boldsymbol{w}, z) - \nabla f(\boldsymbol{v}, z)\| \leqslant \beta \|\boldsymbol{w} - \boldsymbol{v}\|$

Outline

Generalisation bound for convex loss

Theorem

Assume that f is convex, and that the step for SGD are such that $\alpha_t = \frac{c}{\sqrt{t}} \leqslant \frac{1}{\beta}, \forall t \in [T]$. Then SGD is $\epsilon(\mathcal{D}, \boldsymbol{w}_1)$-on-average stable with

$$\epsilon(\mathcal{D}, w_1) = \mathcal{O}\left(\sqrt{c\left(R\left(w_1\right) - R^*\right)} \cdot \frac{\sqrt[4]{T}}{m} + c\sigma\frac{\sqrt{T}}{m}\right)$$

Generalisation bound for convex loss

- The Data-dependant bound is tighter than the uniform bound given by [1] which corresponds to $\epsilon = \mathcal{O}(\sqrt{T/m})$ for the same step

Generalisation bound for convex loss

- The Data-dependant bound is tighter than the uniform bound given by [1] which corresponds to $\epsilon = \mathcal{O}(\sqrt{T/m})$ for the same step
- In the case that σ is not too large, the obtained bound depends on the initial point for the SGD which makes the choice of the starting point crucial to obtaining better stability

Generalisation bound for convex loss

- The Data-dependant bound is tighter than the uniform bound given by [1] which corresponds to $\epsilon = \mathcal{O}(\sqrt{T/m})$ for the same step
- In the case that σ is not too large, the obtained bound depends on the initial point for the SGD which makes the choice of the starting point crucial to obtaining better stability
- On the other hand, for a large σ, the second term in the sum is dominant, and the bound is equivalent to the uniform bound.

Outline

Generalisation bound for non-convex loss

Theorem

Assume that $f(\cdot, z) \in [0,1]$ and has a ρ-Lipschitz Hessian, and that step sizes of a form $\alpha_t = \frac{c}{t}$ satisfy $c \leqslant \min\left\{\frac{1}{\beta}, \frac{1}{4(2\beta\ln(T))^2}\right\}$ Then SGD is $\epsilon(\mathcal{D}, \boldsymbol{w}_1)$-on-average stable with

$$\epsilon(\mathcal{D}, \boldsymbol{w}_1) \leqslant \frac{1 + \frac{1}{c\gamma}}{m}\left(2cL^2\right)^{\frac{1}{1+c\gamma}}\left(\mathop{\mathbb{E}}_{S,A}\left[R\left(A_S\right)\right] \cdot T\right)^{\frac{c\gamma}{1+c\gamma}}$$

where

$$\gamma := \mathcal{O}\left(\min\left\{\beta, \mathop{\mathbb{E}}_{z}\left[\left\|\nabla^2 f\left(\boldsymbol{w}_1, z\right)\right\|_2\right] + \Delta^*_{1,\sigma^2}\right\}\right)$$

$$\Delta^*_{1,\sigma^2} := \rho(c\sigma + \sqrt{c\left(R\left(\boldsymbol{w}_1\right) - R^*\right)})$$

Generalisation bound for non-convex loss

- The quantity γ expresses how the curvature of the loss function at the initial point affects the stability of the algorithm. Consequently, starting from a point that is less curved yields a better bound and thus smaller generalisation error. This statement corroborates the intuition.
- In [1], a similar bound was given, but instead of γ, the given bound included a Lipschitz constant relative to the gradient of f. This constant fails to represent the data dependency.

Generalisation bound for non-convex loss

Corollary

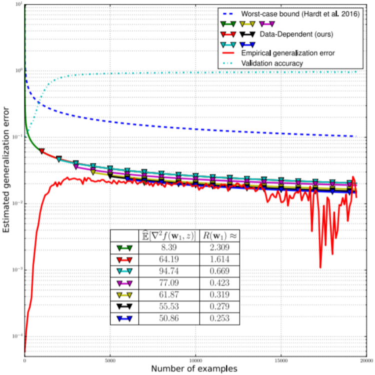
Under conditions of tthe previous theorem we have that SGD is $\epsilon\left(\mathcal{D}, \boldsymbol{w}_1\right)$-on-average stable with

$$\epsilon\left(\mathcal{D}, \boldsymbol{w}_1\right) = \mathcal{O}\left(\frac{1+\frac{1}{c\gamma}}{m}\left(R\left(\boldsymbol{w}_1\right)\cdot T\right)^{\frac{c\gamma}{1+c\gamma}}\right)$$

For $R\left(\boldsymbol{w}_1\right) \to 0$ we have that $\epsilon\left(\mathcal{D}, \boldsymbol{w}_1\right) \to 0$. As the initialisation point error diminishes, the generalisation error is close to zero. The uniform stability is incapable of showing these results seeing that it is distribution independent.

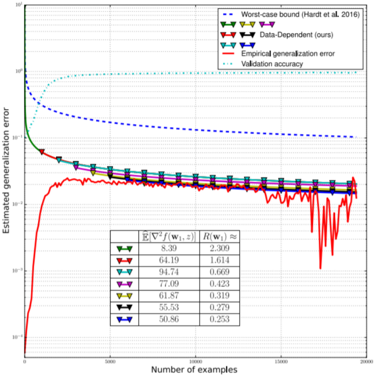
Remarks

- This article gives an average bound, while in [1] an uniform worst-case bound was given, which is more confident but far away from being optimal.



Remarks

- In non-convex case (the case for most Neural Nets), trade-off between curvature of the initial point $\widehat{\mathbb{E}}[\nabla^2 f(\mathbf{w}_1, z)]$ and the risk at the initial point $R(\mathbf{w}_1)$.



Remarks

- This article gives an average bound, while in [1] an uniform worst-case bound was given, which is more confident but far away from being optimal.
- In non-convex case (the case for most Neural Nets), trade-off between curvature of the initial point $\widehat{\mathbb{E}}[\nabla^2 f(\mathbf{w}_1, z)]$ and the risk at the initial point $R(\mathbf{w}_1)$.
- A tighter bound can be obtained by taking the minimum of the two bounds.
- Require a knowledge on the prior distribution of the dataset for exact bound, otherwise only an estimation.

Conclusion

- Provide new data-dependent stability bounds for SGD on convex and non-convex loss functions.
- Better initial point (lower objective) makes the algorithm more stable and generalizes better.
- In non-convex case, starting from a point in a less-curved region yields a better generalisation error.

References I

Ilja Kuzborskij, Christoph H. Lampert.
Data-Dependent Stability of Stochastic Gradient Descent
ICML 2018, arXiv:1703.01678v4, 19 Feb 2018.

Moritz Hardt, Benjamin Recht, and Yoram Singer.
Train faster, generalize better: Stability of stochastic gradient descent, 2015.