Stability

Egor Shulgin

Moscow Institute of Physics and Technology
Department of Control and Applied Math

March 13, 2019

Paper presentation

Train faster, generalize better: Stability of stochastic gradient descent 259 2015
M Hardt, B Recht, Y Singer
ICML 2016, 1225-1234

arXiv link

Overview

Stochastic Gradient Method (SGM) for Machine Learning

Given n labeled examples $S = (z_1, \ldots, z_n)$ where $z_i \in Z$, consider a *decomposable* objective function

$$f(w) = \frac{1}{n} \sum_{i=1}^{n} f(w; z_i),$$

where $f(w; z_i)$ denotes the *loss* of w on the example z_i. The stochastic gradient update for this problem with *learning rate* $\alpha_t > 0$ is given by

$$w_{t+1} = G_{f,\alpha_t}(w_t) = w_t - \alpha_t \nabla_w f(w_t; z_{i_t}).$$

Default method in practice

- scalable
- easy-to-implement
- robust

  works well across many diferent domains

Stability of randomized iterative algorithms

$\mathcal{D}$ – unknown distribution over space Z. We receive a sample $S = (z_1, \ldots, z_n)$ of n examples drawn i.i.d. from $\mathcal{D}$. Our goal is to find a model w with small *population risk*: $R[w] \stackrel{\text{def}}{=} \mathbb{E}_{z\sim\mathcal{D}} f(w; z)$

- *Empirical risk*: $R_S[w] \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f(w; z_i)$
- *Generalization error*: $R_S[w] - R[w]$
- *Expected generalization error*: $\epsilon_{\text{gen}} \stackrel{\text{def}}{=} \mathbb{E}_{S,A}[R_S[A(S)] - R[A(S)]]$

Definition

A randomized algorithm A is ϵ-*uniformly stable* if for all data sets $S, S' \in Z^n$ such that S and S' differ in at most one example, we have $\sup_z \mathbb{E}_A [f(A(S); z) - f(A(S'); z)] \leq \epsilon$.

Theorem [Generalization in expectation]

Let A be ϵ-uniformly stable. Then, $|\mathbb{E}_{S,A} [R_S[A(S)] - R[A(S)]]| \leq \epsilon$.

Properties of update rules

We consider general update rules of the form $G\colon \Omega \to \Omega$ which map a point $w \in \Omega$ in the parameter space to another point $G(w)$. The most common update is the gradient update rule $G(w) = w - \alpha \nabla f(w)$,

$$G(w) = w - \alpha \nabla f(w) \; ,$$

where $\alpha \geq 0$ is a step size and $f\colon \Omega \to \mathbb{R}$ is a function that we want to optimize.

Definition

An update rule is η-*expansive* if $\sup\limits_{v,w\in\Omega} \frac{\|G(v)-G(w)\|}{\|v-w\|} \leq \eta$

Definition

An update rule is σ-*bounded* if $\sup\limits_{w\in\Omega} \|w - G(w)\| \leq \sigma$.

Properties of update rules

Lemma (Growth recursion)

Fix an arbitrary sequence of updates $G_1, \dots, G_T$ and another sequence $G'_1, \dots, G'_T$. Let $w_0 = w'_0$ be a starting point in Ω and define $\delta_t = \|w'_t - w_t\|$ where w_t, w'_t are defined recursively through

$$w_{t+1} = G_t(w_t) \qquad w'_{t+1} = G'_t(w'_t)\,. \qquad (t > 0)$$

Then, we have the recurrence relation
$\delta_0 = 0$,

$$\delta_{t+1} \leq \begin{cases} \eta\delta_t & G_t = G'_t \text{ is } \eta\text{-expansive} \\ \min(\eta, 1)\delta_t + 2\sigma_t & G_t \text{ and } G'_t \text{ are } \sigma\text{-bounded}, \\ & G_t \text{ is } \eta \text{ expansive} \end{cases}$$

Expansion properties of stochastic gradients

Definition

We say that f is *L-Lipschitz* if for all points u in the domain of f we have $\|\nabla f(x)\| \leq L$. This implies that $|f(u) - f(v)| \leq L\|u - v\|$.

Lemma

Assume that f is L-Lipschitz. Then, the gradient update $G_{f,\alpha}$ is (αL)-bounded.

Function properties

- *convex*: $f(u) \geq f(v) + \langle \nabla f(v), u - v \rangle$
- γ*-strongly convex*: $f(u) \geq f(v) + \langle \nabla f(v), u - v \rangle + \frac{\gamma}{2}\|u - v\|^2$
- β*-smooth*: $\|\nabla f(u) - \nabla f(v)\| \leq \beta\|u - v\| \ \forall \ u, v \in \Omega$

Expansion properties of stochastic gradients

Lemma (Growth recursion)

Assume that f is β-smooth. Then the following properties hold.

- $G_{f,\alpha}$ is $(1 + \alpha\beta)$-expansive.
- Assume in addition that f is convex. Then, for any $\alpha \leq 2/\beta$, the gradient update $G_{f,\alpha}$ is 1-expansive.
- Assume in addition that f is γ-strongly convex. Then, for $\alpha \leq \frac{2}{\beta+\gamma}$, $G_{f,\alpha}$ is $\left(1 - \frac{\alpha\beta\gamma}{\beta+\gamma}\right)$-expansive.

Convex optimization

Theorem

Assume that the loss function $f(\cdot\,; z)$ is β-smooth, convex and L-Lipschitz for every z. Suppose that we run SGM with step sizes $\alpha_t \leq 2/\beta$ for T steps. Then, SGM satisfies uniform stability with

$$\epsilon_{\mathrm{stab}} \leq \frac{2L^2}{n} \sum_{t=1}^{T} \alpha_t \, .$$

Strongly Convex optimization

$L = \sup_{w \in \Omega} \sup_z \|\nabla f(w; z)\|_2$

Theorem

Assume that the loss function $f(\cdot\,; z)$ is γ-strongly convex and β-smooth for all z. Suppose we run the projected SGM iteration with constant step size $\alpha \leq 1/\beta$ for T steps. Then, SGM satisfies uniform stability with

$$\epsilon_{\text{stab}} \leq \frac{2L^2}{\gamma n}$$

Theorem

Assume that the loss function $f(\cdot\,; z) \in [0, 1]$ is γ-strongly convex has gradients bounded by L as in, and is β-smooth function for all z. Suppose we run SGM with step sizes $\alpha_t = \frac{1}{\gamma t}$. Then, SGM has uniform stability of

$$\epsilon_{\text{stab}} \leq \frac{2L^2 + \beta\rho}{\gamma n}$$

Non-convex optimization

Theorem

Assume that $f(\cdot; z) \in [0, 1]$ is an L-Lipschitz and β-smooth loss function for every z. Suppose that we run SGM for T steps with monotonically non-increasing step sizes $\alpha_t \leq c/t$. Then, SGM has uniform stability with

$$\epsilon_{\text{stab}} \leq \frac{1 + 1/\beta c}{n - 1}(2cL^2)^{\frac{1}{\beta c+1}} T^{\frac{\beta c}{\beta c+1}}$$

In particular, omitting constant factors that depend on β, c, and L, we get

$$\epsilon_{\text{stab}} \lessapprox \frac{T^{1-1/(\beta c+1)}}{n}.$$

Stability-inducing operations

- Weight Decay and Regularization

Definition

Let $f : \Omega \to \Omega$, be a differentiable function. We define the *gradient update with weight decay at rate* μ as $G_{f,\mu,\alpha}(w) = (1 - \alpha\mu)w - \alpha\nabla f(w)$.

Lemma

Assume that f is β-smooth. Then, $G_{f,\mu,\alpha}$ is $(1 + \alpha(\beta - \mu))$-expansive.

- Gradient Clipping
- Dropout
- Model Averaging.

Convex risk minimization

Definition (Optimization error)

$\epsilon_{\text{opt}}(w) \overset{\text{def}}{=} \mathbb{E}\left[R_S[w] - R_S[w_\star^S]\right]$ where $w_\star^S = \arg\min_w R_S[w]$.

$\mathbb{E}[R[w]] \leq \mathbb{E}[R_S[w]] + \epsilon_{\text{stab}} \leq \mathbb{E}[R_S[w_\star^S]] + \epsilon_{\text{opt}}(w) + \epsilon_{\text{stab}}.$

Lemma

$\mathbb{E}[R_S[w_\star^S]] \leq R[w_\star]$ where $w_\star = \arg\min_w R[w]$.

Theorem (classical result)

Assume we run stochastic gradient descent with constant stepsize α on a convex function $R[w] = \mathbb{E}_z[f(w;z)]$. Assume further that $\|\nabla f(w;z)\| \leq L$ and $\|w_0 - w_\star\| \leq D$ for some minimizer $w_\star$ of R. Let $\bar{w}_T$ denote the average of the T iterates of the algorithm. Then we have

$$R[\bar{w}_T] \leq R[w_\star] + \tfrac{1}{2}\frac{D^2}{T\alpha} + \tfrac{1}{2}L^2\alpha\,.$$

Convex risk minimization

Corollary (from classical result)

Let f be a convex loss function satisfying $\|\nabla f(w,z)\| \le L$ and let $w_\star$ be a minimizer of the population risk $R[w] = \mathbb{E}_z f(w;z)$. Suppose we make a single pass of SGM over the sample $S = (z_1,\dots,z_n)$ with a suitably chosen fixed step size starting from a point w_0 that satisfies $\|w_0 - w_\star\| \le D$. Then, the average $\bar{w}_n$ of the iterates satisfies $\mathbb{E}[R[\bar{w}_n]] \le R[w_\star] + \frac{DL}{\sqrt{n}}$.

Proposition

Let $S = (z_1,\dots,z_n)$ be a sample of size n. Let f be a β-smooth convex loss function satisfying $\|\nabla f(w,z)\| \le L$ and let $w_\star^S$ be a minimizer of the empirical risk $R_S[w] = \frac{1}{n}\sum_{i=1}^n f(w;z_i)$. Suppose we run T steps of SGM with suitably chosen step size from a starting point w_0 that satisfies $\|w_0 - w_\star^S\| \le D$. Then, the average $\bar{w}_T$ over the iterates satisfies

$$\mathbb{E}[R[\bar{w}_T]] \le \mathbb{E}[R_S[w_\star^S]] + \frac{DL}{\sqrt{n}}\sqrt{\frac{n+2T}{T}}\,.$$