Characterization of Convex Objective Functions and Optimal Expected Convergence Rates for SGD

Marten van Dijk [1] Lam M. Nguyen [2] Phuong Ha Nguyen [1] Dzung T. Phan [2]

Abstract

We study Stochastic Gradient Descent (SGD) with diminishing step sizes for convex objective functions. We introduce a definitional framework and theory that defines and characterizes a core property, called curvature, of convex objective functions. In terms of curvature we can derive a new inequality that can be used to compute an optimal sequence of diminishing step sizes by solving a differential equation. Our exact solutions confirm known results in literature and allows us to fully characterize a new regularizer with its corresponding expected convergence rates.

1. Introduction

It is well-known that the following stochastic optimization problem

$$\min_{w\in\mathbb{R}^d} \left\{ F(w) = \mathbb{E}[f(w;\xi)] \right\}, \tag{1}$$

where ξ is a random variable obeying some distribution can be solved efficiently by stochastic gradient descent (SGD) (Robbins & Monro, 1951). The SGD algorithm is described in Algorithm 1.

If we define $f_i(w) := f(w;\xi_i)$ for a given training set $\{(x_i,y_i)\}_{i=1}^n$ and ξ_i is a random variable that is defined by a single random sample (x,y) pulled uniformly from the training set, then empirical risk minimization reduces to

$$\min_{w\in\mathbb{R}^d} \left\{ F(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w) \right\}. \tag{2}$$

Algorithm 1 Stochastic Gradient Descent (SGD) Method

```
**Initialize:** $w_0$
**Iterate:**
**for** $t = 0, 1, 2, \dots$ **do**
    Choose a step size (i.e., learning rate) $\eta_t > 0$.
    Generate a random variable $\xi_t$.
    Compute a stochastic gradient $\nabla f(w_t;\xi_t)$.
    Update the new iterate $w_{t+1} = w_t - \eta_t \nabla f(w_t;\xi_t)$.
**end for**
```

Problem (2), which can also be solved by gradient descent (GD) (Nesterov, 2004; Nocedal & Wright, 2006), has been discussed in many supervised learning applications (Hastie et al., 2009). As an important note, a class of variance reduction methods (Le Roux et al., 2012; Defazio et al., 2014; Johnson & Zhang, 2013; Nguyen et al., 2017) has been proposed for solving (2) in order to reduce the computational cost. Since all these algorithms explicitly use the finite sum form of (2), they and GD may not be efficient for very large scale machine learning problems. In addition, variance reduction methods are not applicable to (1). Hence, SGD is an important algorithm for very large scale machine learning problems and the problems for which we cannot compute the exact gradient. It is proved that SGD has a sub-linear convergence rate with convergence rate $\mathcal{O}(1/t)$ in the strongly convex cases (Bottou et al., 2016; Nguyen et al., 2018; Gower et al., 2019), and $\mathcal{O}(1/\sqrt{t})$ in the general convex cases (Nemirovsky & Yudin, 1983; Nemirovski et al., 2009), where t is the number of iterations.

In this paper we derive convergence properties for SGD applied to (1) for many different flavors of convex objective functions F. We introduce a new notion called ω-convexity where ω denotes a function with certain properties (see Definition 1). Depending on ω, F can be convex or strongly convex, or something in between, i.e., F is not strongly convex but is “better” than “plain” convex. This region between plain convex and strongly convex F will be characterized by a new notion for convex objective functions called curvature (see Definition 3).

Convex and non-convex optimization are well-known problems in the literature (see e.g. (Schmidt et al., 2016; Defazio et al., 2014; Schmidt & Roux, 2013; Reddi et al., 2016)).

[1] Department of Electrical and Computer Engineering, University of Connecticut, CT, USA. [2] IBM Research, Thomas J. Watson Research Center, NY, USA. Correspondence to: Marten van Dijk <marten.van_dijk@uconn.edu>, Lam M. Nguyen <LamNguyen.MLTD@ibm.com>, Phuong Ha Nguyen <phuongha.ntu@gmail.com>, Dzung T. Phan <phandu@us.ibm.com>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

The problem in the middle range of convexity and non-convexity called quasi-convexity has been studied and analyzed (Hazan et al., 2015). Convex optimization is a basic and well studied primitive in machine learning. In some applications, the optimization problems may be non-strongly convex but may have specific structure of convexity. For example, a classical least squares problem with

$$f_i(w) = (a_i^T w - b_i)^2$$

is convex for some data parameters $a_i \in \mathbb{R}^d$ and $b_i \in \mathbb{R}$. When an ℓ_2-norm regularization $\|w\|^2$ is employed (ridge regression), the regularized problem becomes strongly convex. Group sparsity is desired in some domains, one can add an $\ell_{2,1}$ regularization $\sum_i \|w_{[i]}\|$ (Wright et al., 2009). This problem is no longer strongly convex, but it should be "stronger" than plain convex.

To the best of our knowledge, there are no specific results or studies in the middle range of convexity and strong convexity. In this paper, we provide a new definition of convexity and study its convergence analysis.

In our analysis, the following assumptions are required.[1]

Assumption 1 (L-smooth)**.** *$f(w;\xi)$ is L-smooth for every realization of ξ, i.e., there exists a constant $L > 0$ such that, $\forall w, w' \in \mathbb{R}^d$,*

$$\|\nabla f(w;\xi) - \nabla f(w';\xi)\| \le L\|w - w'\|. \tag{3}$$

Assumption 1 implies that F is also L-smooth.

Assumption 2 (convex)**.** *$f(w;\xi)$ is convex for every realization of ξ, i.e., $\forall w, w' \in \mathbb{R}^d$,*

$$f(w;\xi) - f(w';\xi) \ge \langle \nabla f(w';\xi), (w - w')\rangle.$$

Assumption 2 implies that F is also convex.

We assume that $f(w;\xi)$ is L-smooth and convex for every realization of ξ. Then, according to (Nesterov, 2004), for all $w, w' \in \mathbb{R}^d$,

$$\begin{aligned}&\|\nabla f(w;\xi) - \nabla f(w';\xi)\|^2 \\ &\qquad \le L\langle \nabla f(w;\xi) - \nabla f(w';\xi), w - w'\rangle. \end{aligned} \tag{4}$$

The requirement of existence of unbiased gradient estimators, i.e., $\mathbb{E}_\xi[\nabla f(w;\xi)] = \nabla F(w)$, for any fixed w is in need for applying SGD to the general form (1).

Contributions and Outline.

1- Our convergence analysis of SGD for convex objective functions is based on a new recurrence on the expected convergence rate stated in Lemma 1 (Sec. 2). As a side result this recurrence is used to show in Theorem 1 (Sec. 2) that, for convex objective functions, SGD converges with probability 1 (almost surely) to a global minimum (if one exists). The w.p.1 result is an adaptation of the w.p.1 result in (Nguyen et al., 2018) for the strongly convex case.

2- We introduce a new framework and define ω-convex objective functions in Definition 1 (Sec. 3) and the curvature of convex objective functions in Definition 3 (Sec. 3). We show how strongly convex and "plain" convex objective functions fit this picture, as extremes on either end (curvature 1 and 0, respectively).

3- In Theorem 2 we introduce a new regularizer $G(w)$, for $w \in \mathbb{R}^d$, with curvature $1/2$. It penalizes small $\|w\|$ much less than the 2-norm $\|w\|^2$ regularizer and it penalizes large $\|w\|$ much more than the 2-norm $\|w\|^2$ regularizer. This allows us to enforce more tight control on the size of w when minimizing a convex objective function.

4- By using the recurrence of Lemma 1 (Sec. 2) and a new inequality for ω-convex objective functions, we are able to analyze the expected convergence rate of SGD in Sec. 4. We characterize the expected convergence rate as a solution to a differential equation. Our analysis matches existing theory; for strongly convex F we obtain a 2-approximate optimal solution and for "plain" convex F with no curvature we obtain an optimal step size of order $O(t^{-1/2})$. For the new regularizer we get a precise expression for the optimal step size and expected convergence rates.

2. Convex Optimization

In convex optimization we only assume that $f(w;\xi)$ is L-smooth and convex for every realization of ξ. Under these assumptions, the objective function $F(w) = \mathbb{E}_\xi[f(w;\xi)]$ is also L-smooth and convex. However, the assumptions are too weak to guarantee a unique global minimum for $F(w)$. For this reason we introduce

$$\mathcal{W}^* = \{w_* \in \mathbb{R}^d \;:\; \forall_{w\in\mathbb{R}^d}\; F(w_*) \le F(w)\}$$

as the set of all w_* that minimize $F(.)$. The set $\mathcal{W}^*$ may be empty implying that there does not exist a global minimum. If $\mathcal{W}^*$ is not empty, it may contain many vectors w_* implying that a global minimum exists but that it is not unique.

Assumption 3 (global minimum exists)**.** *Objective function F has a global minimum.*

This assumption implies that

$$\forall_{w_*\in\mathcal{W}^*}\; \nabla F(w_*) = 0 \;\text{ and}$$
$$\exists_{F_{min}} \forall_{w_*\in\mathcal{W}^*}\; F(w_*) = F_{min}.$$

[1] Here and in the remainder of the paper $\|\cdot\|$ stands for the 2-norm.

With respect to $\mathcal{W}^*$ we define

$$N = \sup_{w_* \in \mathcal{W}^*} \mathbb{E}_\xi[\|\nabla f(w_*; \xi)\|^2].$$

Assumption 4 (finite N)**.** *We assume N is finite.*

Without explicitly stating, each of the lemmas and theorems in the remainder of this paper assume Assumptions 1, 2, 3, and 4.

For the recursively computed values w_t, we define

$$Y_t = \inf_{w_* \in \mathcal{W}^*} \|w_t - w_*\|^2 \text{ and } E_t = F(w_t) - F(w_*).$$

These quantities measure the convergence rate towards one of the global minima.

Lemma 1. *Let $\mathcal{F}_t$ be a σ-algebra which contains $w_0, \xi_0, w_1, \xi_1, \ldots, w_{t-1}, \xi_{t-1}, w_t$. Assume $\eta_t \leq 1/L$. For any given $w_* \in \mathcal{W}^*$, we have*

$$\mathbb{E}[Y_{t+1}|\mathcal{F}_t] \leq \mathbb{E}[Y_t|\mathcal{F}_t] - 2\eta_t(1 - \eta_t L)E_t + 2\eta_t^2 N. \quad (5)$$

The proof of Lemma 1 is presented in supplemental material A. Moreover, an immediate application is given by the next theorem (its proof is in supplemental material B).

Theorem 1. *Consider SGD with step size sequence such that*

$$0 < \eta_t \leq \frac{1}{L}\ , \sum_{t=0}^{\infty} \eta_t = \infty \textit{ and } \sum_{t=0}^{\infty} \eta_t^2 < \infty.$$

Then, the following holds w.p.1 (almost surely)

$$F(w_t) - F(w_*) \to 0,$$

where w_* is any optimal solution of $F(w)$.

We note that the convergence w.p.1. in (Nguyen et al., 2018) only works in the strongly convex case while our above theorem holds for the case where the objective function is general convex.

3. Convex Flavors

We define functions

$$a(w) = F(w) - F(w_*) = F(w) - F_{min} \quad (6)$$

and

$$b(w) = \inf_{w_* \in \mathcal{W}^*} \|w - w_*\|^2. \quad (7)$$

Notice that $a(w) = 0$ if and only if $w \in \mathcal{W}^*$ and $b(w) = 0$ if and only if $w \in \mathcal{W}^*$.

We introduce a new definition based on $a(w)$ and $b(w)$ which characterizes a multitude of convex flavors of objective functions:

Definition 1 (ω-convex)**.** *Let $a : \mathbb{R}^d \to [0, \infty)$ and $b : \mathbb{R}^d \to [0, \infty)$ be smooth functions. Let $\omega : [0, \infty) \to [0, \infty)$ be $\cap$-convex (i.e. $\omega''(\epsilon) < 0$) and strictly increasing (i.e., $\omega'(\epsilon) > 0$). Let $\mathcal{B} \subseteq \mathbb{R}^d$ be a convex set (e.g., a sphere or $\mathbb{R}^d$ itself) such that, first,*

$$\omega(a(w)) \geq b(w) \textit{ for all } w \in \mathcal{B}$$

and, second, $a(w) = 0$ implies both $b(w) = 0$ and $w \in \mathcal{B}$. Then we call the pair of functions (a, b) ω-separable over $\mathcal{B}$.

If objective function F gives rise to a pair of functions (a, b) as defined by (6) and (7) which is ω-separable over $\mathcal{B}$, then we call F ω-convex over $\mathcal{B}$.

The objective function being ω-convex is a subcase of the Error Bound Condition (see Equation (1) in (Bolte et al., 2017)) which only requires ω to be non-decreasing (i.e., $\omega'(\cdot) \geq 0$). The Holderian Error Bound (HEB) (also called Local Error Bound, Local Error Bound Condition, or Lojasiewicz Error bound) is a subcase of the Error Bound Condition where $\omega(\epsilon) = c\epsilon^p$ where $c > 0$ and $p \in (0, 2]$ (see Definition 1 of (Xu et al., 2016) where the reader should notice that $b(w)$ in (7) represents the *squared* Euclidean distance implying that ω in our notation is the square of the ω in Equation (6) of (Xu et al., 2016)). When $p = 1$, HEB becomes the Quadratic Growth Condition (Drusvyatskiy & Lewis, 2018); in particular, strong convex objective functions satisfy the Quadratic Growth Condition (see also our Lemma 3).

It turns out that our ω-convex notion and HEB are different as they are not a subclass of each other, but they do have an intersection: Notice that for $p \in (1, 2]$, $\omega(\epsilon) = c\epsilon^p$ is not $\cap$-convex and does not satisfy Definition 1, hence HEB is not a subclass of ω-convexity. Also ω-convexity is not a subclass of HEB; for example, our special case of ω-convexity as defined in Definition 3 and later studied in the rest of the paper is different from HEB (only $r = \infty$ in Lemma 9 and Theorem 3 reflects HEB). HEB and ω-convexity intersect for $\omega(\epsilon) = c\epsilon^p$ with $p \in (0, 1]$. The results in this paper imply that $p \in (0, 1]$ corresponds to the range of plain convex to strong convex objective functions for which we analyze the expected convergence rates of SGD with optimal step sizes (given the recurrence of Lemma 1). To the best of our knowledge there is no existing work on analyzing the convergence of SGD with this ω-convex notion or with HEB.

We list a couple of useful insights (proofs are in supplemental material C.1):

Lemma 2. *Let $a : \mathbb{R}^d \to [0, \infty)$ and $b : \mathbb{R}^d \to [0, \infty)$ be smooth functions and let $\mathcal{B} \subseteq \mathbb{R}^d$ such that $a(w) = 0$ implies $b(w) = 0$ and $w \in \mathcal{B}$.*

For $\epsilon \geq 0$, we define

$$\delta(\epsilon) = \sup_{p:\mathbb{E}_p[a(w)]\leq \epsilon} \mathbb{E}_p[b(w)],$$

where p represents a probability distribution over $w \in \mathcal{B}$.

Assuming $\delta(\epsilon) < \infty$ for $\epsilon \geq 0$, $\delta(\cdot)$ is $\cap$-convex and strictly increasing with $\delta(0) = 0$. Furthermore,

1. *The pair of functions (a, b) is δ-separable over $\mathcal{B}$.*
2. *The pair of functions (a, b) is ω-separable over $\mathcal{B}$ if and only if $\omega(\epsilon) \geq \delta(\epsilon)$ for all $\epsilon \geq 0$.*

The lemma shows that δ is the "minimal" function ω for which (a, b) is ω-separable over $\mathcal{B}$.

The lemma also shows that $a(w)$ and $b(w)$ are not separable over $\mathcal{B}$ for any function $\omega(\cdot)$ if and only if $\delta(\epsilon) = \infty$ for $\epsilon > 0$. This is only possible if $\mathcal{B}$ is not bounded within some sphere (e.g., $\mathcal{B} = \mathbb{R}^d$). If $\mathcal{B}$ is bounded, then there always exists a function $\omega(\cdot)$ such that $a(w)$ and $b(w)$ are ω-separable over $\mathcal{B}$ (e.g., $\omega(x) = \delta(x)$ as defined above).

For convex objective functions, we see in practice that the type of distributions p in the definition of $\delta(\cdot)$ can be restricted to having their probability mass within a bounded sphere $\mathcal{B}$ of w vectors. In the analysis of the convergence rate this corresponds to assuming all $w_t \in \mathcal{B}$ (see next section). As discussed above this makes $\delta(\epsilon)$ finite and we are guaranteed to be able to apply the definitional framework as introduced here.

The relationship towards strongly convex objective functions is given below.

Definition 2 (μ-strongly convex)**.** *The objective function $F : \mathbb{R}^d \to \mathbb{R}$ is called μ-strongly convex, if for all $w, w' \in \mathbb{R}^d$,*

$$F(w) - F(w') \geq \langle \nabla F(w'), (w - w')\rangle + \frac{\mu}{2}\|w - w'\|^2.$$

For $w' \in \mathcal{W}^*$, $\nabla F(w') = 0$. So, for a μ-strongly convex objective function f, $F(w) - F(w_*) \geq \frac{\mu}{2}\|w - w_*\|^2$ for all $w_* \in \mathcal{W}^*$ (notice that $\mathcal{W}^*$ has exactly one vector w_* representing the global minimum). This implies that $\frac{2}{\mu}a(w) \geq b(w)$ for (a, b) defined by (6) and (7):

Lemma 3. *If objective function F is μ-strongly convex, then F is ω-convex over $\mathbb{R}^d$ for function $\omega(x) = \frac{2}{\mu}x$.*

We will show that existing convergence results for strongly convex objective functions can be derived from assuming the weaker ω-convexity property for appropriately selected ω as given in the above lemma.

In order to prove bounds on the expected convergence rate for any ω-convex objective function, we will use the following inequality:

Lemma 4. *Let $a : \mathbb{R}^d \to [0, \infty)$ and $b : \mathbb{R}^d \to [0, \infty)$ be smooth functions and assume they are ω-separable over $\mathcal{B}$ for some $\cap$-convex and strictly increasing function ω and convex set $\mathcal{B} \subseteq \mathbb{R}^d$. Let p be a probability distribution over $\mathcal{B}$. Then, for all $0 < x$,*

$$\frac{\mathbb{E}_p[b(w)]}{\omega'(x)} \leq (\frac{\omega(x)}{\omega'(x)} - x) + \mathbb{E}_p[a(w)].$$

Proof. Since $\omega(a(w)) \geq b(w)$ for all $w \in \mathcal{B}$,

$$\mathbb{E}_p[b(w)] \leq \mathbb{E}_p[\omega(a(w))].$$

Since $\omega(\cdot)$ is $\cap$-convex,

$$\mathbb{E}_p[\omega(a(w))] \leq \omega(\mathbb{E}_p[a(w)]).$$

Since ω is $\cap$-convex and strictly increasing, for all $x > 0$ and $y > 0$, $\omega(y) \leq \omega(x) + \omega'(x)(y - x)$. Substituting $y = \mathbb{E}_p[a(w)]$ yields

$$\omega(\mathbb{E}_p[a(w)]) \leq \omega(x) + \omega'(x)[\mathbb{E}_p[a(w)] - x].$$

Combining the sequence of inequalities, rearranging terms, and dividing by $\omega'(x)$ proves the statement. □

When applying Lemma 4 we will be interested in bounding $\frac{\omega(x)}{\omega'(x)} - x$ from above while maximizing $\frac{1}{\omega'(x)}$. That is, we want to investigate the behavior of

$$v(\eta) = \sup\{\frac{1}{\omega'(x)} \ : \ \frac{\omega(x)}{\omega'(x)} - x \leq \eta\}.$$

Notice that the derivative of $\frac{\omega(x)}{\omega'(x)} - x$ is equal to $\frac{-\omega(x)\omega''(x)}{\omega'(x)^2} \geq 0$, and the derivative $\frac{1}{\omega'(x)}$ is equal to $\frac{-\omega''(x)}{\omega'(x)^2} \geq 0$. This implies that $v(\eta)$ is increasing and is alternatively defined as

$$v(\eta) = \frac{1}{\omega'(x)} \text{ where } \eta = \frac{\omega(x)}{\omega'(x)} - x. \tag{8}$$

Corollary 1. *Given the conditions in Lemma 4 with $v(\eta)$ defined as in (8), for all $0 < \eta$,*

$$v(\eta)\mathbb{E}_p[b(w)] \leq \eta + \mathbb{E}_p[a(w)].$$

We are able to use this corollary to provide upper bounds on the expected convergence rate if $v(\eta)$ has a "nice" form as given in the next definition and lemma.

Definition 3. *For $h \in (0, 1]$, $r > 0$, and $\mu > 0$, define*

$$\omega_{h,r,\mu}(x) = \begin{cases} \frac{2}{\mu h}(x/r)^h, & \text{if } x \leq r, \text{ and} \\ \frac{2}{\mu h} + \frac{2}{\mu}((x/r) - 1), & \text{if } x > r. \end{cases}$$

We say functions $a(w)$ and $b(w)$ are separable by a function with curvature $h \in (0, 1]$ over $\mathcal{B}$ if for some $r, \mu > 0$ they *are $\omega_{h,r,\mu}$-separable over $\mathcal{B}$. We define objective function F to have curvature $h \in (0,1]$ over $\mathcal{B}$ if its associated functions $a(w)$ and $b(w)$ are $\omega_{h,r,\mu}$-separable over $\mathcal{B}$ for some $r, \mu > 0$.*

The proof of the following lemma is in supplemental material C.2.

Lemma 5. *For $v(\eta)$ defined as in (8) and $\omega = \omega_{h,r,\mu}$,*

$$v(\eta) = \beta h \eta^{1-h} \text{ with } \beta = \frac{\mu}{2} h^{-h}(1-h)^{-(1-h)} r^h,$$

for $0 \leq \eta \leq r$.

If set $\mathcal{B}$ is bounded by a sphere, then the supremum s_a and s_b of values $a(w)$ and $b(w)$, $w \in \mathcal{B}$, exist (since $a(w)$ and $b(w)$ are assumed smooth and continuous everywhere). If $s_b > 0$, then trivially, for $h \in (0,1]$,

$$\frac{h\eta}{s_b}\mathbb{E}_p[b(w)] \leq \eta + \mathbb{E}_p[a(w)].$$

In other words a linear function $v(\eta) = \beta h \cdot \eta$ for some constant $\beta > 0$ (e.g., the one of Lemma 5 for $h \downarrow 0$) does not give any information. Nevertheless taking the limit $h \downarrow 0$ will turn out useful in showing that, for convex objective functions with no curvature, a $\eta_t = O(t^{-1/2})$ diminishing step size is optimal in the sense that the asymptotic behavior of the expected convergence rate cannot be improved.

Concluding the above discussions, convex objective functions can be classified in different convex flavors: either having a curvature $h \in (0,1]$ (where $h = 1$ is implied by strong convexity) or having no such curvature. In the latter case we abuse notation and say that the objective function has "curvature $h = 0$". With this extended definition, any convex objective function has a curvature $h \in [0,1]$ over $\mathcal{B}$ and, by Corollary 1 and Lemma 5, there exist constants β and r such that, for $0 \leq \eta \leq r$,

$$\beta h \eta^{1-h}\mathbb{E}_p[b(w)] \leq \eta + \mathbb{E}_p[a(w)] \tag{9}$$

for distributions p over $\mathcal{B}$.

In supplemental material C.3 we show the following example which introduces a new regularizer which makes a convex objective function have curvature $h = 1/2$ over $\mathbb{R}^d$:

Theorem 2. *Let*

$$F(w) = H(w) + \lambda G(w)$$

be our objective function where $\lambda > 0$, $H(w)$ is a convex function, and

$$G(w) = \sum_{i=1}^{d}[e^{w_i} + e^{-w_i} - 2 - w_i^2].$$

Then, F is ω-convex over $\mathbb{R}^d$ for $\omega(x) = \frac{2}{\mu h}x^h$ with $h = 1/2$ and $\mu = \frac{\lambda}{9d}$. The associated $v(\eta)$ as defined in (8) is equal to

$$v(\eta) = \beta h \eta^{1-h} \text{ with } \beta = \frac{\mu}{2} h^{-h}(1-h)^{-(1-h)} = \mu,$$

for $\mu \geq 0$.

Function $G(w)$ is of interest as it severely penalizes large $|w_i|$ due to the exponent functions, while for small $|w_i|$ the corresponding term in the sum of $G(w)$ is very small (in fact, we subtract w_i^2 in order to make it smaller; if we would not have subtracted the w_i^2, then G changes into $G(w) + \|w\|^2$ which is strongly convex). This has the possibility to force the global minimum to smaller size when compared to, e.g., $G(w) = \|w\|$ or $G(w) = \|w\|^2$. The price of moving away from using $G(w) = \|w\|^2$ is moving away from having a strong convex objective function, i.e., the curvature over $\mathbb{R}^d$ is reduced from $h = 1$ to $h = 1/2$. In the next section we show that this leads to a slower expected convergence rate.

4. Expected Convergence Rate

We notice that w_t is coming from a distribution determined by the randomness used in the SGD algorithm when computing $w_0, \xi_0, w_1, \xi_1, \ldots, w_{t-1}, \xi_{t-1}$. Let us call this distribution p^t. Then,

$$\begin{aligned}\mathbb{E}[E_t] &= \mathbb{E}[F(w_t) - F(w_*)] \\ &= \mathbb{E}_{p^t}[F(w) - F(w_*)] = \mathbb{E}_{p^t}[a(w)].\end{aligned} \tag{10}$$

Since distribution p^t determines w_t, we also have

$$\begin{aligned}\mathbb{E}[Y_t] &= \mathbb{E}[\inf_{w_* \in \mathcal{W}^*}\|w_t - w_*\|^2] \\ &= \mathbb{E}_{p^t}[\inf_{w_* \in \mathcal{W}^*}\|w - w_*\|^2] = \mathbb{E}_{p^t}[b(w)].\end{aligned} \tag{11}$$

Both $\mathbb{E}_{p^t}[a(w)]$ and $\mathbb{E}_{p^t}[b(w)]$ measure the expected convergence rate. In practice we want to get close to a global minimum and therefore $\mathbb{E}_{p^t}[a(w)]$ is preferred since $a(w_t) = F(w_t) - F(w_*)$.

For $\eta_t \leq \frac{1}{2L}$, Lemma 1 shows

$$\mathbb{E}[Y_{t+1}|\mathcal{F}_t] \leq \mathbb{E}[Y_t|\mathcal{F}_t] - \eta_t E_t + 2\eta_t^2 N.$$

After taking the full expectation and rearranging terms this gives

$$\eta_t \mathbb{E}[E_t] \leq \mathbb{E}[Y_t] - \mathbb{E}[Y_{t+1}] + 2\eta_t^2 N. \tag{12}$$

By assuming F is ω-convex over $\mathcal{B}$ and p^t has zero probability mass outside $\mathcal{B}$, application of Lemma 4 and Corollary 1 after substituting (10) and (11) gives

$$v(\eta)\mathbb{E}[Y_t] \leq \eta + \mathbb{E}[E_t].$$

The right hand side can be upper bounded by using (12):

$$\begin{aligned}\eta_t v(\eta)\mathbb{E}[Y_t] &\leq \eta_t\eta + \eta_t\mathbb{E}[E_t] \\ &\leq \eta_t\eta + \mathbb{E}[Y_t] - \mathbb{E}[Y_{t+1}] + 2\eta_t^2 N.\end{aligned}$$

After reordering terms and using $\eta = \eta_t$ we obtain the recurrence:

Lemma 6. *If F is ω-convex over $\mathcal{B}$ and p^t has zero probability mass outside $\mathcal{B}$ (or equivalently the SGD algorithm never generates a w_t outside $\mathcal{B}$), then for $\eta_t \leq \frac{1}{2L}$,*

$$\mathbb{E}[Y_{t+1}] \leq (1 - v(\eta_t)\eta_t)\mathbb{E}[Y_t] + (2N+1)\eta_t^2, \qquad (13)$$

where $v(\eta_t)$ is defined by (8).

We notice that if the SGD algorithm has proceeded to the t-th iteration, then we know that, due to finite step sizes during the iterations so far, the SGD algorithm has only been able to push the starting vector w_0 to some w_t within some bounded sphere $\mathcal{B}_t$ around w_0. So, if F is ω-convex over $\mathcal{B}_i$ for $1 \leq i \leq t$, then we may apply the above recurrence up to iteration t. Of course, ideally we do not need to assume this and have $\mathcal{B} = \mathbb{R}^d$ as in Theorem 2.

Assumption 5 ($\mathcal{B}$-bounded). *Until sufficient convergence has been achieved, the SGD algorithm never generates a w_t outside $\mathcal{B}$.*

In supplemental material D we prove the following lemmas that solve recurrence (13).

Lemma 7. *Suppose that the objective function is ω-convex over $\mathcal{B}$ and let $v(\eta)$ be defined as in (8). Let $n(\cdot)$ be a decreasing step size function representing $n(t) = \eta_t \leq \frac{1}{2L}$. Define*

$$\begin{aligned}M(t) &= \int_{x=0}^{t} n(x)v(n(x))dx \text{ and} \\ C(t) &= \exp(-M(t))\int_{x=0}^{t}\exp(M(x))n(x)^2dx.\end{aligned}$$

Then recurrence (13) implies

$$\mathbb{E}[Y_t] \leq A\cdot C(t) + B\cdot\exp(-M(t))$$

for constants

$$A = (2N+1)\exp(n(0))$$

and

$$B = (2N+1)\exp(M(1))n(0)^2 + \mathbb{E}[Y_0]$$

(they depend on parameter N and starting vector w_0).

Lemma 8. *A close to optimal step size can be computed by solving the differential equation*

$$\bar{C}(t) = \frac{2[-\bar{C}'(t)]^{1/2}}{v([-\bar{C}'(t)]^{1/2})}$$

and equating

$$n(t) = [-\bar{C}'(t)]^{1/2}.$$

The solution to the differential equation approaches $C(t)$ for t large enough: For all $t \geq 0$, $C(t) \leq \bar{C}(t)$. For t large enough, $C(t) \geq \bar{C}(t)/2$.

Lemma 9. *For $v(\eta) = \beta h\eta^{1-h}$ with $h \in (0,1]$, where $\beta > 0$ is a constant and $0 \leq \eta \leq r$ for some $r \in (0,\infty]$ (including the possibility $r = \infty$), we obtain*

$$\bar{C}(t) = [1/(2-h)]^{h/(2-h)}(2/\beta)^{2/(2-h)}(t+\Delta)^{-h/(2-h)}$$

for

$$n(t) = \left(\frac{2}{\beta(2-h)}\right)^{1/(2-h)}(t+\Delta)^{-1/(2-h)}$$

with

$$\Delta = \frac{2\max\{2L, 1/r\}}{\beta(2-h)}.$$

The above results show that an objective function with curvature $h = 0$ or a very small curvature does not have a fast decreasing expected convergence rate $\mathbb{E}[Y_t]$. Nevertheless, the SGD algorithm does not need to converge in Y_t. For small curvature the objective function looks very flat and we may still approach F_{min} reasonably fast.

We use the following classical argument: By (12),

$$\sum_{i=t+1}^{2t}\eta_i\mathbb{E}[E_i] \leq \mathbb{E}[Y_{t+1}] - \mathbb{E}[Y_{2t+1}] + 2N\sum_{i=t+1}^{2t}\eta_i^2.$$

Define the average

$$A_t = \frac{1}{t}\sum_{i=t+1}^{2t}\mathbb{E}[E_i].$$

For $\eta_t = n(t)$ as defined in the previous lemma,

$$\begin{aligned}n(2t)tA_t &\leq \sum_{i=t+1}^{2t}\eta_t\mathbb{E}[E_i], \\ \sum_{i=t+1}^{2t}\eta_i^2 &\leq \int_{x=t}^{2t} n(x)^2dx = \int_{x=t}^{2t}[-\bar{C}'(x)]dx \\ &= \bar{C}(t) - \bar{C}(2t)\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}[Y_{t+1}] &\leq A\cdot\bar{C}(t+1) + B\cdot\exp(-M(t+1)) \\ &\leq A\cdot\bar{C}(t) + B\cdot\exp(-M(t)).\end{aligned}$$

We derive

$$\begin{aligned}M(t) &= \int_{x=0}^{t} n(x)v(n(x))dx = \beta h\int_{x=0}^{t} n(x)^{2-h}dx \\ &= \frac{2h}{2-h}\int_{x=0}^{t}(t+\Delta)^{-1}dx \\ &= \frac{2h}{2-h}[\ln(t+\Delta) - \ln\Delta],\end{aligned}$$

hence,

$$\exp(-M(t)) = (t+\Delta)^{-2h/(2-h)}\Delta^{2h/(2-h)}.$$

Combining all inequalities yields

$$n(2t)tA_t \leq (2N+A)\cdot\bar{C}(t)+B\Delta^{2h/(2-h)}(t+\Delta)^{-2h/(2-h)}.$$

This proves the following theorem:

Theorem 3. *For an objective function with curvature $h \in (0,1]$ with associated $v(\eta) = \beta h\eta^{1-h}$, where $\beta > 0$ is a constant and $0 \leq \eta \leq r$ for some $r \in (0,\infty]$ (including the possibility $r = \infty$), a close to optimal step size is*

$$\eta_t = \left(\frac{2}{\beta(2-h)}\right)^{1/(2-h)}(t+\Delta)^{-1/(2-h)}.$$

The corresponding expected convergence rates are

$$\begin{aligned}\mathbb{E}[Y_t] &\leq A\frac{[1/(2-h)]^{h/(2-h)}(2/\beta)^{2/(2-h)}}{(t+\Delta)^{h/(2-h)}} + \\ &\quad B\frac{\Delta^{2h/(2-h)}}{(t+\Delta)^{2h/(2-h)}}, \\ \frac{1}{t}\sum_{i=t+1}^{2t}\mathbb{E}[E_i] &\leq (2N+A)A'\frac{(2t+\Delta)^{1/(2-h)}}{(t+\Delta)^{h/(2-h)}t} + \\ &\quad B\Delta^{2h/(2-h)}B'\frac{(2t+\Delta)^{1/(2-h)}}{(t+\Delta)^{2h/(2-h)}t},\end{aligned}$$

where

$$A' = [1/(2-h)]^{-(1-h)/(2-h)}(2/\beta)^{1/(2-h)},$$

and

$$B' = [1/(2-h)]^{-1/(2-h)}(2/\beta)^{-1/(2-h)}.$$

The asymptotic behavior is dominated by the terms with A and A'. This shows independence of the expected convergence rates from the starting point w_0 since $\mathbb{E}[Y_0]$ only occurs in B. We have

$$\mathbb{E}[Y_t] = O(t^{-h/(2-h)}) \text{ and } \frac{1}{t}\sum_{i=t+1}^{2t}\mathbb{E}[E_i] = O(t^{-1/(2-h)}).$$

For μ-strongly convex objective functions we have $v(\eta) = \frac{\mu}{2}h\eta^{1-h}$ for $h = 1$. Theorem 3 (after substituting constants and substituting $r = \infty$) gives, for $A = (2N+1)e^{1/(2L)}$,

$$\mathbb{E}[Y_t] \leq \frac{A}{\mu}\left[\frac{16}{(\mu t+8L)}\right] + O(t^{-2})$$

and

$$\frac{1}{t}\sum_{i=t+1}^{2t}\mathbb{E}[E_i] \leq \frac{2N+A}{\mu}\left[\frac{4(2\mu t+8L)}{(\mu t+8L)t}\right] + O(t^{-2})$$

for step size

$$\eta_t = \frac{2}{\mu t/2+4L}.$$

In (Nguyen et al., 2018), they report an optimal step size of $2/(\mu t+4L)$, hence, η_{2t} is equal to this optimal steps size for the t-th iteration and this implies that it takes a factor 2 slower to converge; this is consistent with our derivation in which we use $\bar{C}(t)$ as a 2-approximate optimal solution.

For the example in Theorem 2 with $v(\eta) = \mu h\eta^{1-h}$ for $h = 1/2$ (and $r = \infty$), we obtain, for $A = (2N+1)e^{1/(2L)}$,

$$\mathbb{E}[Y_t] \leq \frac{A}{\mu}\left[\frac{32}{3\mu t+8L}\right]^{1/3} + O(t^{-2/3})$$

and

$$\frac{1}{t}\sum_{i=t+1}^{2t}\mathbb{E}[E_i] \leq \frac{2N+A}{\mu}\left[\frac{2(6\mu t+8L)^2}{(3\mu t+8L)t^3}\right]^{1/3} + O(t^{-1})$$

for step size

$$\eta_t = \left(\frac{2}{3\mu t/2+4L}\right)^{2/3}.$$

Due to the smaller curvature we need to choose a larger step size. The expected convergence rates are $O(t^{-1/3})$ and $O(t^{-2/3})$, respectively.

For $h \downarrow 0$, we recognize the classical result which holds for all convex objective functions. In this case the theorem shows that a diminishing step size of $O(t^{-1/2})$ is close to optimal.

5. Experiments

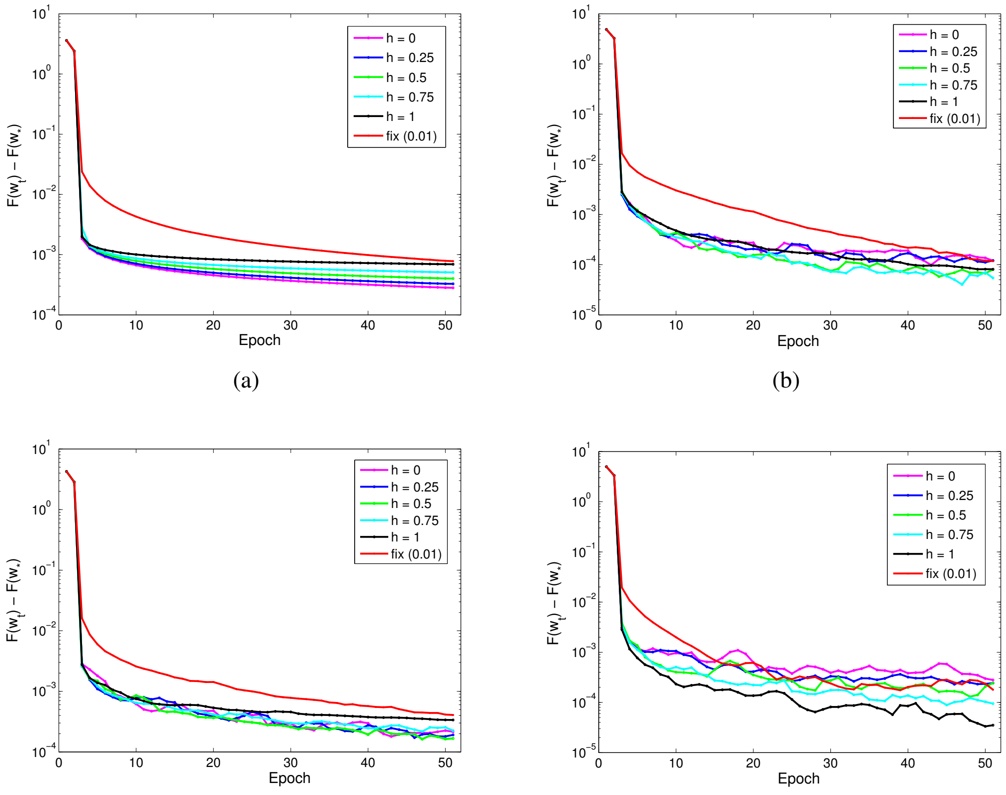
We consider both unregularized and regularized logistic regression problems with different regularizers to account for convex, ω-convex, and strongly convex cases:

$$\begin{aligned}f_i(w) &= \log(1+\exp(-y_i x_i^T w)) \text{ (convex)} \\ f_i^{(a)}(w) &= f_i(w) + \lambda\|w\| \text{ (ω-convex)} \\ f_i^{(b)}(w) &= f_i(w) + \lambda G(w) \text{ (ω-convex)} \\ f_i^{(c)}(w) &= f_i(w) + \frac{\lambda}{2}\|w\|^2 \text{ (strongly convex)},\end{aligned}$$

where the penalty parameter λ is set to 10^{-3}. We have not been able to prove the curvature of the objective function F corresponding to $f_i^{(a)}$, we address this in a general take-away in the conclusion; F corresponding to $f_i^{(b)}$ has curvature $h = 1/2$ by Theorem 2; F corresponding to $f_i^{(c)}$ has curvature $h = 1$ since it is strongly convex.

We conducted experiments on a binary classification dataset `mushrooms` from the LIBSVM website[2]. We ran Algo-

[2] http://www.csie.ntu.edu.tw/∼cjlin/libsvmtools/datasets/



(a)

(b)

(c)

(d)

Figure 1: Convergence rate for (a) $f_i(w)$ $(h = 0)$; (b) $f_i^{(a)}(w)$; (c) the new regularizer $f_i^{(b)}(w)$ $(h = 1/2)$, (d) $f_i^{(c)}(w)$ $(h = 1)$.

rithm 1 using the fixed learning rate $\eta = 10^{-2}$ and diminishing step sizes $\eta_t = 0.1/t^{1/(2-h)}$ for different values of $h = \{0, 0.25, 0.5, 0.75, 1\}$ to validate theoretical convergence rates given in Theorem 3. For each problem, we experimented with 10 seeds and took the average of function values at the end of each epoch. To smooth out function values due to the "noise" from randomness, we reported the moving mean with a sliding window of length 3 for curves in Figure 1.

The plots match the theory closely in terms of curvature values and optimal diminishing step sizes. Figure 1(a) for convex case with curvature $h = 0$ shows the best performance for a step size $\eta_t = 0.1/\sqrt{t}$ corresponding to $h = 0$. Figure 1(b) suggests that the objective function F corresponding to $f_i^{(a)}$ has curvature close to $h = 0.75$; this curvature may be due to convergence to a minimum w_* in a neighborhood where the combination of plain logistic regression and regularizer $\|w\|$ has curvature 0.75. In Figure 1(c), the stepsize rule pertaining to $h = 0.5$ yields the top performance for $f_i^{(b)}$ having curvature $h = 0.5$. Finally, the strongly convex case $f_i^{(c)}$ having curvature $h = 1$, the step size $\eta_t = 0.1/t$, i.e. $h = 1$, gives the fastest convergence.

6. Conclusion

We have provided a solid framework for analyzing the expected convergence rates of SGD for any convex objective function. Experiments match derived optimal step sizes. In particular, our new regularizer fits theoretical predictions.

The proposed framework is useful for analyzing any new regularizer, even if theoretical analysis is out-of-scope. One only needs to experimentally discover the curvature h of a new regularizer once. After curvature h is determined, the regularizer can be used for any convex problem together with a diminishing step size proportional to the optimal one as given by our theory for curvature h. Our theory predicts the resulting expected convergence rates and this can be used together with other properties of regularizers to select the one that best fits a convex problem.

Our framework characterizes a continuum from plain convex to strong convex problems and explains how the expected convergence rates of SGD vary along this continuum. Our metric 'curvature' has a one-to-one correspondence to how to choose an optimal diminishing step size and to the expected and average expected convergence rate.

Acknowledgement

The authors would like to thank the reviewers for useful suggestions which helped to improve the exposition in the paper. Phuong Ha Nguyen and Marten van Dijk were supported in part by AFOSR MURI under award number FA9550-14-1-0351.

References

Auger, A. and Hansen, N. On Proving Linear Convergence of Comparison-based Step-size Adaptive Randomized Search on Scaling-Invariant Functions via Stability of Markov Chains. *CoRR*, abs/1310.7697, 2013.

Bertsekas, D. P. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey, 2015.

Bolte, J., Nguyen, T. P., Peypouquet, J., and Suter, B. W. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, Oct 2017.

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *arXiv:1606.04838*, 2016.

Csiba, D. and Richtárik, P. Global Convergence of Arbitrary-Block Gradient Methods for Generalized Polyak-Lojasiewicz Functions. *arXiv preprint arXiv:1709.03014*, 2017.

Defazio, A., Bach, F., and Lacoste-Julien, S. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *NIPS*, pp. 1646–1654, 2014.

Drusvyatskiy, D. and Lewis, A. S. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018.

Gordji, M. E., Delavar, M. R., and De La Sen, M. On ϕ-convex functions.

Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. SGD: General Analysis and Improved Rates. *CoRR*, abs/1901.09401, 2019.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics, 2nd edition, 2009.

Hazan, E., Levy, K., and Shalev-Shwartz, S. Beyond convexity: Stochastic quasi-convex optimization. In *Advances in Neural Information Processing Systems*, pp. 1594–1602, 2015.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, pp. 315–323, 2013.

Le Roux, N., Schmidt, M., and Bach, F. A stochastic gradient method with an exponential convergence rate for finite training sets. In *NIPS*, pp. 2663–2671, 2012.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Nemirovsky, A. S. and Yudin, D. B. Problem complexity and method efficiency in optimization. 1983.

Nesterov, Y. *Introductory lectures on convex optimization : a basic course*. Applied optimization. Kluwer Academic Publ., Boston, Dordrecht, London, 2004. ISBN 1-4020-7553-7.

Nguyen, L., Nguyen, P. H., van Dijk, M., Richtarik, P., Scheinberg, K., and Takac, M. SGD and Hogwild! convergence without the bounded gradients assumption. In *ICML*, 2018.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *ICML*, 2017.

Nocedal, J. and Wright, S. J. *Numerical Optimization*. Springer, New York, 2nd edition, 2006.

Reddi, S. J., Hefny, A., Sra, S., Póczos, B., and Smola, A. J. Stochastic variance reduction for nonconvex optimization. In *ICML*, pp. 314–323, 2016.

Robbins, H. and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3): 400–407, 1951.

Schmidt, M. and Roux, N. L. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013.

Schmidt, M., Le Roux, N., and Bach, F. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, pp. 1–30, 2016.

Wright, S. J., Nowak, R. D., and Figueiredo, M. A. T. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, July 2009.

Xu, Y., Lin, Q., and Yang, T. Accelerated stochastic subgradient methods under local error bound condition. *arXiv preprint arXiv:1607.01027*, 2016.