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

Marten van Dijk   Lam M. Nguyen   Phuong Ha Nguyen   Dzung T. Phan

October 10, 2018

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) [17]. 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

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

Marten van Dijk, Department of Electrical and Computer Engineering, University of Connecticut, Storrs, CT, USA. Email: marten.van_dijk@uconn.edu

Lam M. Nguyen, IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA. Email: LamNguyen.MLTD@ibm.com

Phuong Ha Nguyen, Department of Electrical and Computer Engineering, University of Connecticut, Storrs, CT, USA. Email: phuongha.ntu@gmail.com

Dzung T. Phan, IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA. Email: phandu@us.ibm.com

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) [11, 15], has been discussed in many supervised learning applications [5]. As an important note, a class of variance reduction methods [8, 4, 7, 13] 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 that we could not compute the exact gradient. It is proved that SGD has a sub-linear convergence rate with the convergence rate of $\mathcal{O}(1/t)$ in the strongly convex cases [3, 12], and $\mathcal{O}(1/\sqrt{t})$ in the general convex cases [10, 9], 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 [18, 4, 1, 16, 14]. The problem in the middle range of convexity and non-convexity called quasi-convexity have been studied and analyzed [6]. 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

$$f_i(w) = (a_i^\top w - b_i)^2$$

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 $\sum_i \|w_{[i]}\|$ [19]. 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, $\forall w, w' \in \mathbb{R}^d$,*

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

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

Assumption 1 implies that F is also L-smooth.

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

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

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 [11], for all $w, w' \in \mathbb{R}^d$,

$$\|\nabla f(w;\xi) - \nabla f(w';\xi)\|^2 \leq L\langle \nabla f(w;\xi) - \nabla f(w';\xi), w - w' \rangle. \tag{4}$$

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- We derive in Theorem 1 (Sec. 2) a new recurrence on a expected convergence rate.

2- This recurrence is used to show in Theorem 2 (Sec. 2) that for convex objective functions SGD converges with probability 1 (almost surely) to a global minimum (if one exists).

3- 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).

4- In Theorem 3 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.

5- By using the recurrence of Theorem 1 (Sec. 2) and a new inequality for ω-convex objective functions, we are able to analyze the expected convergence rate 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

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

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

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

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.*

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.

Theorem 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*

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

Proof. The proof consists of two parts: We first derive a general inequality on Y_t and after this we take the expectation leading to the final result.

We remind the reader that $w_{t+1} = w_t - \eta_t \nabla f(w_t;\xi_t)$ from which we derive, for any given $w_* \in \mathcal{W}^*$,

$$\begin{aligned}\|w_{t+1} - w_*\|^2 &= \|w_t - w_* - \eta_t \nabla f(w_t;\xi_t)\|^2 \\ &= \|w_t - w_*\|^2 - 2\eta_t\langle \nabla f(w_t;\xi_t), w_t - w_*\rangle + \eta_t^2\|\nabla f(w_t;\xi_t)\|^2 \\ &\leq \|w_t - w_*\|^2 - 2\eta_t\langle \nabla f(w_t;\xi_t), w_t - w_*\rangle + 2\eta_t^2\|\nabla f(w_t;\xi_t) - \nabla f(w_*;\xi_t)\|^2 + 2\eta_t^2\|\nabla f(w_*;\xi_t)\|^2,\end{aligned}$$

where the inequality follows from the general inequality $\|x\|^2 \leq 2\|x-y\|^2 + 2\|y\|^2$.

Application of inequality (4), i.e., $\|\nabla f(w_t;\xi_t) - \nabla f(w_*;\xi_t)\|^2 \leq L\langle \nabla f(w_t;\xi_t) - \nabla f(w_*;\xi_t), w_t - w_*\rangle$, gives

$$\begin{aligned}&\|w_{t+1} - w_*\|^2 \\ &\leq \|w_t - w_*\|^2 - 2\eta_t\langle \nabla f(w_t;\xi_t), w_t - w_*\rangle + 2\eta_t^2 L\langle \nabla f(w_t;\xi_t) - \nabla f(w_*;\xi_t), w_t - w_*\rangle + 2\eta_t^2\|\nabla f(w_*;\xi_t)|^2 \\ &= \|w_t - w_*\|^2 - 2\eta_t(1-\eta_t L)\langle \nabla f(w_t;\xi_t), w_t - w_*\rangle + 2\eta_t^2\|\nabla f(w_*;\xi_t)\|^2 - 2\eta_t^2 L\langle \nabla f(w_*;\xi_t), w_t - w_*\rangle.\end{aligned}$$

The convexity assumption states $\langle \nabla f(w_t;\xi_t), w_t - w_*\rangle \geq f(w_t;\xi_t) - f(w_*;\xi_t)$ and this allows us to further develop our derivation: For $\eta_t \leq 1/L$,

$$\begin{aligned}&\|w_{t+1} - w_*\|^2 \\ &\leq \|w_t - w_*\|^2 - 2\eta_t(1-\eta_t L)[f(w_t;\xi_t) - f(w_*;\xi_t)] + 2\eta_t^2\|\nabla f(w_*;\xi_t)\|^2 - 2\eta_t^2 L\langle \nabla f(w_*;\xi_t), w_t - w_*\rangle.\end{aligned}$$

Let w_*^t be such that $\|w_t - w_*^t\|^2 = \inf_{w_* \in \mathcal{W}^*} \|w_t - w_*\|^2 = Y_t$ (here, we assume for simplicity that the infinum can be realized for some w_*^t and we note that the argument below can be made general with small adaptations). Notice that

$$Y_{t+1} = \inf_{w_* \in \mathcal{W}^*} \|w_{t+1} - w_*\|^2 \leq \|w_{t+1} - w_*^t\|^2.$$

By using $w_* = w_*^t$ in the previous derivation, we obtain

$$Y_{t+1} \leq Y_t - 2\eta_t(1-\eta_t L)[f(w_t;\xi_t) - f(w_*^t;\xi_t)] + 2\eta_t^2\|\nabla f(w_*^t;\xi_t)\|^2 - 2\eta_t^2 L\langle \nabla f(w_*^t;\xi_t), w_t - w_*^t\rangle.$$

Now, we take the expectation with respect to $\mathcal{F}_t$. Notice that $F(w_t) = \mathbb{E}_\xi[f(w_t;\xi)] = \mathbb{E}_\xi[f(w_t;\xi)|\mathcal{F}_t]$ and $F(w_*^t) = \mathbb{E}_\xi[f(w_*^t;\xi)] = \mathbb{E}_\xi[f(w_*^t;\xi)|\mathcal{F}_t]$. This yields

$$\mathbb{E}[Y_{t+1}|\mathcal{F}_t] \leq \mathbb{E}[Y_t|\mathcal{F}_t] - 2\eta_t(1-\eta_t L)[F(w_t) - F(w_*^t)] + 2\eta_t^2\mathbb{E}[\|\nabla f(w_*^t;\xi_t)\|^2|\mathcal{F}_t] - 2\eta_t^2 L\langle \nabla F(w_*^t), w_t - w_*\rangle.$$

Since $\nabla F(w_*^t) = 0$ and $F(w_*^t) = F(w_*) = F_{min}$ for all $w_* \in \mathcal{W}^*$, we obtain

$$\begin{aligned}\mathbb{E}[Y_{t+1}|\mathcal{F}_t] &\leq \mathbb{E}[Y_t|\mathcal{F}_t] - 2\eta_t(1-\eta_t L)[F(w_t) - F(w_*)] + 2\eta_t^2\mathbb{E}[\|\nabla f(w_*^t;\xi_t)\|^2|\mathcal{F}_t] \\ &\leq \mathbb{E}[Y_t|\mathcal{F}_t] - 2\eta_t(1-\eta_t L)[F(w_t) - F(w_*)] + 2\eta_t^2 N,\end{aligned}$$

where the last inequality follows from the definition of N. The theorem follows after substituting $E_t = F(w_t) - F(w_*)$. □

An immediate application is given by the next theorem (its proof is in supplemental material A).

Theorem 2. *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 [12] 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

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

and

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

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 $\mathcal{B}$.

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

Lemma 1. *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

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

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 subsection). 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 2. *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 3. *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 3 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 3 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 B.2.

Lemma 4. *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 4 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 function 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 4, there exists a 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 B.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 3. *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 $w_0, \xi_0, w_1, \xi_1, \dots, w_{t-1}, \xi_{t-1}$. Let us call this distribution p^t. Then,

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

Since distribution p^t determines w_t, we also have

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

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 $\eta_t \leq \frac{1}{2L}$, Theorem 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 3 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):

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

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

Lemma 5. *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 $\eta_t \leq \frac{1}{2L}$,*

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

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 3.

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 C we prove the following lemmas that solve recurrence (13).

Lemma 6. *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*

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

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 7. *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 8. *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,

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

and

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

We derive

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

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 4. *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 4 (after substituting constants) 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 [12], 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 3 with $v(\eta) = \mu h\eta^{1-h}$ for $h = 1/2$, 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:

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

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 $f_i^{(a)}$; F corresponding to $f_i^{(b)}$ has curvature $h = 1/2$ by Theorem 3; F corresponding to $f_i^{(c)}$ has curvature $h = 1$ since it is strongly convex.

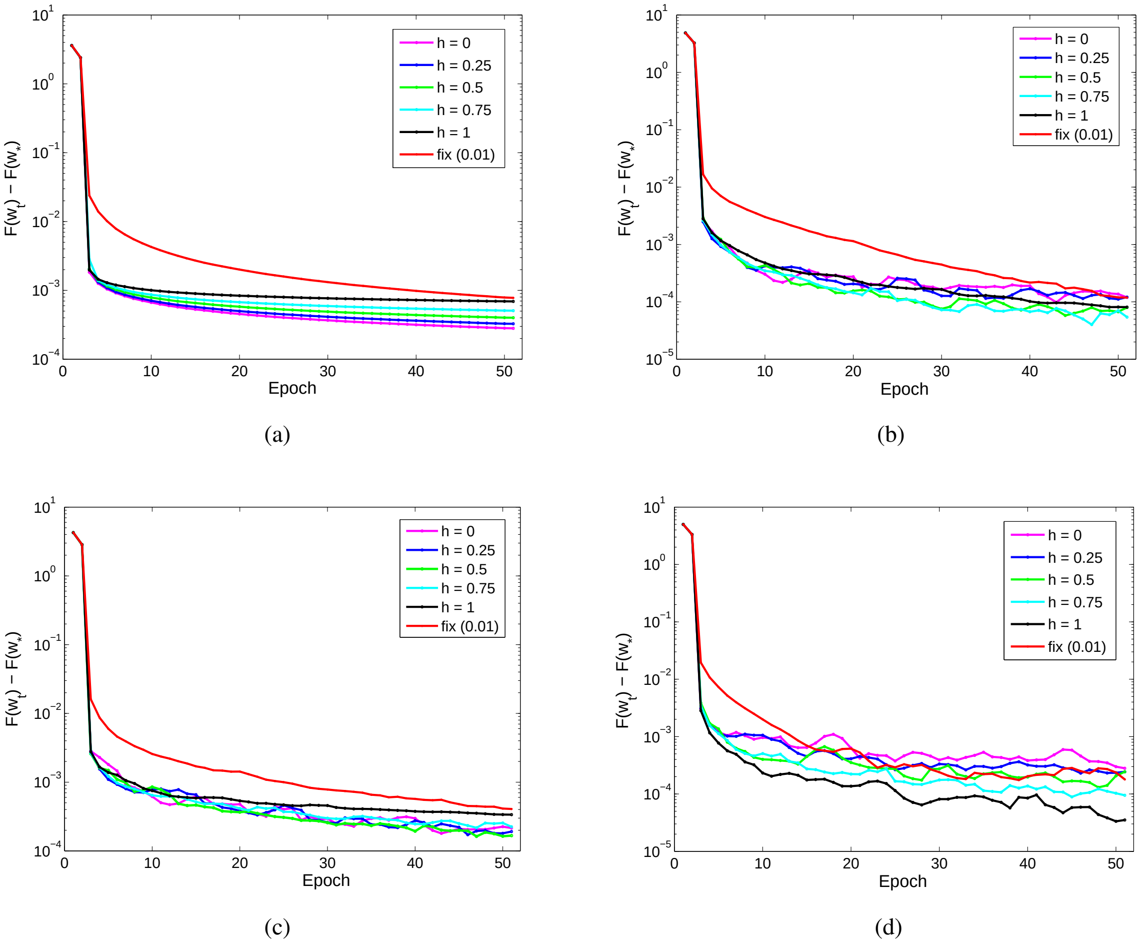
We conducted experiments on a binary classification dataset `mushrooms` from the LIBSVM website[2]. We ran Algorithm 1 using the fixed learning rate $\eta = 10^{-2}$ and diminishing step sizes $\eta_t = 0.1/t^{1/(2-h)}$ for different values of $h = \{0, 0.25, 0.5, 0.75, 1\}$ to validate theoretical convergence rates given in Theorem 4. 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 $h = 0.75$. 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 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.

[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) $f_i^{(c)}(w)$ ($h = 1$).

References

[1] Zeyuan Allen-Zhu and Yang Yuan. Improved SVRG for Non-Strongly-Convex or Sum-of-Non-Convex Objectives. In *ICML*, pages 1080–1089, 2016.

[2] Dimitri P. Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey, 2015.

[3] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *arXiv:1606.04838*, 2016.

[4] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *NIPS*, pages 1646–1654, 2014.

[5] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics, 2nd edition, 2009.

[6] Elad Hazan, Kfir Levy, and Shai Shalev-Shwartz. Beyond convexity: Stochastic quasi-convex optimization. In *Advances in Neural Information Processing Systems*, pages 1594–1602, 2015.

[7] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, pages 315–323, 2013.

[8] Nicolas Le Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *NIPS*, pages 2663–2671, 2012.

[9] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

[10] Arkadii Semenovich Nemirovsky and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

[11] Yurii Nesterov. *Introductory lectures on convex optimization : a basic course*. Applied optimization. Kluwer Academic Publ., Boston, Dordrecht, London, 2004.

[12] Lam Nguyen, Phuong Ha Nguyen, Marten van Dijk, Peter Richtarik, Katya Scheinberg, and Martin Takac. SGD and Hogwild! convergence without the bounded gradients assumption. In *ICML*, 2018.

[13] Lam M. Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *ICML*, 2017.

[14] Lam M. Nguyen, Jie Liu, Katya Scheinberg, and Martin Takác. Stochastic recursive gradient algorithm for nonconvex optimization. *CoRR*, abs/1705.07261, 2017.

[15] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, 2nd edition, 2006.

[16] Sashank J. Reddi, Ahmed Hefny, Suvrit Sra, Barnabás Póczos, and Alexander J. Smola. Stochastic variance reduction for nonconvex optimization. In *ICML*, pages 314–323, 2016.

[17] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[18] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, pages 1–30, 2016.

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

Appendix

A W.p.1. result

Lemma 9 ([2]). *Let Y_k, Z_k, and W_k, $k = 0, 1, \dots$, be three sequences of random variables and let $\{\mathcal{F}_k\}_{k \geq 0}$ be a filtration, that is, σ-algebras such that $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ for all k. Suppose that:*

- *The random variables Y_k, Z_k, and W_k are nonnegative, and $\mathcal{F}_k$-measurable.*
- *For each k, we have $\mathbb{E}[Y_{k+1}|\mathcal{F}_k] \leq Y_k - Z_k + W_k$.*
- *There holds, w.p.1,*

$$\sum_{k=0}^{\infty} W_k < \infty.$$

Then, we have, w.p.1,

$$\sum_{k=0}^{\infty} Z_k < \infty \text{ and } Y_k \to Y \geq 0.$$

Theorem 2. *Consider SGD with a stepsize sequence such that*

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

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

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

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

Proof. The following proof follows the proof in [12]. From Theorem 1 we recall the recursion

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

Let $Z_t = 2\eta_t(1 - \eta_t L)E_t$ and $W_t = 2\eta_t^2 N$. Since $\sum_{t=0}^{\infty} W_t = \sum_{t=0}^{\infty} 2\eta_t^2 N < \infty$, by Lemma 9, we have w.p.1

$$\mathbb{E}[Y_t|\mathcal{F}_t] \to Y \geq 0,$$

$$\sum_{t=0}^{\infty} Z_t = \sum_{t=0}^{\infty} 2\eta_t(1 - \eta_t L)[F(w_t) - F(w_*)] < \infty.$$

We want to show that $[F(w_t) - F(w_*)] \to 0$ w.p.1. Proving by contradiction, we assume that there exists $\epsilon > 0$ and t_0, s.t., $[F(w_t) - F(w_*)] \geq \epsilon$ for $\forall t \geq t_0$. Hence,

$$\begin{aligned}\sum_{t=t_0}^{\infty} 2\eta_t(1-\eta_t L)[F(w_t) - F(w_*)] &\geq \sum_{t=t_0}^{\infty} 2\eta_t(1-\eta_t L)\epsilon \\ &= 2\epsilon \sum_{t=t_0}^{\infty} \eta_t - 2L\epsilon \sum_{t=t_0}^{\infty} \eta_t^2 = \infty.\end{aligned}$$

This is a contradiction. Therefore, $[F(w_t) - F(w_*)] \to 0$ w.p.1.

□

B Convexity and Curvature

B.1 Function $\delta(\cdot)$

For smooth functions $a : w \in \mathbb{R}^d \to [0, \infty)$ and $b : w \in \mathbb{R}^d \to [0, \infty)$ 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 $w \in \mathcal{B} \subseteq \mathbb{R}^d$. We assume $\delta(\epsilon) < \infty$ for $\epsilon \geq 0$.

The next lemmas show that function $\delta(\cdot)$ has a number of interesting properties that are useful for us when applied to

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

and

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

Lemma 10. *Let p be a probability distribution over $w \in \mathcal{B} \subseteq \mathbb{R}^d$. Then, $\mathbb{E}_p[b(w)] \leq \delta(\mathbb{E}_p[a(w)])$.*

Proof. Let $\epsilon = \mathbb{E}_p[a(w)]$. By the definition of function $\delta(\cdot)$, since $\mathbb{E}_p[a(w)] \leq \epsilon$,

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

□

To gain insight about distribution p in the definition of function $\delta(\epsilon)$ we define

$$\mathcal{W}(\hat{\epsilon}) = \{w \ : \ a(w) = \hat{\epsilon}\},$$

$$\hat{p}(w) = \Pr_p(w | w \in \mathcal{W}(\hat{\epsilon})),$$

and

$$q(\hat{\epsilon}) = \Pr_p(w \in \mathcal{W}(\hat{\epsilon})) = p(\mathcal{W}(\hat{\epsilon})).$$

Notice that from q and $\hat{p}$ we can infer p and vice versa.

By Bayes' rule, for $w \in \mathcal{W}(\hat{\epsilon})$,

$$p(w) = q(\hat{\epsilon})\hat{p}(w).$$

Note that the space $\mathbb{R}^d$ is partitioned into the infinitely many subsets $\mathcal{W}(\hat{\epsilon})$ for $\hat{\epsilon} \geq 0$. This proves

$$\mathbb{E}_p[a(w)] \;=\; \int\limits_{\hat{\epsilon}\geq 0} \int\limits_{w\in\mathcal{W}(\hat{\epsilon})} a(w)\hat{p}(w)q(\hat{\epsilon})\, \mathrm{d}w\mathrm{d}\hat{\epsilon}$$

(these integrals make sense because we assume smooth functions $a(w)$ and $b(w)$, which also implies that $\mathcal{W}(\hat{\epsilon})$ is connected and even convex if $a(\cdot)$ is convex). By the definition of $\mathcal{W}(\hat{\epsilon})$, $a(w) = \hat{\epsilon}$ for $w \in \mathcal{W}(\hat{\epsilon})$. This leads to the simplification

$$\mathbb{E}_p[a(w)] \;=\; \int\limits_{\hat{\epsilon}\geq 0} \hat{\epsilon}q(\hat{\epsilon}) \left\{ \int\limits_{w\in\mathcal{W}(\hat{\epsilon})} \hat{p}(w)\, \mathrm{d}w \right\} \mathrm{d}\hat{\epsilon} = \int\limits_{\hat{\epsilon}\geq 0} \hat{\epsilon}q(\hat{\epsilon})\mathrm{d}\hat{\epsilon} = \mathbb{E}_q[\hat{\epsilon}],$$

where q is considered a distribution over parameter $\hat{\epsilon} \in [0, \infty)$.

The above shows that we may rewrite

$$\begin{aligned} \delta(\epsilon) &= \sup_{p:\mathbb{E}_p[a(w)]\leq\epsilon} \mathbb{E}_p[b(w)] \\ &= \sup_{q:\mathbb{E}_q[\hat{\epsilon}]\leq\epsilon} \; \sup_{\hat{p} \text{ over } w\in\mathcal{W}(\hat{\epsilon})\cap\mathcal{B}} \mathbb{E}_{q,\hat{p}}[b(w)] \\ &= \sup_{q:\mathbb{E}_q[\hat{\epsilon}]\leq\epsilon} \; \sup_{\hat{p} \text{ over } w\in\mathcal{W}(\hat{\epsilon})\cap\mathcal{B}} \mathbb{E}_q \left[\int\limits_{w\in\mathcal{W}(\hat{\epsilon})} \hat{p}(w)b(w)\, \mathrm{d}w \right]. \end{aligned}$$

For $\mathcal{W}(\hat{\epsilon}) \cap \mathcal{B} \neq \emptyset$, we implicitly define $w(\hat{\epsilon})$ as a solution of

$$b(w(\hat{\epsilon})) = \sup_{w\in\mathcal{W}(\hat{\epsilon})\cap\mathcal{B}} b(w) \tag{14}$$

(here, we assume for simplicity that the supremum can be realized for some $w(\hat{\epsilon})$ and we note that the argument below can be made general with small adaptations). For $\mathcal{W}(\hat{\epsilon}) \cap \mathcal{B} \neq \emptyset$, integral

$$\int\limits_{w\in\mathcal{W}(\hat{\epsilon})} \hat{p}(w)b(w)\, \mathrm{d}w$$

is maximized by distribution $\hat{p}$ over $\mathcal{B}$ defined by

$$\hat{p}(w(\hat{\epsilon})) = 1$$

and zero elsewhere. This proves the next lemma (in the supremum $q(\hat{\epsilon})$ should be chosen equal to 0 if $\mathcal{W}(\hat{\epsilon}) \cap \mathcal{B} = \emptyset$).

Lemma 11.

$$\delta(\epsilon) \;=\; \sup_{q:\mathbb{E}_q[\hat{\epsilon}]\leq\epsilon} \mathbb{E}_q[b(w(\hat{\epsilon}))] = \sup_{q:\mathbb{E}_q[\hat{\epsilon}]\leq\epsilon} \mathbb{E}_q \left[\sup_{w\in\mathcal{W}(\hat{\epsilon})\cap\mathcal{B}} b(w) \right],$$

where q is a distribution over $\hat{\epsilon} \in [0, \infty)$ and $\mathcal{W}(\hat{\epsilon}) = \{w \;:\; a(w) = \hat{\epsilon}\}$.

It turns out that

$$\rho(\hat{\epsilon}) = b(w(\hat{\epsilon})) = \sup_{w \in \mathcal{W}(\hat{\epsilon}) \cap \mathcal{B}} b(w)$$

is increasing for $\mathcal{W}(\hat{\epsilon}) \cap \mathcal{B} \neq \emptyset$, but it may have convex and concave parts. For this reason we are not able to simplify $\delta(\cdot)$ any further.

Given the reformulation of $\delta(\epsilon)$ we are able to prove the following property of function $\delta(\cdot)$.

Lemma 12. *Suppose that $a(w) = 0$ implies both $b(w) = 0$ and $w \in \mathcal{B}$, and suppose that there exists a w_* such that $a(w_*) = 0$. Then, $\delta(0) = 0$, $\delta(\epsilon)$ is increasing, and $\delta(\epsilon)$ is $\cap$-convex.*

Proof. From Lemma 11 we obtain the expression

$$\delta(0) = \sup_{q:\mathbb{E}_q[\hat{\epsilon}] \leq 0} \mathbb{E}_q \left[\sup_{w \in \mathcal{W}(\hat{\epsilon}) \cap \mathcal{B}} b(w) \right],$$

where q is a distribution over $\hat{\epsilon} \in [0, \infty)$. Therefore, $\mathbb{E}_q[\hat{\epsilon}] \leq 0$ if and only if $\mathbb{E}_q[\hat{\epsilon}] = 0$, i.e., the probability $\hat{\epsilon} = 0$ is equal to 1 according to distribution q. This proves

$$\delta(0) = \sup_{w \in \mathcal{W}(\hat{0}) \cap \mathcal{B}} b(w).$$

By definition, if $w \in \mathcal{W}(\hat{0})$, then $a(w) = 0$ and by our assumption $b(w) = 0$ and $w \in \mathcal{B}$. Given the existence of $a(w_*) = 0$, the set $\mathcal{W}(\hat{0}) \cap \mathcal{B}$ is not empty, hence, $\delta(0) = 0$.

We show that $\delta(.)$ is $\cap$-convex: Lemma 11 shows that for any $\epsilon_1, \epsilon_2 > 0$, there exists distributions q_1 and q_2 such that $\mathbb{E}_{q_1}[\hat{\epsilon}] \leq \epsilon_1$, $\mathbb{E}_{q_2}[\hat{\epsilon}] \leq \epsilon_2$, and

$$\begin{aligned} \delta(\epsilon_1) &= \mathbb{E}_{q_1}[b(w(\hat{\epsilon}))] \text{ and} \\ \delta(\epsilon_2) &= \mathbb{E}_{q_2}[b(w(\hat{\epsilon}))]. \end{aligned}$$

(again we assume for simplicity that the supremums in the reformulation of function values $\delta(\epsilon_1)$ and $\delta(\epsilon_2)$ in Lemma 11 can be realized for some q_1 and q_2; we note that the argument below can be made general with small adaptations).

Note that for $q = \alpha q_1 + (1 - \alpha) q_2$, we have

$$\mathbb{E}_q[\hat{\epsilon}] = \alpha \mathbb{E}_{q_1}[\hat{\epsilon}] + (1 - \alpha) \mathbb{E}_{q_2}[\hat{\epsilon}] \leq \alpha \epsilon_1 + (1 - \alpha) \epsilon_2$$

and

$$\begin{aligned} \mathbb{E}_q[b(w(\hat{\epsilon}))] &= \alpha \mathbb{E}_{q_1}[b(w(\hat{\epsilon}))] + (1 - \alpha) \mathbb{E}_{q_2}[b(w(\hat{\epsilon}))] \\ &= \alpha \delta(\epsilon_1) + (1 - \alpha) \delta(\epsilon_2). \end{aligned}$$

This shows that

$$\begin{aligned} \alpha \delta(\epsilon_1) + (1 - \alpha) \delta(\epsilon_2) &= \mathbb{E}_q[b(w(\hat{\epsilon}))] \\ &\leq \sup_{q:\mathbb{E}_{\bar{q}}[\hat{\epsilon}] \leq \alpha \epsilon_1 + (1-\alpha)\epsilon_2} \mathbb{E}_{\bar{q}}[b(w(\hat{\epsilon}))] \\ &= \delta(\alpha \epsilon_1 + (1 - \alpha) \epsilon_2). \end{aligned}$$

□

Let $\omega(x)$ be increasing and $\cap$-convex such that, for all $w \in \mathbb{R}^d$,

$$\omega(a(w)) \geq b(w).$$

Then, since $\omega(x)$ is $\cap$-convex, $\mathbb{E}[\omega(X)] \leq \omega(\mathbb{E}[X])$ for any real valued random variable. Hence,

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

If we assume $\mathbb{E}_p[a(w)] \leq \epsilon$, then

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

because $\omega(\cdot)$ is increasing. By the definition of $\delta(\cdot)$,

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

The other way around holds true as well: Let $\omega(x)$ be increasing and $\cap$-convex such that $\omega(\epsilon) \geq \delta(\epsilon)$ for $\epsilon \geq 0$. Then, by Lemma 10 using point distribution $p(w) = 1$ and 0 elsewhere, $b(w) \leq \delta(a(w))$ for $w \in \mathcal{B}$. Since $\omega(\epsilon) \geq \delta(\epsilon)$, we have $b(w) \leq \omega(a(w))$. We have the following lemma.

Lemma 13. *Let $\omega(x)$ be increasing and $\cap$-convex. Then, $\omega(a(w)) \geq b(w)$ for all $w \in \mathcal{B} \subseteq \mathbb{R}^d$ if and only if $\delta(\epsilon) \leq \omega(\epsilon)$ for all $\epsilon \geq 0$.*

The above lemma shows that $\delta(\cdot)$ is the 'minimal' increasing and $\cap$-convex function with the property $\omega(a(w)) \geq b(w)$ for all w.

B.2 Relating Expectations of $a(\cdot)$ and $b(\cdot)$

We start by noting that, for all $x > 0$ and $y > 0$,

$$\omega(y) \leq \omega(x) + \omega'(x)(y - x) \tag{15}$$

(irrespective of whether $y \leq x$ or $y \geq x$).

Lemma 14. *Let $\omega(\cdot)$ be increasing and $\cap$-convex with $\omega(0) = \tau \geq 0$. Then,*

$$\begin{aligned}
\omega'(x) &\leq \frac{\omega(x) - \tau}{x} \textit{ for all } x > 0, \textit{ and} && (16)\\
c_\alpha(e)\frac{\omega(x)}{x} &\leq \omega'(x) \textit{ for all } 0 < \alpha \leq x \leq e \textit{ and } 0 < e < \textit{sup}\{z \geq 0 \ : \ \omega'(z) \neq 0\}, && (17)\\
\textit{where} & \\
c_\alpha(e) &= \inf_{x\in[\alpha,e]} (\frac{\omega(2x)}{\omega(x)} - 1) > 0. && (18)
\end{aligned}$$

We notice that (1) $\sup\{z \geq 0 \ : \ \omega'(z) \neq 0\} > 0$ if and only if $\omega(x)$ is not the all zero function, and (2) by combining (16) and (17), $c(e) \leq 1$.

Proof. We start by noting that, for all $x > 0$ and $y > 0$,

$$\omega(y) \leq \omega(x) + \omega'(x)(y - x) \tag{19}$$

(irrespective of whether $y \leq x$ or $y \geq x$).

By substituting $y = 0$ in (19), $\tau = \omega(0) \leq \omega(x) + \omega'(x)[0 - x]$. Therefore, $x \cdot \omega'(x) \leq \omega(x) - \tau$ and (16) follows.

By substituting $y = 2x$ in (19), $\omega(2x) \leq \omega(x) + \omega'(x)[2x - x]$. Therefore, $\omega(2x) - \omega(x) \leq x \cdot \omega'(x)$. By the definition of $c_\alpha(e)$, $c_\alpha(e) \leq \frac{\omega(2x)}{\omega(x)} - 1$ or equivalently $(c_\alpha(e)+1)\omega(x) \leq \omega(2x)$. Combining inequalities yields $c_\alpha(e)\omega(x) \leq x \cdot \omega'(x)$ which proves (17).

Since $\omega(x)$ is increasing, $\omega(2x) \geq \omega(x)$ and $c_\alpha(e) \geq 0$. If $c_\alpha(e) = 0$, then there exists an $x \in [\alpha, e]$ such that $\frac{\omega(2x)}{\omega(x)} - 1 = 0$, i.e., $\omega(2x) = \omega(x)$. This implies that $\omega(\cdot)$ is constant on the non-empty interval $[x, 2x]$. Together with $\omega(\cdot)$ being increasing and $\cap$-convex this implies that $\omega(\cdot)$ is constant on $[x, \infty)$, hence, $\omega'(z) = 0$ for $z \geq x$. This means that

$$e < \sup\{z \geq 0 \ : \ \omega'(z) \neq 0\} \leq x,$$

contradicting $x \in [\alpha, e]$. So, $c_\alpha(e) \neq 0$. □

By substituting $y = \mathbb{E}_p[a(w)]$ in (19), we obtain, for all $x \geq 0$,

$$\omega(\mathbb{E}_p[a(w)]) \leq \omega(x) + \omega'(x)[\mathbb{E}_p[a(w)] - x]. \tag{20}$$

Lemma 10 and Lemma 13, where we assume $\omega(a(w)) \geq b(w)$ for all w, prove

$$\mathbb{E}_p[b(w)] \leq \delta(\mathbb{E}_p[a(w)]) \leq \omega(\mathbb{E}_p[a(w)]) \tag{21}$$

(a more direct proof is given below and is also in the main body). Combination of (20) and (21) yields

$$\begin{aligned}\mathbb{E}_p[b(w)] \leq \omega(\mathbb{E}_p[a(w)]) &\leq \omega(x) + \omega'(x)[\mathbb{E}_p[a(w)] - x] \\ &= (\omega(x) - \omega'(x)x) + \omega'(x)\mathbb{E}_p[a(w)].\end{aligned}$$

We infer from Lemma 14 that

$$\omega'(x) \leq \frac{\omega(x) - \tau}{x}.$$

Lemma 14 also shows that for $0 < \alpha \leq x \leq e$ and e small enough

$$\frac{\omega(x)}{\omega'(x)} - x \leq (c_\alpha(e)^{-1} - 1)x,$$

where $c_\alpha(e) > 0$. Applying these results to the above derivation yields, for $\alpha \leq x \leq e$,

$$\frac{x\mathbb{E}_p[b(w)]}{\omega(x) - \tau} \quad \leq \quad \frac{\mathbb{E}_p[b(w)]}{\omega'(x)} \leq (\frac{\omega(x)}{\omega'(x)} - x) + \mathbb{E}_p[a(w)] \leq (c_\alpha(e)^{-1} - 1)x + \mathbb{E}_p[a(w)].$$

If we assume $\omega'(\epsilon) \neq 0$ for all $\epsilon \geq 0$, then

$$\sup\{z \geq 0 \ : \ \omega'(z) \neq 0\} = \infty,$$

hence, e is unrestricted. By using

$$\sup_{e \in [\alpha, \infty)} \ \inf_{x \in [\alpha, e]} \frac{\omega(2x)}{\omega(x)} - 1$$

instead of $c_\alpha(e)$ we obtain the best bound:

Lemma 15. *Let* $\omega : [0,\infty) \to [0,\infty)$ *be* $\cap$*-convex (i.e.* $\omega''(\epsilon) < 0$*) and strictly increasing (i.e.,* $\omega'(\epsilon) > 0$*) with* $\omega(0) = \tau \geq 0$*. If* $\omega(a(w)) \geq b(w)$ *for all* $w \in \mathbb{R}^d$*, then, (1) for all* $0 < \alpha \leq x$*,*

$$\frac{x\mathbb{E}_p[b(w)]}{\omega(x)-\tau} \leq \frac{2-c_\alpha}{c_\alpha - 1}x + \mathbb{E}_p[a(w)],$$

where

$$c_\alpha = \sup_{e\in[\alpha,\infty)} \inf_{x\in[\alpha,e]} \frac{\omega(2x)}{\omega(x)} > 1,$$

and (2) 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. We repeat a more condensed proof for (2): Since $\omega(a(w)) \geq b(w)$ for all w,

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

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

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

By substituting $y = \mathbb{E}_p[a(w)]$ in (19), we obtain

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

Rearranging terms and dividing by $\omega'(x)$ proves property (2). □

We will apply the above lemma to the class of strictly increasing and $\cap$-convex functions

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

for $h \in (0,1]$, $r > 0$, $\mu > 0$, and $\tau \geq 0$. Notice that $\omega_{h,r,\mu,\tau}(0) = \tau$.

Function $\omega_{h,r,\mu,\tau}(x)$ is curved like x^h for x close enough to zero up to $x \leq r$. For $x > r$, values $\omega_{h,r,\mu,\tau}(x)$ are chosen as large as possible under the constraint that $\omega_{h,r,\mu,\tau}(\cdot)$ remains $\cap$-convex (i.e., $\omega_{h,r,\mu,\tau}(x)$ is equal to the tangent of $\frac{2}{\mu}(x/r)^h$ at $x = r$).

In our analysis of c_α we consider three cases. Let $x \geq \alpha$. First, suppose that $2x \leq r$. Then,

$$\frac{\omega_{h,r,\mu,\tau}(2x)}{\omega_{h,r,\mu,\tau}(x)} = \frac{\tau + \frac{2}{\mu}(2x/r)^h}{\tau + \frac{2}{\mu}(x/r)^h} = 1 + \frac{2^h - 1}{(\mu\tau/2)(r/x)^h + 1}.$$

This is minimized for x as small as possible. Assuming $2\alpha \leq r$, allows $x = \alpha$ and achieves

$$\inf_{x\in[\alpha,r/2]} \frac{\omega_{h,r,\mu,\tau}(2x)}{\omega_{h,r,\mu,\tau}(x)} = 1 + \frac{2^h - 1}{(\mu\tau/2)(r/\alpha)^h + 1}. \tag{22}$$

Second, suppose that $x \leq r \leq 2x$. Then,

$$\frac{\omega_{h,r,\mu,\tau}(2x)}{\omega_{h,r,\mu,\tau}(x)} = \frac{\tau + \frac{2}{\mu} + \frac{2}{\mu}h((2x/r)-1)}{\tau + \frac{2}{\mu}(x/r)^h}. \tag{23}$$

Setting variable $y = r/x \in [1,2]$ and taking the derivative with respect to y (and grouping terms) gives

$$\frac{h\frac{2}{\mu}y^{-h-1}\{\frac{2}{\mu}(1-h)[1-2y^{-1}] - \tau[2y^{h-1}-1]\}}{\tau + \frac{2}{\mu}y^{-h}}.$$

The term $[1 - 2y^{-1}] \leq 0$ for $y \in [1,2]$ and the term $[2y^{h-1} - 1] \geq 0$ for $y \in [1,2]$. This shows that the derivative above is ≤ 0 for $y \in [1,2]$. Therefore (23) is minimized for $y = r/x = 2$; we again assume $2\alpha \leq r$. This achieves

$$\inf_{x\in[r/2,r]} \frac{\omega_{h,r,\mu,\tau}(2x)}{\omega_{h,r,\mu,\tau}(x)} = \frac{\tau + \frac{2}{\mu}}{\tau + \frac{2}{\mu}2^{-h}} = 1 + \frac{2^h - 1}{\tau\frac{\mu}{2}2^h + 1}. \tag{24}$$

Third, suppose that $r \leq x$. Then,

$$\frac{\omega_{h,r,\mu,\tau}(2x)}{\omega_{h,r,\mu,\tau}(x)} = \frac{\tau + \frac{2}{\mu} + \frac{2}{\mu}h((2x/r)-1)}{\tau + \frac{2}{\mu} + \frac{2}{\mu}h((x/r)-1)} = \frac{\tau\frac{\mu}{2} + 1 + h((2x/r)-1)}{\tau\frac{\mu}{2} + 1 + h((x/r)-1)} = 1 + \frac{h}{(\tau\frac{\mu}{2} + 1 - h)(r/x) + h}.$$

This is minimized for r/x as large as possible, i.e., $r = x$, which achieves

$$\inf_{x\in[r,\infty)} \frac{\omega_{h,r,\mu,\tau}(2x)}{\omega_{h,r,\mu,\tau}(x)} = 1 + \frac{h}{(\tau\frac{\mu}{2} + 1 - h) + h} = 1 + \frac{h}{\tau\frac{\mu}{2} + 1}. \tag{25}$$

The above analysis shows that the three cases (22), (24), and (25) neatly fit together in that

$$\frac{\omega_{h,r,\mu,\tau}(2x)}{\omega_{h,r,\mu,\tau}(x)} \text{ is increasing in } x \geq \alpha.$$

This proves that c_α is equal to (22);

$$c_\alpha = 1 + \frac{2^h - 1}{(\mu\tau/2)(r/\alpha)^h + 1}.$$

In Lemma 15 we need

$$\frac{2 - c_\alpha}{c_\alpha - 1} = \frac{(\mu\tau/2)(r/\alpha)^h + 1}{2^h - 1} - 1 = \frac{(\mu\tau/2)(r/\alpha)^h + 2 - 2^h}{2^h - 1}.$$

Lemma 15 provides the following results: (1) For $0 < \alpha \leq x \leq r$ with $\alpha \leq r/2$,

$$\frac{\mu}{2}r^h x^{1-h}\mathbb{E}_p[b(w)] \leq \frac{(\mu\tau/2)(r/\alpha)^h + 2 - 2^h}{2^h - 1}x + \mathbb{E}_p[a(w)].$$

By substituting $\alpha = x$, the tightest inequality is obtained:

$$\frac{\mu}{2}r^h x^{1-h}\mathbb{E}_p[b(w)] \leq \frac{(\mu\tau/2)(r/x)^h + 2 - 2^h}{2^h - 1}x + \mathbb{E}_p[a(w)].$$

(2) By substituting exact expressions for $\omega'(x)$ and $\omega(x)$ in Lemma 15, we obtain for all $0 < x \leq r$,

$$\frac{1}{h}\frac{\mu}{2}r^h x^{1-h}\mathbb{E}_p[b(w)] \leq \frac{(\mu\tau/2)(r/x)^h + 1 - h}{h}x + \mathbb{E}_p[a(w)].$$

This shows that the asymptotic dependency on x obtained by the more accurate derivation in (2) is the same as for the slightly less tight derivation giving (1). The technique that led to (1) may be a useful tool in analyzing other functions $\omega(.)$.

We summarize (2) in the following lemma:

Lemma 16. *Let*

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

Then,

$$\frac{1}{h}\frac{\mu}{2}r^h x^{1-h}\mathbb{E}_p[b(w)] \le \frac{(\mu\tau/2)(r/x)^h + 1 - h}{h}x + \mathbb{E}_p[a(w)].$$

In our analysis of the convergence rate we need

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

Notice that the derivative of $\frac{\omega(x)}{\omega'(x)} - x$ is equal to

$$\frac{-\omega(x)\omega''(x)}{\omega'(x)^2} \ge 0,$$

and the derivative $\frac{1}{\omega'(x)}$ is equal to

$$\frac{-\omega''(x)}{\omega'(x)^2} \ge 0.$$

This implies that $v(\eta)$ is increasing and is alternatively defined as

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

For $\omega_{h,r,\mu,\tau}$ we have

$$v(\eta) = \frac{1}{h}\frac{\mu}{2}r^h x^{1-h} \text{ where } \eta = \frac{(\mu\tau/2)(r/x)^h + 1 - h}{h}x.$$

If $\tau \neq 0$, then

$$x = \frac{\eta h}{(\mu\tau/2)(r/x)^h + 1 - h} \le \frac{\eta h}{(\mu\tau/2)(r/x)^h},$$

hence,

$$x^{1-h} \le \frac{2\eta h}{\mu\tau r^h}.$$

We get the upper bound

$$v(\eta) = \frac{1}{h}\frac{\mu}{2}r^h x^{1-h} \le \frac{1}{h}\frac{\mu}{2}r^h\frac{2\eta h}{\mu\tau r^h} = \frac{\eta}{\tau}.$$

This upper bound is tight for small η.

If $\tau = 0$, then

$$x = \frac{\eta h}{1-h},$$

hence,

$$v(\eta) = \frac{1}{h}\frac{\mu}{2}r^h x^{1-h} = \frac{1}{h}\frac{\mu}{2}r^h\left(\frac{\eta h}{1-h}\right)^{1-h} = \frac{\mu}{2}h^{-h}(1-h)^{-(1-h)}r^h\eta^{1-h}.$$

Lemma 17. *For $\omega_{h,r,\mu,\tau}$,*

$$v(\eta) \leq \frac{\eta}{\tau} \text{ if } \tau \neq 0$$

and

$$v(\eta) = \frac{\mu}{2} h^{-h}(1-h)^{-(1-h)} r^h \eta^{1-h} \text{ if } \tau = 0.$$

Notice that taking the limit $h \downarrow 0$ for $\tau = 0$ gives

$$v(\eta) = \frac{\mu}{2}\eta.$$

The limit $h = 1$ gives

$$v(\eta) = \frac{\mu}{2} r,$$

where $r = 1$ corresponds to μ-strongly objective functions.

In our definition and analysis of curvature (in the main text) we use the functions

$$\omega_{h,r,\mu}(x) = \omega_{h,r,\mu h,\tau=0}(x) = \begin{cases} \frac{2}{\mu h}(x/r)^h, & \text{if } x \leq r, \text{ and} \\ \frac{2}{\mu h} + \frac{2}{\mu}((x/r)-1), & \text{if } x > r, \end{cases}$$

for $h \in (0,1]$, $r > 0$, and $\mu > 0$. We conclude from Lemma 17 that for these functions,

$$v(\eta) = \frac{\mu h}{2} h^{-h}(1-h)^{-(1-h)} r^h \eta^{1-h}.$$

B.3 Example Curvature $h = 1/2$

Let

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

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

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

Since $H(w)$ is convex,

$$H(w) - H(w') \geq \langle \nabla H(w'), (w - w') \rangle.$$

If we can prove, for all $w, w' \in \mathbb{R}^d$,

$$G(w) - G(w') \geq \langle \nabla G(w'), (w - w') \rangle + \gamma \|w - w'\|^{2/h}, \tag{26}$$

then both inequalities can be added to obtain

$$F(w) - F(w') \geq \langle \nabla F(w'), (w - w') \rangle + \lambda\gamma \|w - w'\|^{2/h}.$$

For $w' = w_* \in \mathcal{W}^*$, i.e., $\nabla F(w_*) = 0$, we obtain

$$F(w) - F(w_*) \geq \lambda\gamma \|w - w_*\|^{2/h}.$$

Since this holds for all w_* and $F(w_*) = F_{min}$, we have

$$F(w) - F(w_*) \geq \lambda\gamma \left\{ \inf_{w_* \in \mathcal{W}^*} \|w - w_*\|^2 \right\}^{1/h}.$$

Hence F is ω-convex over $\mathbb{R}^d$ for $\omega(x) = \frac{2}{\mu h} x^h$ with $\mu = \frac{2\lambda\gamma}{h}$. We conclude that F has curvature h over $\mathbb{R}^d$.

We will prove (26) for $h = 1/2$. We derive

$$G(w) - G(w') = \sum_{i=1}^{d} \left\{ [e^{w_i} + e^{-w_i} - 2 - \alpha w_i^2] - [e^{w'_i} + e^{-w'_i} - 2 - \alpha w_i'^2] \right\}$$

and

$$\langle \nabla G(w'), (w - w') \rangle = \sum_{i=1}^{d} [e^{w'_i} - e^{-w'_i} - \alpha 2 w'_i] \cdot (w_i - w'_i).$$

Let $v_i = w_i - w'_i$ and substitute $w'_i = w_i - v_i$ in the above equations. Then,

$$[G(w) - G(w')] - \langle \nabla G(w'), (w - w') \rangle = \sum_{i=1}^{d} \left\{ e^{w_i}[1 - e^{-v_i} - e^{-v_i} v_i] + e^{-w_i}[1 - e^{v_i} + e^{v_i} v_i] - \alpha v_i^2 \right\} \tag{27}$$

and we want to prove that this is at least

$$\geq \gamma \left\{ \sum_{i=1}^{d} v_i^2 \right\}^{1/h} = \gamma \|w - w'\|^{2/h}.$$

Differentiating (27) with respect to w_i yields

$$e^{w_i}[1 - e^{-v_i} - e^{-v_i} v_i] - e^{-w_i}[1 - e^{v_i} + e^{v_i} v_i]. \tag{28}$$

Notice that $1 - e^{-v_i} - e^{-v_i} v_i \geq 0$ since $e^{v_i} \geq 1 + v_i$ for all v_i. Also notice that $1 - e^{v_i} + e^{v_i} v_i \geq 0$ since $e^{-v_i} \geq 1 - v_i$ for all v_i. This shows that (27) is minimized for w_i for which (28) is equal to 0, i.e.,

$$e^{w_i} = \sqrt{\frac{1 - e^{v_i} + e^{v_i} v_i}{1 - e^{-v_i} - e^{-v_i} v_i}}.$$

Plugging this back into (27) shows that (27) is at most

$$\sum_{i=1}^{d} 2\sqrt{[1 - e^{-v_i} - e^{-v_i} v_i][1 - e^{v_i} + e^{v_i} v_i]} - \alpha v_i^2 = \sum_{i=1}^{d} 2\sqrt{2 - v_i^2 - (e^{v_i} + e^{-v_i}) + v_i(e^{v_i} - e^{-v_i})} - \alpha v_i.$$

We substitute the Taylor series expansion of e^{v_i} and e^{-v_i} and get

$$\sum_{i=1}^{d} 2\sqrt{2 \sum_{j=2}^{\infty} \frac{(2j-1)}{(2j)!} v_i^{2j}} - \alpha v_i^2.$$

The i-th term is at least $\bar{\alpha} v_i^4$ if

$$2 \sum_{j=2}^{\infty} \frac{(2j-1)}{(2j)!} v_i^{2j} \geq (\bar{\alpha} v_i^4 + \alpha v_i^2)^2 / 4 = \frac{\alpha^2}{4} v_i^4 + \frac{\alpha\bar{\alpha}}{2} v_i^6 + \frac{\bar{\alpha}^2}{4} v_i^8. \tag{29}$$

The first three terms of the infinite sum are

$$\frac{v_i^4}{4} + \frac{v_i^6}{72} + \frac{v_i^8}{2880}.$$

So, if we set

$$\alpha = 1 \text{ and } \bar{\alpha} = \frac{1}{36},$$

then (29) is satisfied. So,

$$[G(w) - G(w')] - \langle \nabla G(w'), (w - w') \rangle \geq \sum_{i=1}^{d} \bar{\alpha} v_i^4.$$

Since the 2-norm and 4-norm satisfy

$$\|w - w'\|_2 = \left(\sum_{i=1}^{d} v_i^2 \right)^{1/2} \leq d^{1/4} \left(\sum_{i=1}^{d} v_i^4 \right)^{1/4},$$

we obtain

$$\sum_{i=1}^{d} \bar{\alpha} v_i^4 \geq \frac{\bar{\alpha}}{d} \|w - w'\|^4.$$

This proves (26) for $h = 1/2$ and

$$\gamma = \frac{\bar{\alpha}}{d} = \frac{1}{36d}.$$

Hence, F is ω-convex over $\mathbb{R}^d$ for $\omega(x) = \frac{2}{\mu h} x^h$ with $\mu = \frac{2}{\lambda \gamma h}$.

Theorem 5. *Let*

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

be our objective function where $\lambda > 0$, $H(x)$ 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 $\mu = \frac{2\lambda\gamma}{h} = \frac{\lambda}{9d}$.

As the derivation shows, it is not possible to prove a curvature $> 1/2$ (the bounds are tight in that we can always find an example which violates a larger curvature).

The associated $v(\eta)$ as defined in (8) is equal to

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

for $\mu \geq 0$. In effect the r^h term of $\omega_{h,\mu,r}$ is absorbed in μ and $r \to \infty$. Function ω in the above theorem is equal to $\lim_{r \to \infty} \omega_{h,\mu/r^h,r}$.

C Proof Convergence Rate

Lemma 18. *Let $n(\cdot)$ be a decreasing step size function representing $n(t) = \eta_t$. Define*

$$M(y) = \int_{x=0}^{y} n(x)v(n(x))dx \textit{ and } C(t) = \exp(-M(t)) \int_{x=0}^{t} \exp(M(x))n(x)^2 dx.$$

Then recurrence

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

implies

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

for constants $A = (2N+1)\exp(n(0))$ and $B = (2N+1)\exp(M(1))n(0)^2 + \mathbb{E}[Y_0]$ (depending on parameter N and starting vector w_0).

Proof. We first define some notation:

$$y_t = \mathbb{E}[Y_t], \;\; n(t) = \eta_t.$$

Here, y_t measures the expected convergence rate and $n(t)$ is the step size function which we assume to be decreasing in t.

By using induction in t, we can solve the recursion as

$$y_{t+1} \leq \sum_{i=0}^{t} [\prod_{j=i+1}^{t} (1 - n(j)v(n(j)))](2N+1)n(i)^2 + y_0 \prod_{i=0}^{t} (1 - n(i)v(n(i))).$$

Since $1 - x \leq \exp(-x)$ for all $x \geq 0$,

$$\prod_{j=i+1}^{t} (1 - n(j)v(n(j))) \leq \exp(-\sum_{j=i+1}^{t} n(j)v(n(j))).$$

Since $n(j)$ is decreasing in j and $v(\eta)$ is increasing in η, $n(j)v(n(j))$ is decreasing in j and we have

$$\sum_{j=i+1}^{t} n(j)v(n(j)) \geq \int_{x=i+1}^{t+1} n(x)v(n(x))dx.$$

Combining the inequalities above, we have

$$\begin{aligned} y_{t+1} &\leq \sum_{i=0}^{t} \exp(-\sum_{j=i+1}^{t} n(j)v(n(j)))(2N+1)n(i)^2 + y_0 \exp(-\sum_{j=0}^{t} n(j)v(n(i))) \\ &\leq \sum_{i=0}^{t} \exp(-\int_{x=i+1}^{t+1} n(x)v(n(x))dx)(2N+1)n(i)^2 + y_0 \exp(-\int_{x=0}^{t+1} n(x)v(n(x))dx) \\ &= \sum_{i=0}^{t} \exp(-[M(t+1) - M(i+1)])(2N+1)n(i)^2 + \exp(-M(t+1))y_0, \end{aligned}$$

where

$$M(y) = \int_{x=0}^{y} n(x)v(n(x))dx \text{ and } \frac{d}{dy}M(y) = n(y)v(n(y)).$$

We further analyze the sum in the above expression:

$$\begin{aligned} S &= \sum_{i=0}^{t} \exp(-[M(t+1) - M(i+1)])(2N+1)n(i)^2 \\ &= \exp(-M(t+1)) \sum_{i=0}^{t} \exp(M(i+1))(2N+1)n(i)^2. \end{aligned}$$

We know that $\exp(M(x+1))$ increases and $n(x)^2$ decreases, hence, in the most general case either their product first decreases and then starts to increase or their product keeps on increasing. We first discuss the decreasing and increasing case. Let $a(x) = \exp(M(x+1))n(x)^2$ denote this product and let integer $j \geq 0$ be such that $a(0) \geq a(1) \geq \ldots \geq a(j)$ and $a(j) \leq a(j+1) \leq a(j+2) \leq \ldots$ (notice that $j = 0$ expresses the situation where $a(i)$ only increases). Function $a(x)$ for $x \geq 0$ is minimized for some value h in $[j, j+1)$. For $1 \leq i \leq j$, $a(i) \leq \int_{x=i-1}^{i} a(x)\mathrm{d}x$, and for $j+1 \leq i$, $a(i) \leq \int_{x=i}^{i+1} a(x)dx$. This yields the upper bound

$$\begin{aligned} \sum_{i=0}^{t} a(i) = a(0) + \sum_{i=1}^{j} a(i) + \sum_{i=j+1}^{t} a(i) \\ \leq a(0) + \int_{x=0}^{j} a(x)dx + \int_{x=j+1}^{t+1} a(x)dx, \\ \leq a(0) + \int_{x=0}^{t+1} a(x)dx. \end{aligned}$$

The same upper bound holds for the other case as well, i.e., if $a(i)$ is only decreasing. We conclude

$$S \leq (2N+1)\exp(-M(t+1))[\exp(M(1))n(0)^2 + \int_{x=0}^{j+1} \exp(M(x+1))n(x)^2 dx].$$

Combined with

$$M(x+1) = \int_{y=0}^{x+1} n(y)n(v(y))dy \leq \int_{y=0}^{x} n(y)n(v(y))dy + n(x) = M(x) + n(x)$$

we obtain

$$\begin{aligned} S &\leq (2N+1)\exp(-M(t+1))[\exp(M(1))n(0)^2 + \int_{x=0}^{t+1} \exp(M(x))n(x)^2 exp(n(x))dx] \\ &\leq (2N+1)\exp(-M(t+1))[\exp(M(1))n(0)^2 + \exp(n(0)) \int_{x=0}^{t+1} \exp(M(x))n(x)^2 dx]. \end{aligned}$$

This gives

$$\begin{aligned} y_{t+1} &\leq (2N+1)\exp(-M(t+1))[\exp(M(1))n(0)^2 \\ &\quad + \exp(n(0)) \int_{x=0}^{t+1} \exp(M(x))n(x)^2 dx] + y_0 \exp(-M(t+1)) \\ &= (2N+1)\exp(n(0))C(t+1) \\ &\quad + \exp(-M(t+1))[(2N+1)\exp(M(1))n(0)^2 + y_0], \end{aligned}$$

where

$$C(t) = \exp(-M(t)) \int_{x=0}^{t} \exp(M(x)) n(x)^2 dx.$$

□

Notice that if $v(\eta) = c \cdot \eta$ for some constant c, then $C(t) = (1 - \exp(-M(t)))/c$, which approaches $1/c$ rather than 0 for $t \to \infty$. In the main text we already concluded that linear $v(\eta)$ do not contain any information.

We want to minimize $C(t)$ by appropriately choosing the step size function $n(t)$. We first compute the derivative

$$C'(t) = -n(t)v(n(t))C(t) + n(t)^2 = n(t)v(n(t))[\frac{n(t)}{v(n(t))} - C(t)]. \quad (30)$$

Notice that $C(t)$ is decreasing, i.e., $C'(t) < 0$, if and only if

$$C(t) \geq \frac{n(t)}{v(n(t))}. \quad (31)$$

This shows that $C(t)$ can at best approach $\frac{n(t)}{v(n(t))}$. For example,

$$\bar{C}(t) = 2\frac{n(t)}{v(n(t))} \quad (32)$$

would be close to optimal. Substituting (32) back into (30) gives

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

Hence,

$$n(t) = \sqrt{-\bar{C}'(t)}$$

and substituting this back into (32) gives the differential equation

$$\bar{C}(t) = \frac{2\sqrt{-\bar{C}'(t)}}{v(\sqrt{-\bar{C}'(t)})}.$$

For the corresponding step size function $n(t)$ we know that the actual $C(t)$ starts to behave like $\bar{C}(t)$ for large enough t, i.e., as soon as (31) is approached. So, after $C(0) = 0$ (for $t = 0$, the term $B \cdot \exp -M(0)$ will dominate), $C(t)$ increases until it crosses $\frac{n(t)}{v(n(t))}$ after which it starts decreasing and approaches $\bar{C}(t)$, the solution of the differential equation. We have

$$\frac{n(t)}{v(n(t))} = \bar{C}(t)/2 \leq C(t) \text{ for } t \text{ large enough}$$

and

$$C(t) \leq \bar{C}(t) \text{ for all } t \geq 0.$$

The above can be used to show that the actual $C(t)$ is at most a factor 2 larger than the smallest $C(t)$ over all possible step size functions $n(t)$ for t large enough.

Lemma 19. *A close to optimal step size can be computed by solving the differential equation*

$$\bar{C}(t) = \frac{2\sqrt{-\bar{C}'(t)}}{v(\sqrt{-\bar{C}'(t)})}$$

and equating

$$n(t) = \sqrt{-\bar{C}'(t)}.$$

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, $C(t) \geq \bar{C}(t)/2$.

We will solve the differential equation for

$$v(\eta) = \beta h \eta^{1-h}$$

with $h \in (0, 1]$, where β is a constant and $0 \leq \eta \leq r$ for some $r \in (0, \infty]$ (including the possibility $r = \infty$). This gives the differential equation

$$\bar{C}(t) = \frac{2}{\beta h}\left(\sqrt{-\bar{C}'(t)}\right)^h = \frac{2}{\beta h}[-\bar{C}'(t)]^{h/2}.$$

We try the solution

$$\bar{C}(t) = ct^{-h/(2-h)}$$

for some constant c. Plugging this into the differential equation gives

$$ct^{-h/(2-h)} = \frac{2}{\beta h}[\frac{ch}{(2-h)}t^{-h/(2-h)-1}]^{h/2}.$$

Notice that $(-h/(2-h)-1)(h/2) = -(2/(2-h))(h/2) = -h/(2-h)$ so that the t terms cancel. We need to satisfy

$$c = \frac{2}{\beta h}[\frac{ch}{(2-h)}]^{h/2},$$

i.e., we must choose

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

For $n(t)$ we derive

$$n(t) = \sqrt{-\bar{C}'(t)} = [(h/(2-h))ct^{-h/(2-h)-1}]^{1/2} = \sqrt{ch/(2-h)}t^{-1/(2-h)} = \left(\frac{2}{\beta(2-h)}\right)^{1/(2-h)} t^{-1/(2-h)}.$$

We need to be careful about the initial condition: $\eta_0 \leq \frac{1}{2L}$ and also $\eta_t \leq \eta_0 \leq r$. To realize these conditions we make

$$\eta_0 = \min\{\frac{1}{2L}, r\}$$

by defining $\eta_t = n(t + \Delta)$ for some suitable Δ. Notice that by starting with the largest possible step size, $C(t)$ will cross $\frac{n(t)}{v(n(t))}$ as soon as possible so that it starts approaching the close to optimal $\bar{C}(t)$ as soon as possible.

Lemma 20. *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

$$\bar{C}(t) = [1/(2-h)]^{h/(2-h)}(2/\beta)^{2/(2-h)}(t + \frac{2\max\{2L, 1/r\}}{\beta(2-h)})^{-h/(2-h)}$$

with

$$n(t) = \left(\frac{2}{\beta(2-h)}\right)^{1/(2-h)}(t + \frac{2\max\{2L, 1/r\}}{\beta(2-h)})^{-1/(2-h)}.$$

We will apply the lemma to two cases: $v(\eta) = \mu h \eta^{1-h}$ with $h = 1/2$, and $v(\eta) = \frac{\mu}{2} h \eta^{1-h}$ for $h = 1$. In both cases $r = \infty$.

We first consider $h = 1/2$ for which $\beta = \mu$. This gives $\Delta = \frac{8L}{3\mu}$. We obtain

$$\begin{aligned} \bar{C}(t) &= (2^5/3)^{1/3}\frac{1}{\mu^{4/3}}(t + \frac{8L}{3\mu})^{-1/3}, \\ n(t) &= (\frac{4}{3\mu})^{2/3}(t + \frac{8L}{3\mu})^{-2/3} = (\frac{4}{3\mu t + 8L})^{2/3}, \\ A &= (2N+1)e^{1/(2L)}. \end{aligned}$$

We can apply this to $\mathbb{E}[Y_t]$ and $\frac{1}{t}\sum_{i=t+1}^{2t}\mathbb{E}[E_i]$ in Theorem 4:

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

and

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

Next we consider the case $h = 1$ for which $\beta = \frac{\mu}{2}$. This gives $\Delta = \frac{8L}{\mu}$. We obtain

$$\begin{aligned} \bar{C}(t) &= (4/\mu)^2(t + \frac{8L}{\mu})^{-1}, \\ n(t) &= \frac{4}{\mu}(t + \frac{8L}{\mu})^{-1} = \frac{4}{\mu t + 8L}, \\ A &= (2N+1)e^{1/(2L)}. \end{aligned}$$

We can apply this to $\mathbb{E}[Y_t]$ and $\frac{1}{t}\sum_{i=t+1}^{2t}\mathbb{E}[E_i]$ in Theorem 4:

$$\mathbb{E}[Y_t] \leq (2N+1)e^{1/(2L)}\frac{1}{\mu}\frac{16}{(\mu t + 8L)} + O(t^{-2})$$

and

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