Restricted-dimension subgradient descent: asymptotic bounds on error

Emmanuel Sales
University of British Columbia
emsal@cs.ubc.ca

April 2020

Abstract

Convex optimization, the study of minimizing convex functions over convex sets, is host to a multitude of methods with far-reaching applications in machine learning. For methods in convex optimization, it is often of interest to analyze the asymptotic bounds of the error of the method, or in other words, how close the result of the method gets to the minimum value after some set period of time.

Gradient descent, an iterative procedure that involves taking the gradients of convex functions at a sequence of points, is one of the most important algorithms in the field of convex optimization. Subgradient descent refers to gradient descent that can be applied to functions that need not be smooth. The primary focus of this text is this particular type of gradient descent.

This text will explore error estimates of the asymptotic bounds of the final iterate error of subgradient descent; that is, the error between the result of the subgradient descent procedure after a set number of steps T. Prior work has established tight asymptotic error bounds that are independent of the dimension of the underlying set; this work explores the possibility of there existing tighter bounds when the dimension of the underlying set is restricted to a finite constant d that is independent of T. In this work we have proven that in the case of $d = 1$, the final iterate error has an upper bound of $O\left(\frac{1}{\sqrt{T}}\right)$.

1 Introduction: Convex optimization

1.1 Convex sets and functions

The domain of convex optimization is concerned with problems of the following form: given a *convex function* f on a *convex* set $\mathcal{S}$, devise an algorithm for finding an approximate minimum of $f(x)$ on $\mathcal{S}$. Definitions of the relevant concepts are as below:

Definition 1.1 (Convex set)**.** A set $\mathcal{S} \subseteq \mathbb{R}$ is **convex** if $\forall x, y \in \mathcal{S}, \forall \theta \in [0, 1], \theta x + (1 - \theta)y \in \mathcal{S}$.

Definition 1.2 (Convex function)**.** A function $f : \mathcal{S} \to \mathbb{R}$ is **convex** if $\forall x, y \in \mathcal{S}, \forall \theta \in [0, 1], f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y)$.

1.2 Convex optimization problems

In general, convex optimization methods are concerned with functions wherein we have limited or no prior information about its global landscape and optima. Instead, different methods are concerned with scenarios in which, given a point x in the set $\mathcal{S}$ of interest, we can query the value of the function $f(x)$ and possibly its first- or second-order derivatives. For functions that are not smooth, first-order methods can make use of elements of the *subdifferential* of the function at a point.

Definition 1.3 (subdifferential)**.** Let $\mathcal{S} \subset \mathbb{R}^d$ be convex. The **subdifferential** of a function $f : \mathbb{R}^d \to \mathbb{R}$ at the point $\mathbf{x}$, denoted $\partial f(x)$ is the set of directions $g \in \mathbb{R}^d$ such that, for all $y \in \mathcal{S}$, $f(y) \geq f(x) + g^T(y - x)$.

1.2.1 Convex optimization in machine learning

In machine learning, convex optimization problems often take the form of finding $\min_{w\in\mathbb{R}^n} \sum_{i=1}^{m} f_i(w) + \lambda\mathcal{R}(w)$, where the functions $f_1, \ldots, f_m, \mathcal{R}$ are convex and $\lambda \geq 0$ is a known parameter (Bubeck, [1]). Convex optimization is particularly useful in these cases because local minima of convex functions are guaranteed to be global minima, thus simplifying the criteria needed to find global minima.

In machine learning these functions f_i are determined by a *data set* $(x_i, y_i) \in \mathbb{R}^n \times \mathcal{Y}$ for $i = 1, \ldots, m$. The f_i generally represent a *loss function* that quantifies how close a function $g_w(x_i)$, parameterized by w is able to return a value close to y_i for each x_i. The lower f_i is for each i, in general, the closer $g_w(x_i)$ is able to attain the results y_i for every i. The function $\mathcal{R}(w)$ acts as a regularizer, which imposes additional constraints imposed on x in the form of a penalty.

In the problem of *linear regression* for example the $g_w(x)$ we are interested in is $w^T x_i$. Setting $f_i(w, x_i) = (w^T x_i - y_i)^2$ and $\mathcal{R}(w) = 0$, we obtain the least squares minimization problem, which can be reframed in matrix notation as $\min\|Xw - y\|_2$, where $X \in \mathbb{R}^{m\times n}$ is the matrix with the ith row equivalent to x_i^T, and $y \in mathbbR^m$ is the vector of y_i values. The least-squares problem has been thoroughly theoretically explored and it has an analytical closed-form solution of $w = (X^T X)^{-1} X^T y$ when the matrix X has full rank.

Many other problems also fall into this framework but do not reap the benefits of a closed-form solution. For example, setting the f_i as in the case for linear regression but adding $\mathcal{R}(w) = \|w\|_1$ produces the *lasso* problem ([2]), which requires iterative methods in order to solve.

1.3 Convex optimization methods and gradient descent

First-order methods are convex optimization methods that involve the first derivative or gradient of the function. These are useful when the global landscape of the function (and thus, its optimum points) cannot be determined analytically, but it is computationally easy to compute the values of the functions at any given point x.

Central to the analysis of first-order methods is the notion of *oracle complexity*, which Bubeck ([1]) defines as how many queries to an "oracle" that, given an input x, either produces either the value of $f(x)$ (zeroth order oracle) or the value of a subgradient (more in Section 2) of f at x.

Gradient descent is an iteration on the equation

$$x_{t+1} = x_t - \eta_t \nabla f(x_t) \tag{1}$$

Results have been shown for lower bounds that would apply to any gradient descent optimization method, mostly by proving the existence of functions for which any black box procedure would only get within a certain distance of the minimum value of the function within t iterations (or oracle queries). In this case, a *gradient descent method* is defined as an iterative procedure $\{x_i\}$ and $\{g_i\}$ where $g_i = \nabla f(x_i)$, the initial $x_1 = 0$, and $x_{t+1} \in \text{Span}(g_1, \ldots, g_t)$.

These theorems make use of the important additional concept of strongly convex functions, noted as being significant because functions of this class significantly speed up the performance of first-order methods.

Definition 1.4 (Strongly convex function)**.** A function $f : \mathcal{S} \to \mathbb{R}$ is **α-strongly convex** if $\forall x, y \in \mathcal{S}$, f satisfies the inequality $f(x) - f(y) \leq \nabla f(x)^T(x-y) - \frac{\alpha}{2}\|x-y\|^2$. If f is not smooth then $\nabla f(x)$ can be replaced by $g_x \in \partial f(x)$.

These are Theorems 3.13, 3.14, and 3.15 in Bubeck.

Theorem 1.1. *Let $t \leq n$ and $L, R > 0$. There exists a convex and L-Lipschitz function f such that for any gradient descent black-box procedure,*

$$\min_{1 \leq s \leq t} f(x_s) - \min_{\|x\|_2 \leq R} f(x) \geq \frac{RL}{2(1+\sqrt{t})} \tag{2}$$

Theorem 1.2. *Let $t \leq (n-1)/2, \beta > 0$. There exists a continuously differentiable function f for which $\|\nabla f(x) - \nabla f(y)\| \leq \beta\|x-y\|$ for all x and y, such that for any gradient descent procedure,*

$$\min_{1 \leq s \leq t} f(x_s) - f(x^*) \geq \frac{3\beta}{32} \frac{\|x_1 - x^*\|^2}{(t+1)^2} \tag{3}$$

Theorem 1.3. *Let $\kappa > 1$. There exists an α-strongly convex continuously differentiable function f for which $\|\nabla f(x) - \nabla f(y)\| \leq \beta\|x-y\|$ for all x and y, with $\kappa = \beta/\alpha$ such that for any gradient descent procedure, we have*

$$f(x_t) - f(x^*) \geq \frac{\alpha}{2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2(t-1)} \|x_1 - x^*\|^2 \tag{4}$$

2 Subgradient descent

We examine subgradient descent on convex and bounded sets.

2.1 Definition of subgradient descent

Subgradient descent is a type of gradient descent that works on functions that are not necessarily smooth. One important usage of subgradient descent is in the case of L_1 minimization and regularization; loss functions of the form $\sum_{i=1}^n |g(X) - y|$ are nonsmooth and require the use of subgradients, as a gradient cannot be selected at every point. We have

$$x_{t+1} = x_t - \eta_t g_t \tag{5}$$

where $g_t \in \partial f(x_t)$, not requiring the gradient $\nabla f(x_t)$ to exist as in (1).

The algorithm can be written as follows:

Algorithm 1: SubgradientDescent

---

x$_1$ $\leftarrow$ some initial guess in the set $\mathcal{S}$
for t *in* $2 \ldots T$ **do**
  $\eta_t \leftarrow \frac{1}{\sqrt{t}}$
  g$_t$ $\leftarrow$ an element in $\partial f(x)$
  x$_{t+1}$ $\leftarrow \Pi_S(x_t - \eta_t g_t)$
end

---

where Π, the projection operator, is defined by $\Pi_{\mathcal{S}}(z) = \arg\min_{x \in \mathcal{S}} \|x - z\|_2$.

2.1.1 Choices of step size

There are many possible choices of η_t. Boyd, Xia, and Mutapcic ([3]) list four:

- Constant step size: $\eta_t = h$, independent of t
- Constant step length: $\eta_t = h/g_t$, such that $\|x_{t+1} - x_t\|$ is the same for all t
- Square summable but not summable: choices of η_t such that $\sum_{t=1}^{\infty} \eta_t^2 < \infty$ and $\sum_{t=1}^{\infty} \eta_t = \infty$.
- Nonsummable diminishing: choices of η_t such that $\lim_{t\to\infty} \eta_t = 0$ and $\sum_{i=1}^{\infty} \eta_t = \infty$.

In Section 3, where we provide a proof related to final iterate error for the case where $\dim(\mathcal{S}) = 1$, we examine the algorithm with $\eta_t = \frac{1}{\sqrt{t}}$, which is an example of a nonsummable diminishing step size rule.

2.2 Convergence results

Definition 2.1. The final iterate error of a run of subgradient descent is the value of $f(x_T) - f(x^*)$.

Boyd ([3]) states that for the diminishing step size rule, the limit $\lim_{t\to\infty} f(x_t) = f(x^*)$, i.e. the algorithm is guaranteed to converge to the optimal value.

Prior work has established that for a run of SGD with T iterations on a T-dimensional space, the final iterate error is $\Omega(\log(T)/\sqrt{T})$ [4].

An upper bound for the expected final iterate error for stochastic gradient descent without smoothness assumptions was established by Shamir and Zhang [5] as $O(\log T/\sqrt{T})$ over non-smooth convex objective functions.

2.3 Goals of this work

This work's primary aim is to examine the properties of subgradient descent on convex functions in arbitrary finite-dimensional cases. We aim to examine the possibility that tighter bounds can be achieved than existing results if we restrict the dimension of the space we are analyzing to a finite constant d.

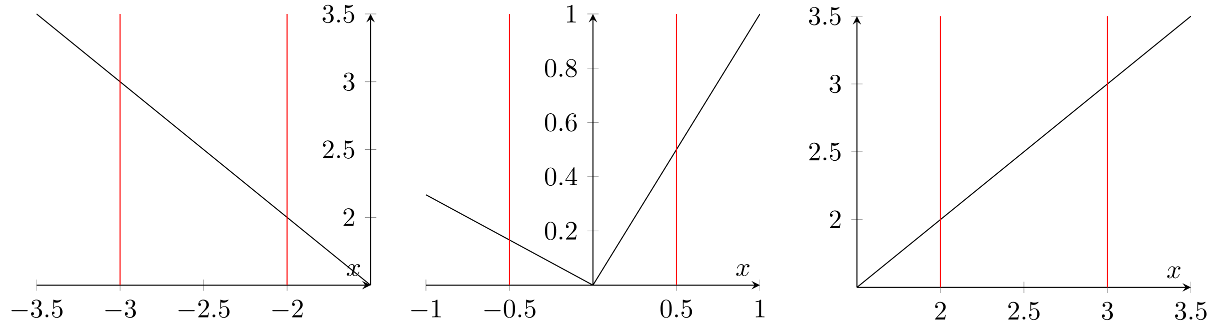
In Section 3, we establish a tight upper bound on the final iterate error of the algorithm as $O(1/\sqrt{T})$ in the case where $d = 1$.

3 Upper bound for final iterate error where $d = 1$

The primary result is the following theorem on the final iterate error of subgradient descent in the restricted 1-dimensional case. This is a tighter result than the general upper bound achieved by Shamir and Zhang ([5]) of $O(\log(T)/\sqrt{T})$.

Theorem 3.1. *Let $\mathcal{S} \subset \mathbb{R}$ be convex and bounded (without loss of generality, let diam$(S) = 1$). Let $f : \mathcal{S} \to \mathbb{R}$ be a convex and 1-Lipschitz.*

Then the final iterate error $|f(x_T) - f(x^*)|$ of a run of subgradient descent is in $\frac{O(1)}{\sqrt{T}}$.



(a) (b) (c)

Figure 1: Example of three cases where Theorem 3.1 applies. Figures 1a and 1c are examples of convex functions that are monotonically decreasing on their underlying set, and hence has the minimizer x^* lying at one of the endpoints. Figure 1b denotes the case where the minimizer x^* resides in the middle.

Proof. We prove this by induction on the number of iterations.

The base case is on the first iteration, $T = 1$, or the initial value. Set $c = \max(\sqrt{2}, f(x_1) - f(x^*))$. Any initial starting point will have error that is less than or equal to $\frac{c}{\sqrt{1}}$.

The inductive step: Suppose the iteration error $f(x_{T-1}) - f(x^*) \leq \frac{c}{\sqrt{T-1}}$. We will show that this implies

$$f(x_T) - f(x^*) \leq \frac{c}{\sqrt{T}} \tag{6}$$

We have

$$f(x_T) - f(x^*) = f(x_T) - f(x_{T-1}) + f(x_{T-1}) - f(x^*) \tag{7}$$

which is known to be less than $(f(x_T) - f(x_{T-1})) + \frac{c}{\sqrt{T-1}}$ by the inductive hypothesis.

We then examine the quantity $f(x_T) - f(x_{T-1})$. By the subgradient descent algorithm we have

$$x_T = \Pi_S(x_{T-1} - \eta_{T-1} g_{T-1}) \tag{8}$$

with $\eta_{T-1} = \frac{1}{\sqrt{T-1}}$ and g_{T-1} in the subgradient of $f(x_{T-1})$.

Let $\dot{x}_T = x_{T-1} - \eta_{T-1} g_{T-1}$, with $x_T = \Pi_{\mathcal{S}}(\dot{x}_T)$.

Because f is 1-Lipschitz, we know that $|g_i| \leq 1$ for all i.

An important property used in showing the inequality is the monotonicity of the subgradient on convex functions:

Claim 3.1.1 (Monotonicity of the subgradient)**.** For convex f, for all $x, y \in \mathcal{S}, g_x \in \partial f(x), g_y \in \partial f(y)$, $(g_x - g_y)(x - y) \geq 0$.

Proof. By the subgradient inequality we have both $f(y) \geq f(x) + g_x(y - x)$ and $f(x) \geq f(y) + g_y(x - y)$. Combining these two inequalities we have $0 \geq (g_x - g_y)(y - x)$, or, by chaging signs, our intended inequality of $(g_x - g_y)(x - y) \geq 0$. □

To prove the induction we condition on three cases depending on the presence of 0 in $\partial f(x_T)$ and $\partial f(x_{T-1})$.

Case 1: If $0 \in \partial f(x_T)$ then we have $f(x_T) = f(x^*)$, which makes $f(x_T) - f(x^*) = 0$, which is less than $\frac{c}{\sqrt{T}}$ for all T.

Case 2: If $0 \notin \partial f(x_T)$ but $0 \in \partial f(x_{T-1})$, we have $x_{T-1} = x^*$ (the optimum) and thus $|x_T - x_{T-1}| = \eta_{T-1}|g_{T-1}|$. Because f is 1-Lipschitz we have $|x_T - x_{T-1}| \leq \eta_{T-1} = \frac{1}{\sqrt{T-1}}$; again by Lipschitz we thus conclude $f(x_T) - f(x^*) = f(x_T) - f(x_{T-1}) \leq \frac{c}{\sqrt{T}}$.

Case 3: $0 \notin \partial f(x_{T-1})$ and $0 \notin \partial f(x_T)$. For this case we show that (6) holds by showing that it holds on all possible cases: $\text{sign}(\max \partial f(x_T)) = \text{sign}(\max \partial f(x_{T-1}))$, and $\text{sign}(\max \partial f(x_T)) \neq \text{sign}(\max \partial f(x_{T-1}))$.

Case 3.1: $\text{sign}(\max \partial f(x_T)) = \text{sign}(\max \partial f(x_{T-1}))$.

There is a special case for when x_T is on the boundary of the set $\mathcal{S}$. As such, we condition on whether or not x_T is on the boundary of $\mathcal{S}$.

Case 3.1a: x_T is on the boundary of $\mathcal{S}$. In that case, we are able to show the following:

Claim 3.1.2. If x_T is on the boundary of $\mathcal{S}$ and $\text{sign}(\max(\partial f(x_T))) = \text{sign}(\max(\partial f(x_{T-1})))$, then x_T must be a global minimizer for f on $\mathcal{S}$.

Proof. Because x_T is on the boundary of a one-dimensional set, either $x_T \geq x \forall x \in \mathcal{S}$ or $x_T \leq x \forall x \in \mathcal{S}$. Because $x_T = x_{T-1} - \eta_{T-1} g_{T-1}$, and $\eta_{T-1} > 0$, we are able to conclude by algebra that $\text{sign}(g_T) = \text{sign}(g_{T-1}) = -\text{sign}(x_T - x_{T-1}) = \text{sign}(x_{T-1} - x_T)$.

By the subgradient inequality we have $f(x) \geq f(x_T) + g_T(x - x_T), \forall x \in \mathcal{S}$. Because $\text{sign}(g_T) = \text{sign}(x_{T-1} - x_T) = \text{sign}(x - x_T) \forall x \in \mathcal{S}$, we have $f(x_T) + g_T(x - x_T) \geq f(x_T)$ and therefore $f(x) \geq f(x_T) \forall x \in \mathcal{S}$. $\square$

Thus $f(x_T) = f(x^*)$ and thus we have $f(x_T) - f(x^*) = 0 \leq \frac{c}{\sqrt{T}}$.

Case 3.1b: x_T is in the interior of $\mathcal{S}$. For this case we use the following claim:

Claim 3.1.3. If $\text{sign}(\max \partial f(x_T)) = \text{sign}(\max \partial f(x_{T-1}))$, then $f(x_T) \leq f(x_{T-1})$ and $|g_T| \leq |g_{T-1}|$.

Proof. Applying Claim 3.1.1 to x_T and x_{T-1}, we have

$$(g_T - g_{T-1})(x_T - x_{T-1}) = (g_T - g_{T-1})(-\eta_{T-1} g_{T-1}) \geq 0 \tag{9}$$

We thus have $-g_T g_{T-1} + g_{T-1}^2 \geq 0$. Because we know that g_T and g_{T-1} are the same sign, we know that $g_T g_{T-1} \geq 0$. Thus we know that $|g_{T-1}^2| \geq |g_T g_{T-1}|$ and thus $|g_T| \leq |g_{T-1}|$.

Using this fact, and the fact that g_T and g_{T-1} have the same sign, we know that either $g_{T-1} \leq g_T < 0$ or $0 < g_T \leq g_{T-1}$. We remark that 0 is in the subgradient of a point z if and only if z is a global minimizer (a point x^*) for f, as f is convex. Thus by monotonicity of the subgradient we have either $x_{T-1} < x_T \leq x^*$ or $x^* \leq x_T < x_{T-1}$, associated with each case respectively. We are thus able to conclude

$$\text{sign}(g_T) = \text{sign}(x_{T-1} - x_T)$$

We know that by the subgradient inequality that $f(x_{T-1}) \geq f(x_T) + g_T(x_{T-1} - x_T)$. Because g_T and $(x_{T-1} - x_T)$ have the same sign, the quantity is nonnegative; thus we have

$$f(x_{T-1}) \geq f(x_T) \quad \square$$

Corollary 3.1.1. *sign*$(g_T) =$ *sign*(g_{T-1})*, because* $0 \notin \partial f(x_T)$ *and* $0 \notin \partial f(x_{T-1})$.

To show that $f(x_T) \leq \frac{c}{\sqrt{T}}$, it is sufficient to show that

$$f(x_{T-1}) - f(x_T) \geq \frac{c}{\sqrt{T-1}} - \frac{c}{\sqrt{T}} \tag{10}$$

We examine the quantity $\eta_T g_T$, examining the properties of the cases where $|g_T| \leq \frac{\alpha}{\sqrt{T}}$ and where $g_T > \frac{\alpha}{\sqrt{T}}$, where α is a number chosen in the interval $[\sqrt{c}, c]$, which is a nonempty interval because $c \geq 1$.

If $|g_T| \leq \frac{\alpha}{\sqrt{T}}$: we have

$$\begin{aligned}
f(x_T) - f(x^*) &\leq |g_T||x_T - x^*| && \text{by convexity/monotonicity of the gradient} \\
&\leq \frac{\alpha}{\sqrt{T}}(1) && (|x_T - x^*| \leq 1 \text{ by the diameter of the domain}) \\
&\leq \frac{c}{\sqrt{T}}
\end{aligned}$$

If $|g_T| > \frac{\alpha}{\sqrt{T}}$:

By convexity we know that

$$\begin{aligned}
|g_T| &\leq \frac{f(x_{T-1}) - f(x_T)}{|x_{T-1} - x_T|} && \text{(By Corollary 3.1.1)} \\
&= \frac{f(x_{T-1}) - f(x_T)}{|\eta_{T-1} g_{T-1}|} \leq \frac{f(x_{T-1}) - f(x_T)}{|\eta_{T-1} g_T|} && \text{(By Claim 3.1.3)} \\
\eta_{T-1}|g_T|^2 &\leq f(x_{T-1}) - f(x_T) = [f(x_{T-1}) - f(x^*)] - [f(x_T) - f(x^*)] \\
&\leq \frac{c}{\sqrt{T-1}} - [f(x_T) - f(x^*)]
\end{aligned}$$

By rearranging we deduce

$$f(x_T) - f(x^*) \leq \frac{c}{\sqrt{T-1}} - \eta_{T-1}|g_T|^2 \tag{11}$$

By assumption we have $|g_T| \geq \frac{\alpha}{\sqrt{T}}$. We thus have

$$\begin{aligned}
f(x_T) - f(x^*) &\leq \frac{c}{\sqrt{T-1}} - \frac{1}{\sqrt{T-1}}\frac{\alpha^2}{T} \\
&\leq \frac{c}{\sqrt{T-1}} - \frac{\alpha^2/2}{(T-1)^{1.5}}
\end{aligned} \tag{12}$$

This is true because for $T \geq 2$, $\frac{1}{T} \geq \frac{1}{2(T-1)}$.

By applying Claim A.1 to (12) we thus have

$$f(x_T) - f(x^*) \leq \frac{c}{\sqrt{T}} \tag{13}$$

as intended.

Case 3.2: $\text{sign}(\max \partial f(x_T)) \neq \text{sign}(\max \partial f(x_{T-1}))$

We know that if 0 is in the subgradient of a point $x \in \mathcal{S}$, then x is a global minimizer for f.

Combining this optimality condition with fact with the monotonicity of the subgradient (Claim 3.1.1) we are able to conclude that a global minimizer for f, x^*, is in the open interval with endpoints x_T and x_{T-1}; this can either be (x_T, x_{T-1}) or (x_{T-1}, x_T). Thus we know that

$$\begin{aligned}
&|x_T - x^*| \leq |x_T - x_{T-1}| = |\eta_{T-1} g_{T-1}| \leq \eta_T \\
&\qquad (|g_{T-1}| \leq 1 \text{ because } f \text{ is 1-Lipschitz}) \\
\implies &f(x_T) - f(x^*) \leq |x_T - x^*| \leq \eta_{T-1} \\
&\qquad = \frac{1}{\sqrt{T-1}} \leq \frac{c}{\sqrt{T}} \\
&\qquad (\text{as } c \geq \sqrt{2})
\end{aligned} \tag{14}$$

Thus in all cases we have $f(x_T) - f(x^*) \leq \frac{c}{\sqrt{T}}$. Thus, we complete the induction. We thus conclude that $f(x_T) - f(x^*) \leq \frac{O(1)}{\sqrt{T}}$. $\square$

4 Future directions

The natural extension of the established work is to attempt further theoretical ground work for higher-dimensional cases and possibly a generalization to any arbitrary finite dimension d. Lower bounds have been established for arbitrary-dimensional cases where $d > T$, namely that the error $f(x_T) - f(x^*) = \Omega(\frac{\log(d)}{\sqrt{T}})$, shown in ([4]). Following that, it is natural to speculate that the bound is tight, such that the error is $O\left(\frac{\log \min(d,T)}{\sqrt{T}}\right)$.

It is also of interest to study lower bounds. As shown in the theorems from [1] listed in Section 1.3, lower bounds can be established by proving the existence of a function $f : \mathcal{S} \subset \mathbb{R}^d \to \mathbb{R}$, such that running subgradient descent on f produces a final iterate error after T iterations that is bounded below by some function of d and T.

References

[1] S. Bubeck *et al.*, "Convex optimization: Algorithms and complexity," *Foundations and Trends® in Machine Learning*, vol. 8, no. 3-4, pp. 231–357, 2015.

[2] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[3] S. Boyd, L. Xiao, and A. Mutapcic, "Subgradient methods," 2003.

[4] N. J. A. Harvey, C. Liaw, Y. Plan, and S. Randhawa, "Tight analyses for non-smooth stochastic gradient descent," *CoRR*, vol. abs/1812.05217, 2018.

[5] O. Shamir and T. Zhang, "Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes," in *Proceedings of the 30th International Conference on Machine Learning* (S. Dasgupta and D. McAllester, eds.), vol. 28 of *Proceedings of Machine Learning Research*, (Atlanta, Georgia, USA), pp. 71–79, PMLR, 17–19 Jun 2013.

A Scalar inequalities

Claim A.1. If $T \geq 2$, then $\frac{\alpha^2/2}{(T-1)^{1.5}} \geq \frac{c}{\sqrt{T-1}} - \frac{c}{\sqrt{T}}$.

Proof. We show this by the following:

$$
\begin{aligned}
c\left(\frac{1}{\sqrt{T-1}} - \frac{1}{\sqrt{T}}\right) &= c\left(\frac{\sqrt{T} - \sqrt{T-1}}{\sqrt{T(T-1)}}\right) \\
&= \frac{c}{\sqrt{T(T-1)}(\sqrt{T} + \sqrt{T-1})} && \text{(Using } a - b = \frac{a^2 - b^2}{a+b}\text{)} \\
&\leq \frac{c}{(T-1)(2\sqrt{T-1})} && \text{(replace every } T \text{ with } T-1 \text{ in the denominator)} \\
&= \frac{c}{2(T-1)^{1.5}} \leq \frac{\alpha^2/2}{(T-1)^{1.5}}
\end{aligned}
$$

(as $\alpha^2 \geq c$). $\square$